Міністерство освіти і науки України

Дніпровський національний університет імені Олеся Гончара

Я. С. Бондаренко, Д. О. Рачко, А. О. Розливан

ПОСІБНИК ДО ВИВЧЕННЯ ДИСЦИПЛІНИ “ІМОВІРНІСНІ ГРАФІЧНІ МОДЕЛІ” ЧАСТИНА 2. НАВЧАННЯ БАЙЄСІВСЬКОЇ МЕРЕЖІ

Дніпро

2020

УДК 519.21:004.032.26
Б81

Рецензенти: канд. фіз.-мат. наук, доц. М.Є. Ткаченко,
канд. фіз.-мат. наук, доц. А.М. Пасько

Б81 Бондаренко Я. С. Посібник до вивчення дисципліни “Імовірнісні графічні моделі”. Частина 2. Навчання байєсівської мережі [Текст] / Я.С. Бондаренко, Д.О. Рачко, А.О. Розливан. – Дніпро: Ліра, 2020. – 40 с.

Викладено теоретичні положення щодо оцінювання невідомих параметрів умовних імовірнісних розподілів вершин байєсівської мережі за умови відомої структури мережі та повних даних.

Для студентів механіко-математичного факультету ДНУ спеціальності “Статистика”.

Рекомендовано до друку вченою радою
механіко-математичного факультету
Дніпровського національного університету імені Олеся Гончара
протокол №5 від 15.12.2020 року

Навчальне видання

Яна Сергіївна Бондаренко
Деніс Олексійович Рачко
Анастасія Олександрівна Розливан

Посібник до вивчення дисципліни
“Імовірнісні графічні моделі”
Частина 2. Навчання байєсівської мережі

Друкується за авторською редакцією

Підписано до друку 28.12.2020. Формат 60×84/16. Папір друкарський. Друк плоский. Ум. друк. арк. 2,33. Тираж 20 пр. Зам. № 335.

Друкарня «Ліра», вул. Наукова, 5, м. Дніпро, 49107. Свідоцтво про внесення до Державного реєстру серія ДК №6042 від 26.02.2018 р.

ВСТУП

Байєсівські мережі застосовуються для побудови систем прийняття рішень в медицині, генетиці, фінансах та банківській справі, військовій справі, космічних дослідницьких програмах, системах розпізнавання зображень та мовних сигналів, освіті. Успішність побудови байєсівської мережі для дослідження реального процесу залежить від вміння коректно поставити задачу, встановити причинно-наслідкові зв'язки між величинами, які в повній мірі характеризують процес, зібрати статистичні дані, навчити мережу і застосувати точні та/або наближені алгоритми формування ймовірнісного висновку для побудови моделей міркувань на основі мережі.

Задача навчання байєсівської мережі полягає в знаходженні оцінок невідомих параметрів умовних імовірнісних розподілів вершин мережі за умови: 1) відомої структури мережі та повних даних; 2) невідомої структури мережі та повних даних; 3) відомої структури мережі та неповних даних; 4) невідомої структури мережі та неповних даних; 5) наявності прихованих змінних в структурі мережі.

Ми розглянемо найпростішу задачу оцінювання параметрів байєсівської мережі за умови відомої структури мережі та повних даних, розв'язання якої виступає підґрунтям для навчання байєсівської мережі за умови невідомої структури та/або неповних даних.

Метод максимальної правдоподібності та байєсівський метод статистичного оцінювання застосовуються для оцінювання невідомих параметрів дискретних умовних імовірнісних розподілів вершин мережі. Ідея декомпозиції функції максимальної правдоподібності у вигляді добутку локальних функцій максимальної правдоподібності відіграє ключову роль при оцінюванні методом максимальної правдоподібності та дозволяє здобути оцінки параметрів аналітично для кожної вершини окремо. Ідея використання розподілу Діріхле як спряженого апріорного розподілу до мультиноміального розподілу вершин мережі відіграє основну роль при оцінюванні байєсівським методом і дозволяє здобути оцінки параметрів аналітично та оновлювати їх протягом онлайн-навчання байєсівської мережі [1-4].

1. Оцінювання параметрів методом максимальної правдоподібності

1.1. Метод максимальної правдоподібності

Нехай $\xi = (\xi_1, \xi_2, \dots, \xi_n)$ – вибірка з розподілом $F(\,\cdot\,;\theta) = F(\,\cdot\,;\,\theta_1, \dots, \theta_s)$, який залежить від параметра $\theta = (\theta_1, \theta_2, \dots, \theta_s) \in \Theta \subset \mathbb{R}^s$. Параметр $\theta \in \Theta$ невідомий і його необхідно оцінити за вибіркою $(\xi_1, \xi_2, \dots, \xi_n)$.

Загальним (важливим як з точки зору теорії, так і застосувань) методом побудови оцінок є метод максимальної правдоподібності, запропонований Р. Фішером. Функцією максимальної правдоподібності вибірки $\xi = (\xi_1, \xi_2, \dots, \xi_n)$ називають функцію $L(\theta) = L(\theta_1, \theta_2, \dots, \theta_s)$ параметра $\theta \in \Theta$, яка визначається рівністю:

$$L(\theta) = f(\xi; \theta), \;\; \theta \in \Theta,$$

якщо вибірковий вектор $\xi = (\xi_1, \xi_2, \dots, \xi_n)$ абсолютно неперервний зі щільністю

$$f(x; \theta) = f(x_1, x_2, \dots, x_n; \theta),$$

та рівністю

$$L(\theta) = P(\xi; \theta), \;\; \theta \in \Theta,$$

якщо вибірковий вектор $\xi = (\xi_1, \xi_2, \dots, \xi_n)$ дискретний з розподілом

$$P(x; \theta) = P(x_1, x_2, \dots, x_n; \theta).$$

Метод максимальної правдоподібності побудови оцінок полягає в тому, що в якості оцінки параметра $\theta = (\theta_1, \theta_2, \dots, \theta_s)$ обирається точка $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_s)$ в якій функція максимальної правдоподібності $L(\theta)$ набуває найбільшого значення.

Оцінкою максимальної правдоподібності називають точку $\hat{\theta}$, в якій функція максимальної правдоподібності $L(\theta)$ набуває найбільшого

значення. Іншими словами, оцінкою максимальної правдоподібності параметра θ називають відмінні від константи розв'язки рівняння:

$$L(\hat{\theta}) = \max_{\theta \in \Theta} L(\theta),$$

якщо такі розв'язки існують. Корені, які не залежать від вибірки $\xi_1, \xi_2, \dots, \xi_n$, тобто які мають вигляд $\hat{\theta} = c$, де c – константа, варто відкинути (оцінка – це функція від вибірки).

Логарифм $\ln L(\theta)$ від функції максимальної правдоподібності $L(\theta)$ називають логарифмічною функцією максимальної правдоподібності.

Зауважимо, що функції $L(\theta)$ та $\ln L(\theta)$ досягають найбільшого значення в одній і тій самій точці. А знайти точку, в якій функція $\ln L(\theta)$ досягає найбільшого значення, часто простіше.

Тому якщо функція $L(\theta) = L(\theta_1, \theta_2, \dots, \theta_s)$ диференційовна за $\theta_1, \theta_2, \dots, \theta_s$, то для розв'язку рівняння

$$L(\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_s) = \max_{\theta_1, \theta_2, \dots, \theta_s \in \Theta} L(\theta_1, \theta_2, \dots, \theta_s) \qquad (1.1.1)$$

достатньо знайти стаціонарні точки функції

$$\ln L(\theta_1, \theta_2, \dots, \theta_s).$$

Розв'язавши рівняння

$$\frac{\partial}{\partial \theta_i} \ln L(\theta_1, \theta_2, \dots, \theta_s) = 0, \;\; i = 1,2, \dots, \mathrm{s}, \qquad (1.1.2)$$

та порівнюючи значення функції $\ln L(\theta_1, \theta_2, \dots, \theta_s)$ в стаціонарних точках і на границях множини Θ, обрати точку $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_s)$, в якій функція $\ln L(\theta_1, \theta_2, \dots, \theta_s)$ досягає найбільшого значення. Ця точка і буде розв'язком рівняння (1.1.1). Рівняння (1.1.2) називають рівнянням максимальної правдоподібності [5].

1.2. Мультиноміальний розподіл

Нехай проводиться M незалежних випробувань, у кожному з яких з однаковою ймовірністю, що не залежить від результатів інших випробувань, відбувається одна з подій $A_1, A_2, \dots, A_K$. Імовірність того, що в даному випробуванні відбудеться подія A_k дорівнює θ_k:

$$P(A_k) = \theta_k, k = 1, 2, \dots K; \quad \theta_1 + \cdots + \theta_K = 1.$$

Імовірність того, що в M незалежних випробуваннях подія A_k відбудеться $M[k]$ разів, $k = 1, 2, \dots K$, дорівнює:

$$\frac{M!}{M[1]!\,M[2]!\,\dots M[K]!}\theta_1^{M[1]}\theta_2^{M[2]}\dots\theta_K^{M[K]}. \tag{1.2.1}$$

Набір імовірностей (1.2.1) визначає розподіл K-вимірної дискретної випадкової величини X, який називається мультиноміальним з параметрами $(M, \theta_1, \theta_2, \dots, \theta_K)$:

$$P\big(X = (M[1], M[2], \dots, M[K])\big) = \frac{M!}{M[1]!\,M[2]!\,\dots M[K]!}\theta_1^{M[1]}\theta_2^{M[2]}\dots\theta_K^{M[K]},$$

$$M[1] + M[2] + \dots + M[K] = M.$$

Надалі за подію A_k розглядатимемо подію «випадкова величина X набуває значення x^k».

Запишемо функцію максимальної правдоподібності мультиноміального розподілу:

$$L(\boldsymbol{\theta}, D) = \frac{M!}{M[1]!\,M[2]!\,\dots M[K]!}\prod_{k=1}^{K}\theta_k^{M[k]} = M!\prod_{k=1}^{K}\frac{\theta_k^{M[k]}}{M[k]!}\,. \tag{1.2.2}$$

Тоді логарифмічна функція правдоподібності дорівнює:

$$l(\boldsymbol{\theta}) = \ln L(\boldsymbol{\theta}, D) = \ln M! + \ln\prod_{k=1}^{K}\frac{\theta_k^{M[k]}}{M[k]!},$$

$$l(\boldsymbol{\theta}) = \ln M! + \sum_{k=1}^{K} M[k]\ln\theta_k - \sum_{k=1}^{K}\ln(M[k]!).$$

Оскільки $\theta_1 + \theta_2 + \ldots + \theta_K = 1$, застосуємо метод множників Лагранжа для знаходження екстремуму функції:

$$l'(\boldsymbol{\theta}, \lambda) = l(\boldsymbol{\theta}) + \lambda\left(1 - \sum_{k=1}^{K} \theta_k\right). \tag{1.2.3}$$

Знайдемо похідні та прирівняємо їх нулеві:

$$\frac{\partial}{\partial \theta_i} l'(\boldsymbol{\theta}, \lambda) = \frac{\partial}{\partial \theta_i} l(\boldsymbol{\theta}) + \frac{\partial}{\partial \theta_i} \lambda\left(1 - \sum_{k=1}^{K} \theta_k\right) = 0,$$

$$\frac{\partial}{\partial \theta_i} l'(\boldsymbol{\theta}, \lambda) = \frac{\partial}{\partial \theta_i} \sum_{k=1}^{K} M[k] \ln \theta_k - \lambda \frac{\partial}{\partial \theta_i} \sum_{k=1}^{K} \theta_k = 0,$$

$$\frac{M[i]}{\theta_i} - \lambda = 0,$$

$$\theta_i = \frac{M[i]}{\lambda}.$$

Скористаємося умовою $\theta_1 + \theta_2 + \ldots + \theta_K = 1$ та здобутою рівністю

$$\theta_i = \frac{M[i]}{\lambda}$$

для знаходження оцінок $\hat{\theta}_i, i = 1, \ldots, K$:

$$1 = \sum_{k=1}^{K} \theta_k = \sum_{k=1}^{K} \frac{M[k]}{\lambda},$$

$$\lambda = \sum_{k=1}^{K} M[k],$$

$$\lambda = M.$$

Отже,

$$\hat{\theta}_i = \frac{M[i]}{M}, \qquad i = 1, \ldots, K.$$

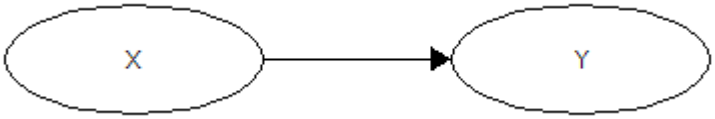
Оцінки максимальної правдоподібності параметрів $\theta_1, \theta_2, \ldots, \theta_K$ дорівнюють

$$\widehat{\boldsymbol{\theta}} = \left(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_K\right) = \left(\frac{M[1]}{M}, \frac{M[2]}{M}, \ldots, \frac{M[K]}{M}\right), \tag{1.2.4}$$

де $M = M[1] + M[2] + \ldots + M[K]$.

1.3. Метод максимальної правдоподібності для байєсівської мережі з двома вершинами

Розглянемо байєсівську мережу:



Імовірнісний розподіл вершини X задається:

$$X \sim \begin{pmatrix} x^0 & x^1 \\ \theta_{x^0} & \theta_{x^1} \end{pmatrix}, \qquad \theta_{x^0} + \theta_{x^1} = 1.$$

Умовний імовірнісний розподіл вершини Y задається:

Y		
X	x^0	x^1
y^0	$\theta_{y^0\|x^0}$	$\theta_{y^0\|x^1}$
y^1	$\theta_{y^1\|x^0}$	$\theta_{y^1\|x^1}$

$$\theta_{y^0|x^0} + \theta_{y^1|x^0} = 1,$$

$$\theta_{y^0|x^1} + \theta_{y^1|x^1} = 1.$$

Позначимо через

$$\boldsymbol{\theta} = \left(\boldsymbol{\theta}_X, \boldsymbol{\theta}_{Y|X}\right) = \left(\theta_{x^0}, \theta_{x^1}, \theta_{y^0|x^0}, \theta_{y^0|x^1}, \theta_{y^1|x^0}, \theta_{y^1|x^1}\right)$$

вектор невідомих параметрів, які необхідно оцінити за методом максимальної правдоподібності за вибіркою $D = \{(x[1], y[1]), \dots, (x[M], y[M])\}$.

Запишемо функцію максимальної правдоподібності:

$$L(\boldsymbol{\theta}, D) = P(z[1], z[2], \dots, z[M]; \boldsymbol{\theta})$$

$$= P(x[1], y[1]; \boldsymbol{\theta}) P(x[2], y[2]; \boldsymbol{\theta}) \cdot \dots \cdot P(x[M], y[M]; \boldsymbol{\theta}).$$

Скористаємося формулою множення для байєсівської мережі:

$$L(\boldsymbol{\theta}, D) = P(x[1]; \boldsymbol{\theta}) P(y[1]|x[1]; \boldsymbol{\theta}) \dots P(x[M]; \boldsymbol{\theta}) P(y[M]|x[M]; \boldsymbol{\theta})$$

$$= P(x[1]; \boldsymbol{\theta}) \dots P(x[M]; \boldsymbol{\theta}) \cdot P(y[1]|x[1]; \boldsymbol{\theta}) \dots P(y[M]|x[M]; \boldsymbol{\theta})$$

$$= \prod_{m=1}^{M} P(x[m]; \boldsymbol{\theta}) \prod_{m=1}^{M} P(y[m]|x[m]; \boldsymbol{\theta}).$$

Ми здобули добуток двох локальних функцій максимальної правдоподібності. Розглянемо кожну окремо. Перша локальна функція максимальної правдоподібності набуває вигляду:

$$\prod_{m=1}^{M} P(x[m];\boldsymbol{\theta}) = \prod_{m=1}^{M} P(x[m];\boldsymbol{\theta}_X)$$

$$= \prod_{m:\, x[m]=x^0} P(x[m];\theta_{x^0}) \prod_{m:\, x[m]=x^1} P(x[m];\theta_{x^1})$$

$$= \prod_{m:\, x[m]=x^0} \theta_{x^0} \prod_{m:\, x[m]=x^1} \theta_{x^1} = \theta_{x^0}^{M[x^0]} \theta_{x^1}^{M[x^1]},$$

де $M[x^0]$ – число вибіркових значень таких, що $x[m]$ набуває значення x^0; $M[x^1]$ – число вибіркових значень таких, що $x[m]$ набуває значення x^1.

Друга локальна функція максимальної правдоподібності запишеться:

$$\prod_{m=1}^{M} P(y[m]|x[m];\boldsymbol{\theta}) = \prod_{m=1}^{M} P\big(y[m]|x[m];\boldsymbol{\theta}_{Y|X}\big)$$

$$= \prod_{m:\, x[m]=x^0} P\big(y[m]|x[m];\boldsymbol{\theta}_{Y|x^0}\big) \prod_{m:\, x[m]=x^1} P\big(y[m]|x[m];\boldsymbol{\theta}_{Y|x^1}\big)$$

$$= \prod_{m:\, x[m]=x^0, y[m]=y^0} P\big(y[m]|x[m];\theta_{y^0|x^0}\big) \prod_{m:\, x[m]=x^0, y[m]=y^1} P\big(y[m]|x[m];\theta_{y^1|x^0}\big)$$

$$\cdot \prod_{m:\, x[m]=x^1, y[m]=y^0} P\big(y[m]|x[m];\theta_{y^0|x^1}\big) \prod_{m:\, x[m]=x^1, y[m]=y^1} P\big(y[m]|x[m];\theta_{y^1|x^1}\big)$$

$$= \prod_{m:\, x[m]=x^0, y[m]=y^0} \theta_{y^0|x^0} \prod_{m:\, x[m]=x^0, y[m]=y^1} \theta_{y^1|x^0}$$

$$\cdot \prod_{m:\, x[m]=x^1, y[m]=y^0} \theta_{y^0|x^1} \prod_{m:\, x[m]=x^1, y[m]=y^1} \theta_{y^1|x^1} =$$

$$= \theta_{y^0|x^0}^{M[x^0,y^0]} \theta_{y^1|x^0}^{M[x^0,y^1]} \theta_{y^0|x^1}^{M[x^1,y^0]} \theta_{y^1|x^1}^{M[x^1,y^1]},$$

де $M[x^0, y^0]$ – число вибіркових значень, для яких $x[m] = x^0, y[m] = y^0$; $M[x^0, y^1]$ – число вибіркових значень, для яких $x[m] = x^0, y[m] = y^1$; $M[x^1, y^0]$ – число вибіркових значень, для яких $x[m] = x^1, y[m] = y^0$; $M[x^1, y^1]$ – число вибіркових значень, для яких $x[m] = x^1, y[m] = y^1$.

Отже, функція максимальної правдоподібності запишеться:

$$L(\boldsymbol{\theta}, D) = \left(\theta_{x^0}^{M[x^0]}\theta_{x^1}^{M[x^1]}\right)\left(\theta_{y^0|x^0}^{M[x^0,y^0]}\theta_{y^1|x^0}^{M[x^0,y^1]}\right)\left(\theta_{y^0|x^1}^{M[x^1,y^0]}\theta_{y^1|x^1}^{M[x^1,y^1]}\right),$$

де $\theta_{x^0}^{M[x^0]}\theta_{x^1}^{M[x^1]}$ – функція максимальної правдоподібності мультиноміального розподілу з параметрами $(\theta_{x^0}, \theta_{x^1})$; $\theta_{y^0|x^0}^{M[x^0,y^0]}\theta_{y^1|x^0}^{M[x^0,y^1]}$ – функція максимальної правдоподібності мультиноміального розподілу з параметрами $\left(\theta_{y^0|x^0}, \theta_{y^1|x^0}\right)$; $\theta_{y^0|x^1}^{M[x^1,y^0]}\theta_{y^1|x^1}^{M[x^1,y^1]}$ – функція максимальної правдоподібності мультиноміального розподілу з параметрами $\left(\theta_{y^0|x^1}, \theta_{y^1|x^1}\right)$.

Оцінки максимальної правдоподібності невідомих параметрів мультиноміальних розподілів знайдемо для кожної локальної функції правдоподібності окремо. Маємо:

$$\hat{\theta}_{x^0} = \frac{M[x^0]}{M[x^0] + M[x^1]}, \quad \hat{\theta}_{x^1} = \frac{M[x^1]}{M[x^0] + M[x^1]},$$

$$\hat{\theta}_{y^0|x^0} = \frac{M[x^0, y^0]}{M[x^0, y^0] + M[x^0, y^1]}, \quad \hat{\theta}_{y^1|x^0} = \frac{M[x^0, y^1]}{M[x^0, y^0] + M[x^0, y^1]},$$

$$\hat{\theta}_{y^0|x^1} = \frac{M[x^1, y^0]}{M[x^1, y^0] + M[x^1, y^1]}, \quad \hat{\theta}_{x^1|y^1} = \frac{M[x^1, y^1]}{M[x^1, y^0] + M[x^1, y^1]}.$$

Для навчання байєсівської мережі з двома вершинами необхідно обчислити величини $M[x^i, y^i]$, $i = 0{,}1$, для кожної комбінації станів вершини Y та станів її батьківської вершини X, а також здобути суми цих значень по всіх можливих станах вершини Y [1, 4].

1.4. Метод максимальної правдоподібності для байєсівської мережі

Розглянемо байєсівську мережу $B = (G, P)$ з відомою структурою G та невідомими параметрами $\boldsymbol{\theta}$ умовних імовірнісних розподілів вершин мережі. Необхідно оцінити невідомі параметри згідно з методом максимальної правдоподібності за вибіркою $\xi[1], \dots, \xi[M]$, де $\xi \in \mathbb{R}^n$. Функція максимальної правдоподібності як функція від параметрів $\boldsymbol{\theta}$ запишеться:

$$L(\boldsymbol{\theta}, D) = P(\,\xi[1], \dots, \xi[M]; \boldsymbol{\theta}) = \prod_{m=1}^{M} P(\xi[m]; \boldsymbol{\theta}).$$

Згідно з формулою множення для байєсівської мережі спільний розподіл вершин $X_1, \dots, X_n$ можна подати як добуток умовних імовірнісних розподілів вершин мережі:

$$P(X_1, \dots, X_n) = \prod_{i=1}^{n} P\left(X_i \middle| Parents_{X_i}\right).$$

Тоді функція максимальної правдоподібності перепишеться у вигляді добутку локальних функцій максимальної правдоподібності :

$$L(\boldsymbol{\theta}, D) = \prod_{m=1}^{M} \prod_{i=1}^{n} P\left(x_i[m] \middle| Pa_{X_i}[m]; \boldsymbol{\theta}\right) =$$

$$= \prod_{i=1}^{n} \prod_{m=1}^{M} P\left(x_i[m] \middle| Pa_{X_i}[m]; \boldsymbol{\theta}\right) = \prod_{i=1}^{n} L_i\left(\boldsymbol{\theta}_{X_i|Pa_{X_i}}\right). \tag{1.4.1}$$

Оскільки параметри $\boldsymbol{\theta}_{X_i|Pa_{X_i}}$ та $\boldsymbol{\theta}_{X_j|Pa_{X_j}}$, $i \neq j$, різних вершин не пов'язані між собою, то оцінки максимальної правдоподібності знаходяться для кожної локальної функції максимальної правдоподібності окремо.

Розглянемо одну локальну функцію максимальної правдоподібності

$$L_i\left(\boldsymbol{\theta}_{X_i|Pa_{X_i}}\right) = P\left(x_i[m] \middle| Pa_{X_i}[m]; \boldsymbol{\theta}_{X_i|Pa_{X_i}}\right)$$

та знайдемо оцінки максимальної правдоподібності.

Локальна функція запишеться як добуток добутків умовних імовірностей по всіх різних можливих значеннях, які набувають батьківські вершини Pa_{X_i} для вершини X_i:

$$L_i\left(\boldsymbol{\theta}_{X_i|Pa_{X_i}}\right) = \prod_{m:Pa_{X_i}[m]=pa_{X_i}^1} P\left(x_i[m] \middle| Pa_{X_i}[m]; \boldsymbol{\theta}_{X_i|pa_{X_i}^1}\right) \cdot \dots$$

$$\cdot \prod_{m:Pa_{X_i}[m]=pa_{X_i}^K} P\left(x_i[m] \middle| Pa_{X_i}[m]; \boldsymbol{\theta}_{X_i|pa_{X_i}^K}\right)$$

Далі локальна функція перепишеться як добуток добутків умовних імовірностей по всіх комбінаціях різних можливих значень, які набувають батьківські вершини Pa_{X_i} для вершини X_i та сама вершина X_i:

$$L_i\left(\boldsymbol{\theta}_{X_i|Pa_{X_i}}\right) = \prod_{m:Pa_{X_i}[m]=pa^1_{X_i},\ x_i[m]=x^1} P\left(x_i[m]\middle|Pa_{X_i}[m];\theta_{x^1|pa^1_{X_i}}\right)\cdot\ldots$$

$$\cdot \prod_{m:Pa_{X_i}[m]=pa^K_{X_i},\ x_i[m]=x^S} P\left(x_i[m]\middle|Pa_{X_i}[m];\theta_{x^S|pa^K_{X_i}}\right).$$

Підставимо умовні ймовірності в останню формулу (це і є наші невідомі параметри, для яких необхідно здобути оцінки):

$$L_i\left(\boldsymbol{\theta}_{X_i|Pa_{X_i}}\right) = \left(\prod_{\substack{m:Pa_{X_i}[m]=pa^1_{X_i},\\ x_i[m]=x^1}} \theta_{x^1|pa^1_{X_i}}\cdot\ldots\cdot \prod_{\substack{m:Pa_{X_i}[m]=pa^1_{X_i},\\ x_i[m]=x^S}} \theta_{x^S|pa^1_{X_i}}\cdot\right)\cdot\ldots$$

$$\cdot\left(\prod_{\substack{m:Pa_{X_i}[m]=pa^K_{X_i},\\ x_i[m]=x^1}} \theta_{x^1|pa^K_{X_i}}\cdot\ldots\cdot \prod_{\substack{m:Pa_{X_i}[m]=pa^K_{X_i},\\ x_i[m]=x^S}} \theta_{x^S|pa^K_{X_i}}\cdot\right)$$

Підраховуємо число вибіркових значень, які задовольняють умовам по яких рахуються добутки і запишемо ці числа як ступені відповідних умовних ймовірностей:

$$L_i\left(\boldsymbol{\theta}_{X_i|Pa_{X_i}}\right) = \left(\theta^{M\left[x^1,pa^1_{X_i}\right]}_{x^1|pa^1_{X_i}}\cdot\ldots\cdot\theta^{M\left[x^S,pa^1_{X_i}\right]}_{x^S|pa^1_{X_i}}\right)\cdot\ldots$$

$$\cdot\left(\theta^{M\left[x^1,pa^K_{X_i}\right]}_{x^1|pa^K_{X_i}}\cdot\ldots\cdot\theta^{M\left[x^S,pa^K_{X_i}\right]}_{x^S|pa^K_{X_i}}\right)$$

Отже, здобули добуток функцій максимальної правдоподібності мультиноміальних розподілів. Оцінки максимальної правдоподібності мультиноміальних розподілів мають вигляд:

$$\hat{\theta}_{x_i|pa_{X_i}} = \frac{M\left[x_i,\ pa_{X_i}\right]}{M\left[pa_{X_i}\right]},\qquad M\left[pa_{X_i}\right] = \sum_{x_i} M\left[x_i, pa_{X_i}\right]. \tag{1.4.2}$$

де $M\left[x_i,\ pa_{X_i}\right]$ – число спостережень таких, що $x_i[m] = x_i,\ Pa_{X_i}[m] = pa_{X_i}$.

Для навчання мережі нам необхідно обчислити величини $M[x_i,\ pa_{X_i}]$ для кожної комбінації станів вершини X_i та станів її батьківських вершин Pa_{X_i}, а також здобути суми $M[pa_{X_i}]$ по всіх можливих станах вершини X_i.

Здобуті оцінки (1.4.2) вказують на основну проблему, яка виникає при оцінюванні параметрів умовних імовірнісних розподілів вершин байєсівської мережі. При збільшенні числа батьківських вершин, число різних можливих комбінацій значень батьківських вершин в умовних імовірнісних розподілах зростає експоненційно. Дана властивість називається *фрагментацією даних*. Інтуїтивно, якщо обсяг навчальної вибірки для оцінювання параметрів малий, то оцінки параметрів умовних імовірнісних розподілів вершин можуть бути поганими, а деякі з великою ймовірністю дорівнюватимуть нулеві. Отже, можливість здобуття поганих оцінок параметрів збільшується зі збільшенням числа батьківських вершин та/або числа їх можливих станів, що є значущим обмеженням при застосуванні методу максимальної правдоподібності для оцінювання параметрів байєсівської мережі [1, 4].

1.5. Байєсівська мережа Credit

Байєсівська мережа *Credit* представлена орієнтованим ациклічним графом, в вершинах якого знаходяться характеристики клієнтів, а орієнтовані ребра представляють вплив однієї характеристики на іншу. Структура мережі та експертні оцінки параметрів умовних імовірнісних розподілів вершин наведені в онлайн-курсі [2]. Так вік клієнта *Age* та відношення боргу до доходу *Ratio of Debts to Income* впливають на історію платежів *Payment History,* вік клієнта *Age* та історія платежів *Payment History* впливають на надійність клієнта *Reliability*, дохід *Income* та активи *Assets* впливають на майбутній дохід *Future Income,* майбутній дохід *Future Income,* відношення боргу до доходу *Ratio of Debts to Income* та надійність клієнта *Reliability* впливають на кредитоспроможність клієнта *Credit Worthiness.*

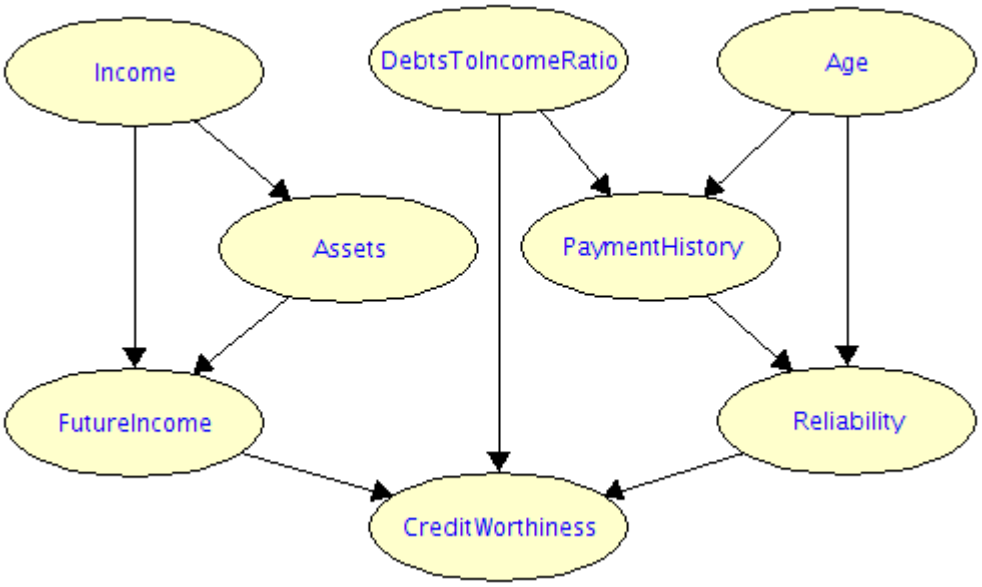


Рис. 1.5.1. Байєсівська мережа Credit [2]

Проведемо дослідження оцінок максимальної правдоподібності параметрів $\boldsymbol{\theta}_{X_i|Pa_{X_i}}$ умовних імовірнісних розподілів вершин байєсівської мережі порівняно з експертними оцінками.

1. Змоделюємо навчальну вибірку даних з експертними оцінками параметрів умовних імовірнісних розподілів вершин мережі та застосуємо метод максимальної правдоподібності для знаходження оцінок максимальної правдоподібності. Кожна вершина X_i має мультиноміальний розподіл з вектором параметрів $\boldsymbol{\theta}_{X_i|Pa_{X_i}}$. Оцінки максимальної правдоподібності мультиноміального розподілу дорівнюють

$$\hat{\theta}_{x_i|pa_{X_i}} = \frac{M[x_i,\ pa_{X_i}]}{M[pa_{X_i}]},$$

де величини $M[x_i,\ pa_{X_i}]$ обчислюються для кожної комбінації значень вершини X_i та значень її батьківських вершин Pa_{X_i}, а суми $M[pa_{X_i}]$ здобуваються по всіх можливих значеннях вершини X_i.

2. Обчислимо дивергенцію Кульбака-Лейблера між експертними та емпіричними умовними ймовірнісними розподілами вершин мережі.

3. Порівняємо експертні оцінки та оцінки максимальної правдоподібності параметрів умовних імовірнісних розподілів вершин.

Залежність відстані Кульбака-Лейблера $D(P \parallel \hat{P})$ від обсягу M навчальної вибірки наведена на рис. 1.5.2. При збільшенні обсягу вибірки відстань Кульбака-Лейблера між експертними та емпіричними умовними ймовірнісними розподілами вершин зменшується, при цьому суттєве скорочення відстані спостерігається для малих обсягів вибірки і повільніше – для великих обсягів.

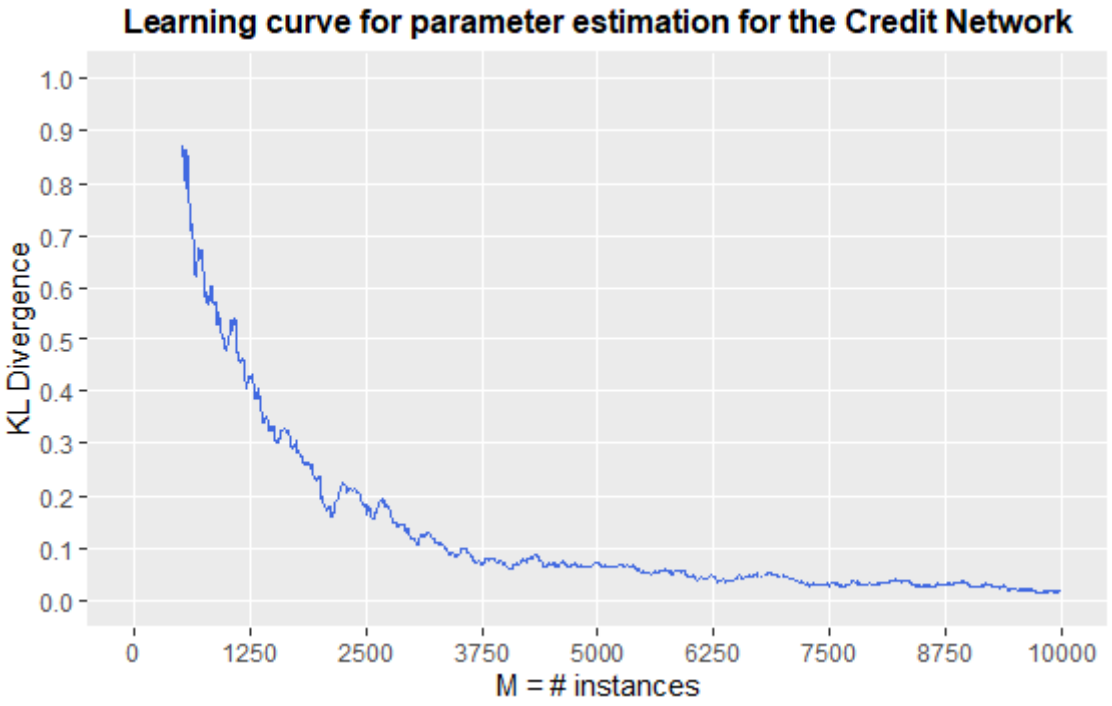


Рис. 1.5.2. Крива навчання байєсівської мережі Credit

Оцінки максимальної правдоподібності – спроможні та асимптотично ефективні оцінки параметрів умовних імовірнісних розподілів вершин байєсівської мережі. Спроможність гарантує збіжність оцінки за ймовірністю до справжнього значення параметра зі зростанням обсягу вибірки. Асимптотична ефективність гарантує прямування дисперсії оцінки до нуля зі зростанням обсягу вибірки. Асимптотична поведінка оцінок максимальної правдоподібності наведена на рис. 1.5.3, 1.5.4. Неперервними лініями позначено оцінки максимальної правдоподібності параметрів, пунктирними лініями – експертні оцінки параметрів.

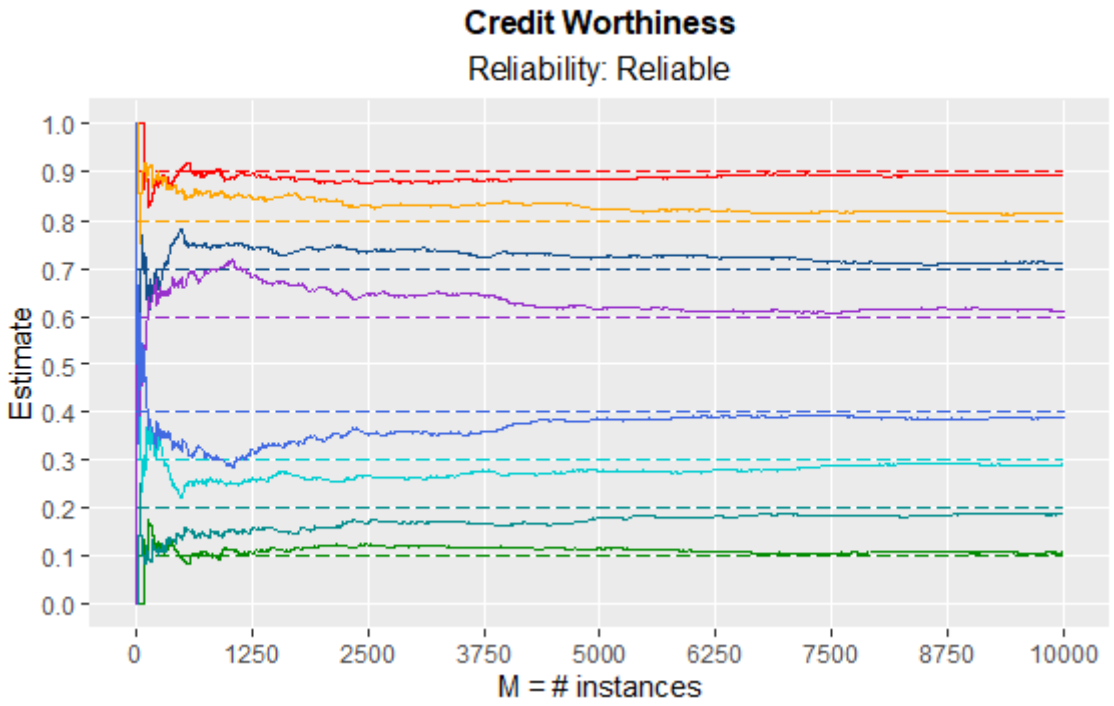


Рис. 1.5.3. Оцінки параметрів вершини Credit Worthiness

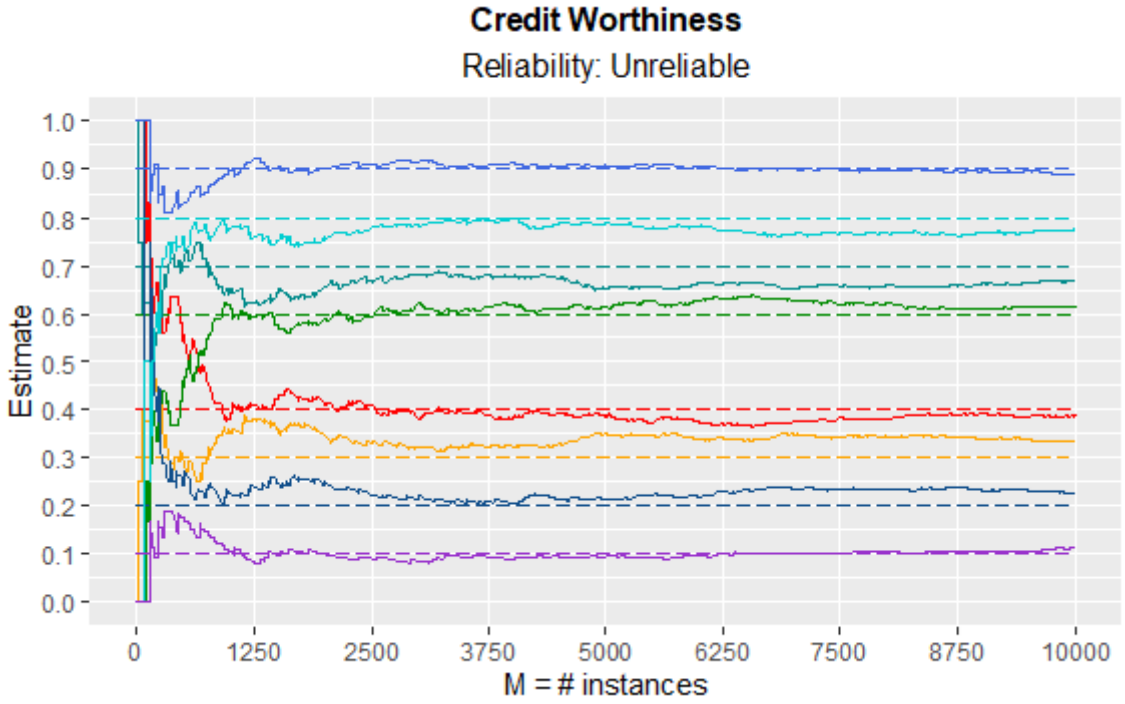


Рис. 1.5.4. Оцінки параметрів вершини Credit Worthiness

Для надійних клієнтів (Reliability: Reliable) збіжність оцінок максимальної правдоподібності до експертних оцінок набагато швидша, ніж для ненадійних клієнтів (Reliability: Unreliable) за рахунок незбалансованості навчальної вибірки даних.

1.6. Якість навчання байєсівської мережі

Оцінки максимальної правдоподібності – спроможні оцінки для оцінювання параметрів умовних імовірнісних розподілів вершин байєсівської мережі. Спроможність гарантує збіжність оцінок до справжніх значень параметрів, якщо обсяг вибірки прямує до нескінченності. На практиці обсяг вибірки обмежений, тому при оцінювання якості навчання моделі як функції обсягу вибірки необхідно відповісти на питання: *яким повинен бути мінімальний обсяг вибірки для здобуття результатів із заданою точністю* ε *та надійністю* $1-\delta$.

Нехай $P(X_1,\dots,X_n)$ та $Q(X_1,\dots,X_n)$ – імовірнісні розподіли дискретних випадкових величин $X_1,\dots,X_n$. Дивергенцією Кульбака-Лейблера називають міру відстані між імовірнісними розподілами, яка визначається так:

$$D(P(X_1,\dots,X_n) \parallel Q(X_1,\dots,X_n)) = \sum P(X_1,\dots,X_n) \ln \frac{P(X_1,\dots,X_n)}{Q(X_1,\dots,X_n)}.$$

Наступна теорема та наслідок використовують дивергенцію Кульбака-Лейблера між теоретичним розподілом $P(X)$ випадкової величини X та емпіричним розподілом $\hat{P}(X)$ випадкової величини X, як *міру якості навчання однієї вершини байєсівської мережі*.

Вибірка $D=\{X[1],\dots,X[M]\}$ утворена незалежними, однаково розподіленими випадковими величинами, кожна з яких має розподіл $P(X)$.

Теорема 1 [1]. Нехай $P(X)$ – мультиноміальний розподіл випадкової величини X такий, що $P(x)\geq\lambda$ для всіх можливих значень $x\in Val(X)$. Тоді для довільних $\varepsilon>0$, $\delta>0$ справедлива нерівність:

$$P\{D(P(X) \parallel \hat{P}(X)) > \varepsilon\} \leq |Val(X)| exp\left\{-2M\lambda^2\varepsilon^2 \frac{1}{(1+\varepsilon)^2}\right\},$$

де $\hat{P}(X)$ – емпіричний розподіл випадкової величини з параметрами, здобутими методом максимальної правдоподібності.

Наслідок [1]. Нехай виконуються умови теореми 1 і обсяг вибірки M задовольняє нерівності:

$$M \geq \frac{1}{2}\frac{1}{\lambda^2}\frac{(1+\varepsilon)^2}{\varepsilon^2}\ln\frac{|Val(X)|}{\delta}.$$

Тоді

$$P\{D(P(X) \parallel \hat{P}(X)) \leq \varepsilon\} \geq 1-\delta.$$

Наступна теорема та наслідок використовують дивергенцію Кульбака-Лейблера між теоретичним умовним розподілом $P(X_i|Pa_{X_i})$ випадкової величини X_i та емпіричним умовним розподілом $\hat{P}(X_i|Pa_{X_i})$ випадкової величини X_i, як *міру якості навчання байєсівської мережі вцілому*.

Теорема 2 [1]. Нехай $P(X_i|Pa_{X_i})$ – мультиноміальний розподіл випадкової величини X_i такий, що $P(x_i|pa_{X_i}) \geq \lambda$ для всіх можливих значень $x_i \in Val(X_i), pa_{X_i} \in Val(Pa_{X_i}), i = 1 \ldots, n$. Тоді для довільних $\varepsilon > 0, \delta > 0$ справедлива нерівність:

$$P\left\{\sum_{i=1}^{n} D(P(X_i|Pa_{X_i}) \parallel \hat{P}(X_i|Pa_{X_i})) > n\varepsilon\right\} \leq nK^{d+1}exp\left\{-2M\lambda^{2(d+1)}\frac{\varepsilon^2}{(1+\varepsilon)^2}\right\},$$

де $\hat{P}(X_i|Pa_{X_i})$ – емпіричний розподіл випадкової величини X_i з параметрами, здобутими методом максимальної правдоподібності, K – максимальне значення можливих значень випадкової величини X_i, d – максимальне число батьківських вершин в байєсівській мережі.

Наслідок [1]. Нехай виконуються умови теореми 2 і обсяг вибірки M задовольняє нерівності:

$$M \geq \frac{1}{2}\frac{1}{\lambda^{2(d+1)}}\frac{(1+\varepsilon)^2}{\varepsilon^2}\ln\frac{nK^{d+1}}{\delta}.$$

Тоді

$$P\left\{\sum_{i=1}^{n} D(P(X_i|Pa_{X_i}) \parallel \hat{P}(X_i|Pa_{X_i})) < n\varepsilon\right\} > 1-\delta.$$

2. Байєсівське оцінювання параметрів розподілів

2.1. Байєсівський метод статистичного оцінювання

Планування експерименту без апріорної інформації про статистичні властивості випадкової величини, яка спостерігатиметься, відбувається досить рідко. Класичні методи незміщеного оцінювання та оцінювання за методом максимальної правдоподібності не дають способів врахування апріорного знання. Теорія байєсівського оцінювання дозволяє об'єднати апріорну інформацію з величинами, які спостерігаються в експерименті. Основна відмінність між байєсівським і небайєсівським підходами полягає в тому, що байєсівський підхід розглядає параметр розподілу як випадкову величину, в той час як небайєсівський вважає його фіксованою точкою [6].

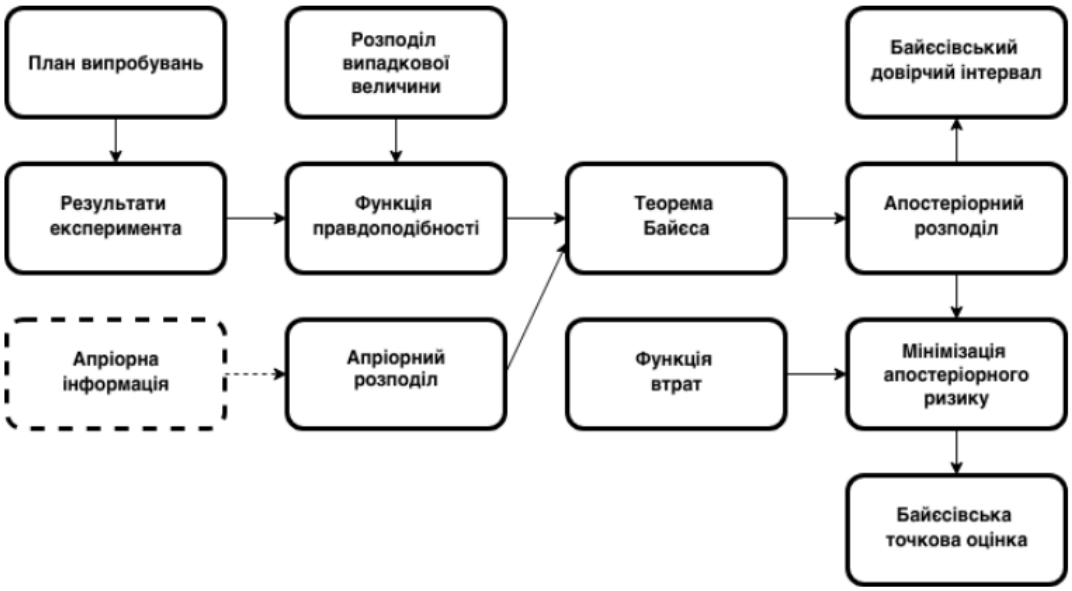


Рис. 2.1.1. Байєсівський метод статистичного оцінювання

Розглянемо схему реалізації байєсівського оцінювання невідомих параметрів, наведену на рис. 2.1.1. Апріорна інформація може бути здобута під час попередніх теоретичних або експериментальних досліджень. Апріорна інформація задається апріорним розподілом вектору параметрів $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)$. Емпірична інформація задається реалізацією вибіркового

вектору $D = \{X[1], \cdots, X[M]\}$. Сформулюємо теорему Байєса для абсолютно неперервних випадкових величин.

Позначимо через $p(D, \boldsymbol{\theta})$ спільну щільність розподілу ймовірностей для вибіркового вектору D і вектору параметрів $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$.

Тоді спільна щільність дорівнює добутку функції максимальної правдоподібності $p(\boldsymbol{\theta} \mid D)$ та апріорної щільності $p(\boldsymbol{\theta})$:

$$p(D, \boldsymbol{\theta}) = p(D|\boldsymbol{\theta})p(\boldsymbol{\theta}), \tag{2.1.1}$$

або добутку апостеріорної щільності $p(\boldsymbol{\theta} \mid D)$ та маргінальної правдоподібності $p(D)$:

$$p(D, \boldsymbol{\theta}) = p(\boldsymbol{\theta}|D)p(D). \tag{2.1.2}$$

Тоді апостеріорна щільність розподілу вектору параметрів $\boldsymbol{\theta}$ за умови заданої вибірки D дорівнює

$$p(\boldsymbol{\theta} \mid D) = \frac{p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(D)}, \tag{2.1.3}$$

де маргінальна правдоподібність визначається

$$p(D) = \int_{\boldsymbol{\theta}} p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}. \tag{2.1.4}$$

Теорему Байєса часто записують у вигляді:

$$p(\boldsymbol{\theta}|D) \sim p(D|\boldsymbol{\theta})p(\boldsymbol{\theta}), \tag{2.1.5}$$

де знак $\sim$ означає пропорційність, $p(\boldsymbol{\theta}|D)$ – апостеріорна щільність розподілу ймовірностей вектору параметрів за умови заданої вибірки D, $p(\boldsymbol{\theta})$ – апріорна щільність розподілу ймовірностей вектору параметрів $\boldsymbol{\theta}$, а $p(D|\boldsymbol{\theta})$, як функція від $\boldsymbol{\theta}$, є функцією максимальної правдоподібності.

Апостеріорна щільність розподілу ймовірностей $p(\boldsymbol{\theta}|D)$ містить як апріорну, так і емпіричну інформацію: апріорна інформація подана апріорною щільністю розподілу ймовірностей, емпірична інформація – функцією максимальної правдоподібності. Апостеріорна щільність розподілу ймовірностей $p(\boldsymbol{\theta}|D)$ використовується в байєсівському аналізі для здобуття точкових та інтервальних оцінок невідомих параметрів [6, 7].

2.2. Апріорний розподіл параметра θ – розподіл Діріхле з гіперпараметрами (1,1)

Нехай X – випадкова величина з розподілом Бернуллі

$$X \sim \begin{pmatrix} x^0 & x^1 \\ 1-\theta & \theta \end{pmatrix}.$$

Нехай $D = \{x[1], x[2], \dots, x[M]\}$ – реалізація випадкової величини X.

Запишемо функцію максимальної правдоподібності розподілу Бернуллі:

$$p(x[1], \cdots, x[M]|\theta) = p(x[1]|\theta) \cdots p(\, x[M]|\theta)$$

$$= \prod_{m=1}^{M} p(x[m]|\theta) = \prod_{m:x[m]=x^0} p(x[m]|\theta) \cdot \prod_{m:x[m]=x^1} p(x[m]|\theta)$$

$$= \prod_{m:x[m]=x^0} (1-\theta) \cdot \prod_{m:x[m]=x^1} \theta = (1-\theta)^{M[x^0]} \theta^{M[x^1]},$$

де $M[x^0]$ – число вибіркових значень таких, що $x[m] = x^0$, $M[x^1]$ – число вибіркових значень таких, що $x[m] = x^1$, при цьому $M[x^0] + M[x^1] = M$.

Нехай апріорна щільність розподілу параметра θ (апріорна інформація) задається щільністю розподілу Діріхле з гіперпараметрами (1, 1) або, що те ж саме, рівномірним розподілом на відрізку [0,1]:

$$p(\theta) = 1 \text{ для } \theta \in [0, 1].$$

Останнє відображає малість апріорного знання щодо невідомого параметра. Згідно з теоремою Байєса апостеріорна щільність розподілу параметра θ дорівнює:

$$p(\theta|x[1], \cdots, x[M]) = \frac{p(x[1], \cdots, x[M]|\theta) p(\theta)}{p(x[1], \cdots, x[M])}$$

$$= \frac{(1-\theta)^{M[x^0]} \theta^{M[x^1]} \cdot 1}{\int_0^1 (1-\theta)^{M[x^0]} \theta^{M[x^1]} \cdot 1 d\theta} = \frac{(1-\theta)^{M[x^0]} \theta^{M[x^1]}}{B(M[x^0]+1, M[x^1]+1)}.$$

Від апріорної щільності – щільності розподілу Діріхле з гіперпараметрами (1, 1) – перейшли до апостеріорної щільності – щільності розподілу Діріхле з гіперпараметрами $(M[x^0]+1, M[x^1]+1)$ [1, 4].

Здобудемо байєсівську точкову оцінку параметра θ. Для цього спрогнозуємо ймовірність того, що наступне вибіркове значення $x[M+1]$ набуватиме значення x^1 за умови відомих значень $x[1], \dots, x[M]$:

$$P(x[M+1] = x^1 | x[1], \cdots, x[M]) =$$

$$= \int_0^1 \theta \frac{(1-\theta)^{M[x^0]} \theta^{M[x^1]}}{B(M[x^0]+1, M[x^1]+1)} d\theta$$

$$= \int_0^1 \frac{(1-\theta)^{M[x^0]} \theta^{M[x^1]+1}}{B(M[x^0]+1, M[x^1]+1)} d\theta$$

$$= \frac{B(M[x^0]+1, M[x^1]+2)}{B(M[x^0]+1, M[x^1]+1)} \int_0^1 \frac{(1-\theta)^{M[x^0]} \theta^{M[1]+1}}{B(M[x^0]+1, M[x^1]+2)} d\theta$$

$$= \frac{\Gamma(M[x^0]+1)\Gamma(M[x^1]+2)}{\Gamma(M[x^0]+1+M[x^1]+2)} \cdot \frac{\Gamma(M[x^0]+1+M[x^1]+1)}{\Gamma(M[x^0]+1)\Gamma(M[x^1]+1)}$$

$$= \frac{(M[x^1]+1)!}{(M[x^0]+M[x^1]+2)!} \cdot \frac{(M[x^0]+M[x^1]+1)!}{(M[x^1])!}$$

$$= \frac{M[x^1]+1}{M[x^0]+M[x^1]+2} = \frac{M[x^1]+1}{M+2}.$$

Отже, байєсівська точкова оцінка параметра θ дорівнює

$$P(x[M+1] = x^1 | x[1], \cdots, x[M]) = \frac{M[x^1]+1}{M+2}.$$

2.3. Апріорний розподіл параметра $\boldsymbol{\theta}$ – розподіл Діріхле з гіперпараметрами $(\boldsymbol{\alpha_1}, \boldsymbol{\alpha_0})$

Нехай апріорна щільність розподілу параметра θ (апріорна інформація) задається щільністю розподілу Діріхле з гіперпараметрами (α_1, α_0), або, що те ж саме, бета-розподілом з параметрами (α_1, α_0):

$$p(\theta) = \frac{1}{B(\alpha_1, \alpha_0)} \theta^{\alpha_1 - 1} (1-\theta)^{\alpha_0 - 1}, \alpha_1 > 0, \alpha_0 > 0,$$

де $B(\alpha_1, \alpha_0)$ – бета-функція з параметрами (α_1, α_0). Відмітимо, що бета-розподіл є спряженим апріорним розподілом до розподілу Бернуллі.

Згідно з теоремою Байєса апостеріорна щільність розподілу параметра θ дорівнює

$$p(\theta|x[1],\cdots,x[M]) = \frac{p(x[1],\cdots,x[M]|\theta)p(\theta)}{p(x[1],\cdots,x[M])}$$

$$= \frac{\theta^{M[x^1]}(1-\theta)^{M[x^0]}\theta^{\alpha_1-1}(1\ -\ \theta)^{\alpha_0-1}}{\int_0^1 \theta^{M[x^1]}(1-\theta)^{M[x^0]}\theta^{\alpha_1-1}(1\ -\ \theta)^{\alpha_0-1}d\theta}$$

$$= \frac{\theta^{\alpha_1+M[x^1]-1}(1-\theta)^{\alpha_0+M[x^0]-1}}{B(\alpha_1+M[x^1],\alpha_0+\ M[x^0])}.$$

Від апріорної щільності – щільності розподілу Діріхле з гіперпараметрами $(\alpha_1;\ \alpha_0)$ – перейшли до апостеріорної щільності – щільності розподілу Діріхле з гіперпараметрами $(\alpha_1 + M[x^1];\ \alpha_0 + M[x^0])$ [1, 4]. Цей результат ілюструє основну властивість розподілу Діріхле: якщо апріорним розподілом є розподіл Діріхле, то апостеріорний розподіл – розподіл Діріхле.

Здобудемо байєсівську точкову оцінку параметра θ. Для цього спрогнозуємо ймовірність того, що наступне вибіркове значення $x[M+1]$ набуватиме значення x^1 за умови відомих вибіркових значень $x[1],\cdots,x[M]$:

$$P(x[M+1]=x^1|x[1],\cdots,x[M]) = \int_0^1 \theta\,\frac{\theta^{\alpha_1+M[x^1]-1}(1-\theta)^{\alpha_0+M[x^0]-1}}{B(\alpha_1+M[x^1],\alpha_0+\ M[x^0])}\,d\theta$$

$$= \int_0^1 \frac{\theta^{\alpha_1+M[x^1]}(1-\theta)^{\alpha_0+M[x^0]-1}}{B(\alpha_1+M[x^1],\alpha_0+\ M[x^0])}\,d\theta$$

$$= \frac{B(\alpha_1+M[x^1]+1,\alpha_0+\ M[x^0])}{B(\alpha_1+M[x^1],\alpha_0+\ M[x^0])}\int_0^1 \frac{\theta^{\alpha_1+M[x^1]}(1-\theta)^{\alpha_0+M[x^0]-1}}{B(\alpha_1+M[x^1]+1,\alpha_0+\ M[x^0])}\,d\theta$$

$$= \frac{\Gamma(\alpha_1+M[x^1]+1)\Gamma(\alpha_0+\ M[x^0])}{\Gamma(\alpha_1+M[x^1]+1+\alpha_0+\ M[x^0])}\cdot\frac{\Gamma(\alpha_1+M[x^1]+\alpha_0+\ M[x^0])}{\Gamma(\alpha_1+M[x^1])\Gamma(\alpha_0+\ M[x^0])}$$

$$= \frac{(\alpha_1+M[x^1])!}{(\alpha_1+\alpha_0+M[x^1]+\ M[x^0])!}\cdot\frac{(\alpha_1+\alpha_0+M[x^1]+\ M[x^0]-1)!}{(\alpha_1+M[x^1]-1)!}$$

$$= \frac{\alpha_1+M[x^1]}{\alpha_1+\alpha_0+M[x^1]+\ M[x^0]} = \frac{\alpha_1+M[x^1]}{\alpha+M}.$$

2.4. Апріорний розподіл параметра $\boldsymbol{\theta}$ – розподіл Діріхле з гіперпараметрами $(\boldsymbol{\alpha_1, \dots, \alpha_K})$

Нехай апріорна щільність параметра $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)$ (апріорна інформація) задається щільністю розподілу Діріхле з гіперпараметрами $(\alpha_1, \dots, \alpha_K)$:

$$p(\theta_1, \dots, \theta_K) = \frac{1}{B(\alpha_1, \dots, \alpha_K)} {\theta_1}^{\alpha_1 - 1} \dots {\theta_K}^{\alpha_K - 1}, \qquad \theta_k \geq 0, \qquad \sum_{k=1}^{K} \theta_k = 1,$$

де $B(\alpha_1, \dots, \alpha_K) = \prod_{i=1}^{K} \Gamma(\alpha_i) / \Gamma\left(\sum_{i=1}^{K} \alpha_i\right)$ – багатовимірна бета-функція. Відмітимо, що розподіл Діріхле є спряженим апріорним розподілом до мультиноміального розподілу.

Згідно з теоремою Байєса апостеріорна щільність параметра $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)$ дорівнює

$$p(\boldsymbol{\theta} | x[1], \dots, x[M]) = \frac{{\theta_1}^{M[1]} \dots {\theta_K}^{M[K]} {\theta_1}^{\alpha_1 - 1} \dots {\theta_K}^{\alpha_K - 1}}{\int {\theta_1}^{M[1]} \dots {\theta_K}^{M[K]} {\theta_1}^{\alpha_1 - 1} \dots {\theta_K}^{\alpha_K - 1} d\boldsymbol{\theta}} =$$

$$= \frac{{\theta_1}^{M[1]} \dots {\theta_K}^{M[K]} {\theta_1}^{\alpha_1 - 1} \dots {\theta_K}^{\alpha_K - 1}}{B(\alpha_1 + M[1], \dots, \alpha_K + M[K]) \int \frac{{\theta_1}^{\alpha_1 + M[1] - 1} \dots {\theta_K}^{\alpha_K + M[K] - 1}}{B(\alpha_1 + M[1], \dots, \alpha_K + M[K])} d\boldsymbol{\theta}} =$$

$$= \frac{{\theta_1}^{M[1]} \dots {\theta_K}^{M[K]} {\theta_1}^{\alpha_1 - 1} \dots {\theta_K}^{\alpha_K - 1}}{B(\alpha_1 + M[1], \dots, \alpha_K + M[K])} = \frac{{\theta_1}^{\alpha_1 + M[1] - 1} \dots {\theta_K}^{\alpha_K + M[K] - 1}}{B(\alpha_1 + M[1], \dots, \alpha_K + M[K])}.$$

Від апріорної щільності – щільності розподілу Діріхле з гіперпараметрами $(\alpha_1, \dots, \alpha_K)$ – перейшли до апостеріорної щільності – щільності розподілу Діріхле з гіперпараметрами $(\alpha_1 + M[1], \dots, \alpha_K + M[K])$ [1, 4].

Здобудемо байєсівські точкові оцінки параметрів $\theta_1, \dots, \theta_K$. Для цього спрогнозуємо ймовірність того, що наступне вибіркове значення набуватиме значення x^k за умови відомих вибіркових значень $x[1], \cdots, x[M]$:

$$P(x[M+1] = x^k \big| x[1], \cdots, x[M]) = \frac{\alpha_k + M[k]}{\alpha + M},$$

де $\alpha = \alpha_1 + \dots + \alpha_K$.

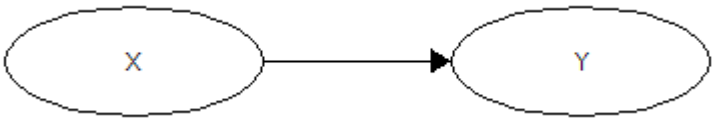
Байєсівські точкові оцінки подібні до оцінок максимальної правдоподібності параметрів $\theta_1, \dots, \theta_K$ мультиноміального розподілу:

$$P(x[M+1] = x^k \big| x[1], \cdots, x[M]) = \frac{M[k]}{M}.$$

Для знаходження оцінок ми можемо скористатися результатами пілотного експерименту, а саме обчислити гіперпараметри α_k як число вибіркових значень, для яких випадкова величина *X* набуває значення x^k у вибірці D'. В цьому випадку байєсівське оцінювання параметрів θ_k еквівалентно оцінювання параметрів θ_k методом максимальної правдоподібності за об'єднанням вибірок $D' \cup D$, при цьому α називають еквівалентним обсягом вибірки [1, 4].

2.5. Байєсівське оцінювання параметрів для байєсівської мережі з двома вершинами

Розглянемо байєсівську мережу:



Імовірнісний розподіл вершини *X* задається:

$$\mathrm{X} \sim \begin{pmatrix} x^0 & x^1 \\ \theta_{x^0} & \theta_{x^1} \end{pmatrix}, \qquad \theta_{x^0} + \theta_{x^1} = 1.$$

Умовний імовірнісний розподіл вершини Y задається:

Y		
X	x^0	x^1
y^0	$\theta_{y^0 \mid x^0}$	$\theta_{y^0 \mid x^1}$
y^1	$\theta_{y^1 \mid x^0}$	$\theta_{y^1 \mid x^1}$

$$\theta_{y^0 \mid x^0} + \theta_{y^1 \mid x^0} = 1,$$

$$\theta_{y^0 \mid x^1} + \theta_{y^1 \mid x^1} = 1.$$

Позначимо через $\boldsymbol{\theta} = \left(\boldsymbol{\theta}_X, \boldsymbol{\theta}_{Y \mid x^0}, \boldsymbol{\theta}_{Y \mid x^1}\right)$ вектор невідомих параметрів, які необхідно оцінити.

Апріорна інформація задається апріорними розподілами параметрів

$\boldsymbol{\theta}_X = (\theta_{x^0}, \theta_{x^1}),\ \boldsymbol{\theta}_{Y|x^0} = (\theta_{y^0|x^0}, \theta_{y^1|x^0}),\ \boldsymbol{\theta}_{Y|x^1} = (\theta_{y^0|x^1}, \theta_{y^1|x^1}).$

Нехай апріорна щільність розподілу параметра $\boldsymbol{\theta}_X = (\theta_{x^0}, \theta_{x^1})$ задається щільністю розподілу Діріхле з гіперпараметрами $(\alpha_{x^0}, \alpha_{x^1})$, $\alpha_{x^0} > 0,\ \alpha_{x^1} > 0$:

$$p(\theta_{x^0}, \theta_{x^1}) = \frac{1}{B(\alpha_{x^0}, \alpha_{x^1})}\, \theta_{x^0}^{\alpha_{x^0}-1} \theta_{x^1}^{\alpha_{x^1}-1}.$$

Апріорна щільність розподілу параметра $\boldsymbol{\theta}_{Y|x^0} = (\theta_{y^0|x^0}, \theta_{y^1|x^0})$ задається щільністю розподілу Діріхле з гіперпараметрами $(\alpha_{y^0|x^0}, \alpha_{y^1|x^0})$, $\alpha_{y^0|x^0} > 0,\ \alpha_{y^1|x^0} > 0$:

$$p\left(\theta_{y^0|x^0}, \theta_{y^1|x^0}\right) = \frac{1}{B(\alpha_{y^0|x^0}, \alpha_{y^1|x^0})} \theta_{y^0|x^0}^{\alpha_{y^0|x^0}-1}\, \theta_{y^1|x^0}^{\alpha_{y^1|x^0}-1}.$$

Апріорна щільність розподілу параметра $\boldsymbol{\theta}_{Y|x^1} = \left(\theta_{y^0|x^1}, \theta_{y^1|x^1}\right)$ задається щільністю розподілу Діріхле з гіперпараметрами $(\alpha_{y^0|x^1}, \alpha_{y^1|x^1})$, $\alpha_{y^0|x^1} > 0,\ \alpha_{y^1|x^1} > 0$:

$$p\left(\theta_{y^0|x^1}, \theta_{y^1|x^1}\right) = \frac{1}{B(\alpha_{y^0|x^1}, \alpha_{y^1|x^1})} \theta_{y^0|x^1}^{\alpha_{y^0|x^1}-1}\, \theta_{y^1|x^1}^{\alpha_{y^1|x^1}-1}.$$

Тоді апріорна щільність розподілу вектору параметрів

$$\boldsymbol{\theta} = \left(\boldsymbol{\theta}_X, \boldsymbol{\theta}_{Y|x^0}, \boldsymbol{\theta}_{Y|x^1}\right)$$

запишеться

$$p(\boldsymbol{\theta}) = p\left(\theta_{x^0}, \theta_{x^1}, \theta_{y^0|x^0}, \theta_{y^1|x^0}, \theta_{y^0|x^1}, \theta_{y^1|x^1}\right)$$

$$= p(\theta_{x^0}, \theta_{x^1}) p\left(\theta_{y^0|x^0}, \theta_{y^1|x^0}\right) p\left(\theta_{y^0|x^1}, \theta_{y^1|x^1}\right)$$

$$= \frac{\theta_{x^0}^{\alpha_{x^0}-1}\, \theta_{x^1}^{\alpha_{x^1}-1}}{B(\alpha_{x^0}, \alpha_{x^1})} \frac{\theta_{y^0|x^0}^{\alpha_{y^0|x^0}-1}\, \theta_{y^1|x^0}^{\alpha_{y^1|x^0}-1}}{B(\alpha_{y^0|x^0}, \alpha_{y^1|x^0})} \frac{\theta_{y^0|x^1}^{\alpha_{y^0|x^1}-1}\, \theta_{y^1|x^1}^{\alpha_{y^1|x^1}-1}}{B(\alpha_{y^0|x^1}, \alpha_{y^1|x^1})}.$$

Запишемо функцію максимальної правдоподібності:

$$L(\boldsymbol{\theta}, D) = P(z[1], z[2], \dots, z[M];\ \boldsymbol{\theta})$$

$$= P(x[1], y[1];\ \boldsymbol{\theta}) P(x[2], y[2];\ \boldsymbol{\theta}) \cdot \dots \cdot P(x[M], y[M];\ \boldsymbol{\theta}).$$

Скористаємось формулою множення для байєсівської мережі:

$$P(X, Y) = P(X) P(Y|X).$$

Тоді

$$L(\boldsymbol{\theta}, D) = P(x[1];\boldsymbol{\theta})P(y[1]|x[1];\boldsymbol{\theta}) \dots P(x[M];\ \boldsymbol{\theta})P(y[M]|x[M];\boldsymbol{\theta})$$

$$= P(x[1];\boldsymbol{\theta}) \dots P(x[M];\boldsymbol{\theta})P(y[1]|x[1];\boldsymbol{\theta}) \dots\ P(y[M]|x[M];\boldsymbol{\theta})$$

$$= \prod_{m=1}^{M} P(x[m];\boldsymbol{\theta}) \cdot \prod_{m=1}^{M} P(y[m]|x[m];\boldsymbol{\theta}).$$

Ми здобули добуток двох локальних функцій максимальної правдоподібності. Розглянемо кожну окремо. Перша локальна функція максимальної правдоподібності – це функція максимальної правдоподібності мультиноміального розподілу з параметрами $(\theta_{x^0}, \theta_{x^1})$:

$$\prod_{m=1}^{M} P(x[m];\boldsymbol{\theta}) = \prod_{m=1}^{M} P(x[m];\boldsymbol{\theta}_X)$$

$$= \prod_{m:x[m]=x^0} P(x[m];\theta_{x^0}) \prod_{m:x[m]=x^1} P(x[m];\theta_{x^1})$$

$$= \prod_{m:x[m]=x^0} \theta_{x^0} \prod_{m:x[m]=x^1} \theta_{x^1} = \theta_{x^0}^{M[x^0]}\, \theta_{x^1}^{M[x^1]},$$

де $M[x^0]$ – число вибіркових значень таких, що $x[m]$ набуває значення x^0, $M[x^1]$ – число вибіркових значень таких, що $x[m]$ набуває значення x^1.

Друга локальна функція максимальної правдоподібності – це добуток функцій максимальної правдоподібності мультиноміальних розподілів з параметрами $(\theta_{y^0|x^0}, \theta_{y^1|x^0})$ та $(\theta_{y^0|x^1}, \theta_{y^1|x^1})$:

$$\prod_{m=1}^{M} P(y[m]|x[m];\boldsymbol{\theta}) = \prod_{m=1}^{M} P\big(y[m]|x[m];\boldsymbol{\theta}_{Y|X}\big)$$

$$= \prod_{m:x[m]=x^0} P\big(y[m]|x[m];\boldsymbol{\theta}_{Y|x^0}\big) \prod_{m:x[m]=x^1} P\big(y[m]|x[m];\boldsymbol{\theta}_{Y|x^1}\big)$$

$$= \prod_{\substack{m:x[m]=x^0,\\ y[m]=y^0}} P\big(y[m]|x[m];\theta_{y^0|x^0}\big) \prod_{\substack{m:x[m]=x^0\\ y[m]=y^1}} P\big(y[m]|x[m];\theta_{y^1|x^0}\big) \cdot$$

$$\cdot \prod_{\substack{m:x[m]=x^1,\\ y[m]=y^0}} P\big(y[m]|x[m];\theta_{y^0|x^1}\big) \prod_{\substack{m:x[m]=x^1\\ y[m]=y^1}} P\big(y[m]|x[m];\theta_{y^1|x^1}\big)$$

$$= \prod_{\substack{m:x[m]=x^0,\\ y[m]=y^0}} \theta_{y^0|x^0} \prod_{\substack{m:x[m]=x^0\\ y[m]=y^1}} \theta_{y^1|x^0} \prod_{\substack{m:x[m]=x^1,\\ y[m]=y^0}} \theta_{y^0|x^1} \prod_{\substack{m:x[m]=x^1\\ y[m]=y^1}} \theta_{y^1|x^1}$$

$$= \theta_{y^0|x^0}^{M[x^0,y^0]} \theta_{y^1|x^0}^{M[x^0,y^1]} \theta_{y^0|x^1}^{M[x^1,y^0]} \theta_{y^1|x^1}^{M[x^1,y^1]},$$

де $M[x^0, y^0]$ – число вибіркових значень, для яких $x[m] = x^0$, $y[m] = y^0$;

$M[x^0, y^1]$ – число вибіркових значень, для яких $x[m] = x^0$, $y[m] = y^1$;

$M[x^1, y^0]$ – число вибіркових значень, для яких $x[m] = x^1$, $y[m] = y^0$;

$M[x^1, y^1]$ – число вибіркових значень, для яких $x[m] = x^1$, $y[m] = y^1$.

Отже, функція максимальної правдоподібності запишеться:

$$L(\boldsymbol{\theta}, D) = \theta_{x^0}^{M[x^0]} \theta_{x^1}^{M[x^1]} \theta_{y^0|x^0}^{M[x^0,y^0]} \theta_{y^1|x^0}^{M[x^0,y^1]} \theta_{y^0|x^1}^{M[x^1,y^0]} \theta_{y^1|x^1}^{M[x^1,y^1]},$$

Згідно з теоремою Байєса апостеріорна щільність розподілу параметра $\boldsymbol{\theta}$ дорівнює:

$$P(\boldsymbol{\theta}|D) = \frac{P(D|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(D)},$$

або, що те ж саме,

$$P(\boldsymbol{\theta}|x[1], \dots, x[M]) = \frac{P(x[1], \dots, x[M]|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(x[1], \dots, x[M])}$$

$$= \frac{\theta_{x^0}^{M[x^0]} \theta_{x^1}^{M[x^1]} \theta_{y^0|x^0}^{M[x^0,y^0]} \theta_{y^1|x^0}^{M[x^0,y^1]} \theta_{y^0|x^1}^{M[x^1,y^0]} \theta_{y^1|x^1}^{M[x^1,y^1]}}{P(x[1], \dots, x[M])} \cdot$$

$$\cdot \frac{\theta_{x^0}^{\alpha_{x^0}-1} \theta_{x^1}^{\alpha_{x^1}-1} \theta_{y^0|x^0}^{\alpha_{y^0|x^0}-1} \theta_{y^1|x^0}^{\alpha_{y^1|x^0}-1} \theta_{y^0|x^1}^{\alpha_{y^0|x^1}-1} \theta_{y^1|x^1}^{\alpha_{y^1|x^1}-1}}{B(\alpha_{x^0}, \alpha_{x^1}) B\left(\alpha_{y^0|x^0}, \alpha_{y^1|x^0}\right) B\left(\alpha_{y^0|x^1}, \alpha_{y^1|x^1}\right)}$$

$$= \frac{\theta_{x^0}^{M[x^0]+\alpha_{x^0}-1} \theta_{x^1}^{M[x^1]+\alpha_{x^1}-1} \theta_{y^0|x^0}^{M[x^0,y^0]+\alpha_{y^0|x^0}-1} \theta_{y^1|x^0}^{M[x^0,y^1]+\alpha_{y^1|x^0}}}{B(\alpha_{x^0} + M[x^0], \alpha_{x^1} + M[x^1]) B\left(\alpha_{y^0|x^0} + M[x^0, y^0], \alpha_{y^1|x^0} + M[x^0, y^1]\right)} \cdot$$

$$\cdot \frac{\theta_{y^0|x^1}^{M[x^1,y^0]+\alpha_{y^0|x^1}-1} \theta_{y^1|x^1}^{M[x^1,y^1]+\alpha_{y^1|x^1}-1}}{B\left(\alpha_{y^0|x^1} + M[x^1, y^0], \alpha_{y^1|x^1} + M[x^1, y^1]\right)}.$$

Від апріорної щільності розподілу параметра $\boldsymbol{\theta}$ поданої як добуток щільностей розподілів Діріхле з гіперпараметрами $(\alpha_{x^0}, \alpha_{x^1})$, $(\alpha_{y^0|x^0}, \alpha_{y^1|x^0})$, $(\alpha_{y^0|x^1}, \alpha_{y^1|x^1})$ перейшли до апостеріорної щільності розподілу параметру $\boldsymbol{\theta}$

поданої як добуток щільностей розподілів Діріхле з гіперпараметрами $(\alpha_{x^0} + M[x^0], \alpha_{x^1} + M[x^1])$, $\left(\alpha_{y^0|x^0} + M[x^0, y^0], \alpha_{y^1|x^0} + M[x^0, y^1]\right)$, $\left(\alpha_{y^0|x^1} + M[x^1, y^0], \alpha_{y^1|x^1} + M[x^1, y^1]\right)$.

Здобудемо байєсівські точкові оцінки параметрів $\theta_{x^1}, \theta_{y^1|x^1}$. Для цього спрогнозуємо ймовірність того, що наступне вибіркове значення $(x[M+1], y[M+1])$ набуватиме значення (x^1, y^1) за умови відомих вибіркових значень $(x[1], y[1]), \dots, (x[M], y[M])$:

$$P(x[M+1] = x^1, y[M+1] = y^1|D)$$

$$= \int_{\boldsymbol{\theta}} P(x[M+1] = x^1, y[M+1] = y^1|D, \boldsymbol{\theta})\, P(\boldsymbol{\theta}|D) d\boldsymbol{\theta}.$$

Згідно з правилом множення для байєсівської мережі

$$P(x[M+1] = x^1, y[M+1] = y^1|D, \boldsymbol{\theta})$$
$$= P(x[M+1] = x^1, y[M+1] = y^1|\boldsymbol{\theta})$$
$$= P\left(x[M+1] = x^1|\boldsymbol{\theta}_X\right), P\left(y[M+1] = y^1 \middle| x[M+1] = x^1, \boldsymbol{\theta}_{Y|X}\right).$$

Апостеріорний розподіл вектору параметрів $\boldsymbol{\theta} = (\boldsymbol{\theta}_X, \boldsymbol{\theta}_{Y|X})$ дорівнює добутку апостеріорних розподілів параметрів $\boldsymbol{\theta}_X$ та $\boldsymbol{\theta}_{Y|X}$:

$$P(\boldsymbol{\theta}|D) = P(\boldsymbol{\theta}_X|D) P\left(\boldsymbol{\theta}_{Y|X} \middle| D\right).$$

Тоді

$$P(x[M+1] = x^1, y[M+1] = y^1|D)$$
$$= \iint P\left(x[M+1] = x^1|\boldsymbol{\theta}_X\right) P\left(y[M+1] = y^1 \middle| x[M+1] = x^1, \boldsymbol{\theta}_{Y|X}\right) \cdot$$
$$\cdot P(\boldsymbol{\theta}_X|D) P\left(\boldsymbol{\theta}_{Y|X} \middle| D\right) d\boldsymbol{\theta}_X d\boldsymbol{\theta}_{Y|X}$$
$$= \left(\int_{\boldsymbol{\theta}_X} P(x[M+1] = x^1|\boldsymbol{\theta}_X) P(\boldsymbol{\theta}_X|D) d\boldsymbol{\theta}_X \right) \cdot$$
$$\cdot \left(\int_{\boldsymbol{\theta}_{Y|X}} P\left(y[M+1] = y^1 \middle| x[M+1] = x^1, \boldsymbol{\theta}_{Y|X}\right) P\left(\boldsymbol{\theta}_{Y|X} \middle| D\right) d\boldsymbol{\theta}_{Y|X} \right)$$
$$= P(x[M+1] = x^1|D) P(y[M+1] = y^1|x[M+1] = x^1, D).$$

Отже, ми можемо розв'язати задачу прогнозування ймовірностей для вершин X та Y окремо. Порахуємо перший інтеграл:

$$P(x[M+1]=x^1|D)=\int\limits_{\boldsymbol{\theta}_X} P(x[M+1]=x^1|\boldsymbol{\theta}_X)P(\boldsymbol{\theta}_X|D)d\boldsymbol{\theta}_X$$

$$=\int\limits_{\theta_{x^0}}\int\limits_{\theta_{x^1}}\theta_{x^1}\frac{\theta_{x^0}^{M[x^0]+\alpha_{x^0}-1}\theta_{x^1}^{M[x^1]+\alpha_{x^1}-1}}{B(\alpha_{x^0}+M[x^0],\alpha_{x^1}+M[x^1])}d\theta_{x^0}d\theta_{x^1}$$

$$=\frac{B(\alpha_{x^0}+M[x^0],\alpha_{x^1}+M[x^1]+1)}{B(\alpha_{x^0}+M[x^0],\alpha_{x^1}+M[x^1])}\cdot$$

$$\cdot\int\limits_{\theta_{x^0}}\int\limits_{\theta_{x^1}}\frac{\theta_{x^0}^{M[x^0]+\alpha_{x^0}-1}\theta_{x^1}^{M[x^1]+\alpha_{x^1}}}{B(\alpha_{x^0}+M[x^0],\alpha_{x^1}+M[x^1]+1)}d\theta_{x^0}d\theta_{x^1}$$

$$=\frac{B(\alpha_{x^0}+M[x^0],\alpha_{x^1}+M[x^1]+1)}{B(\alpha_{x^0}+M[x^0],\alpha_{x^1}+M[x^1])}=$$

$$=\frac{\Gamma(\alpha_{x^0}+M[x^0])\Gamma(\alpha_{x^1}+M[x^1]+1)\Gamma(\alpha_{x^0}+M[x^0]+\alpha_{x^1}+M[x^1])}{\Gamma(\alpha_{x^0}+M[x^0]+\alpha_{x^1}+M[x^1]+1)\Gamma(\alpha_{x^0}+M[x^0])\Gamma(\alpha_{x^1}+M[x^1])}$$

$$=\frac{(\alpha_{x^0}+M[x^0]-1)!\,(\alpha_{x^1}+M[x^1])!\,(\alpha_{x^0}+M[x^0]+\alpha_{x^1}+M[x^1]-1)!}{(\alpha_{x^0}+M[x^0]+\alpha_{x^1}+M[x^1])!\,(\alpha_{x^0}+M[x^0]-1)!\,(\alpha_{x^1}+M[x^1]-1)!}$$

$$=\frac{\alpha_{x^1}+M[x^1]}{\alpha_{x^0}+M[x^0]+\alpha_{x^1}+M[x^1]}=\frac{\alpha_{x^1}+M[x^1]}{\alpha+M}.$$

Порахуємо другий інтеграл:

$$P(y[M+1]=y^1|x[M+1]=x^1,D)=$$

$$=\int\limits_{\boldsymbol{\theta}_{Y|X}}P(y[M+1]=y^1|x[M+1]=x^1,\boldsymbol{\theta}_{Y|X})P\left(\boldsymbol{\theta}_{Y|X}\middle|D\right)d\boldsymbol{\theta}_{Y|X}$$

$$=\int\limits_{y^0|x^0}\int\limits_{y^1|x^0}\frac{\theta_{y^0|x^0}^{M[x^0,y^0]+\alpha_{y^0|x^0}-1}\theta_{y^1|x^0}^{M[x^0,y^1]+\alpha_{y^1|x^0}-1}}{B\left(\alpha_{y^0|x^0}+M[x^0,y^0],\alpha_{y^1|x^0}+M[x^0,y^1]\right)}d\theta_{y^0|x^0}d\theta_{y^1|x^0}\cdot$$

$$\cdot \int_{y^0|x^1} \int_{y^1|x^1} \theta_{y^1|x^1} \frac{\theta_{y^0|x^1}^{M[x^1,y^0]+\alpha_{y^0|x^1}-1} \theta_{y^1|x^1}^{M[x^1,y^1]+\alpha_{y^1|x^1}-1}}{B\left(\alpha_{y^0|x^1}+M[x^1,y^0], \alpha_{y^1|x^1}+M[x^1,y^1]\right)} d\theta_{y^0|x^1} d\theta_{y^1|x^1}$$

$$= \frac{B\left(\alpha_{y^0|x^1}+M[x^1,y^0], \alpha_{y^1|x^1}+M[x^1,y^1]+1\right)}{B\left(\alpha_{y^0|x^1}+M[x^1,y^0], \alpha_{y^1|x^1}+M[x^1,y^1]\right)}$$

$$= \frac{\Gamma\left(\alpha_{y^0|x^1}+M[x^1,y^0]\right)\Gamma(\alpha_{y^1|x^1}+M[x^1,y^1]+1)}{\Gamma\left(\alpha_{y^0|x^1}+M[x^1,y^0]+\alpha_{y^1|x^1}+M[x^1,y^1]+1\right)} \cdot$$

$$\cdot \frac{\Gamma\left(\alpha_{y^0|x^1}+M[x^1,y^0]+\alpha_{y^1|x^1}+M[x^1,y^1]\right)}{\Gamma\left(\alpha_{y^0|x^1}+M[x^1,y^0]\right)\Gamma(\alpha_{y^1|x^1}+M[x^1,y^1])}$$

$$= \frac{\left(\alpha_{y^1|x^1}+M[x^1,y^1]\right)!\left(\alpha_{y^0|x^1}+M[x^1,y^0]+\alpha_{y^1|x^1}+M[x^1,y^1]-1\right)!}{\left(\alpha_{y^0|x^1}+M[x^1,y^0]+\alpha_{y^1|x^1}+M[x^1,y^1]\right)!\left(\alpha_{y^1|x^1}+M[x^1,y^1]-1\right)!}$$

$$= \frac{\alpha_{y^1|x^1}+M[x^1,y^1]}{\alpha_{y^0|x^1}+\alpha_{y^1|x^1}+M[x^1,y^0]+M[x^1,y^1]}.$$

Оцінки параметрів θ_{x^0}, $\theta_{y^0|x^0}$, $\theta_{y^1|x^0}$, $\theta_{y^0|x^1}$ здобуваються аналогічно. Байєсівські точкові оцінки параметрів байєсівської мережі з двома вершинами запишуться:

$$\hat{\theta}_{x^0} = \frac{\alpha_{x^0}+M[x^0]}{\alpha_{x^0}+\alpha_{x^1}+M[x^0]+M[x^1]} = \frac{\alpha_{x^0}+M[x^0]}{\alpha+M},$$

$$\hat{\theta}_{x^1} = \frac{\alpha_{x^1}+M[x^1]}{\alpha_{x^0}+\alpha_{x^1}+M[x^0]+M[x^1]} = \frac{\alpha_{x^1}+M[x^1]}{\alpha+M},$$

$$\hat{\theta}_{y^0|x^0} = \frac{\alpha_{y^0|x^0}+M[x^0,y^0]}{\alpha_{y^0|x^0}+\alpha_{y^1|x^0}+M[x^0,y^0]+M[x^0,y^1]},$$

$$\hat{\theta}_{y^1|x^0} = \frac{\alpha_{y^1|x^0}+M[x^0,y^1]}{\alpha_{y^0|x^0}+\alpha_{y^1|x^0}+M[x^0,y^0]+M[x^0,y^1]},$$

$$\hat{\theta}_{y^0|x^1} = \frac{\alpha_{y^0|x^1}+M[x^1,y^0]}{\alpha_{y^0|x^1}+\alpha_{y^1|x^1}+M[x^1,y^0]+M[x^1,y^1]},$$

$$\hat{\theta}_{y^1|x^1} = \frac{\alpha_{y^1|x^1}+M[x^1,y^1]}{\alpha_{y^0|x^1}+\alpha_{y^1|x^1}+M[x^1,y^0]+M[x^1,y^1]}.$$

2.6. Оцінювання параметрів байєсівської мережі (загальний випадок)

Розглянемо байєсівську мережу $B = (G, P)$ з відомою структурою G та невідомим вектором параметрів $\boldsymbol{\theta}$ умовних імовірнісних розподілів вершин мережі. Необхідно оцінити невідомі параметри $\boldsymbol{\theta}$ згідно з байєсівським методом статистичного оцінювання.

Згідно з теоремою Байєса апостеріорна щільність розподілу вектору параметрів $\boldsymbol{\theta}$ за умови заданої вибірки D дорівнює

$$p\,(\boldsymbol{\theta} \mid D) = \frac{p\,(D \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p\,(D)}, \tag{2.6.1}$$

де $p(D|\boldsymbol{\theta})$ – функція максимальної правдоподібності, $p(\boldsymbol{\theta})$ – апріорна щільність розподілу вектору параметрів $\boldsymbol{\theta}$, $p(D)$ – маргінальна правдоподібність.

Функція максимальної правдоподібності дорівнює спільному розподілу вибіркових значень $D = (\xi[1], \cdots, \xi[M])$:

$$p(D, \boldsymbol{\theta}) = P(\xi[1], \cdots, \xi[M], \boldsymbol{\theta}) = \prod_{m=1}^{M} P(\xi[m], \boldsymbol{\theta}). \tag{2.6.2}$$

Згідно з формулою множення для байєсівської мережі спільний розподіл вершин $X_1, \ldots, X_n$ можна подати як добуток умовних імовірнісних розподілів:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P\left(X_i \mid Parents_{X_i}\right). \tag{2.6.3}$$

Тоді функція максимальної правдоподібності перепишеться у вигляді добутку локальних функцій максимальної правдоподібності (2.6.4):

$$p(D, \boldsymbol{\theta}) = \prod_{m=1}^{M} \prod_{i=1}^{n} P\left(x_i[m] \mid Pa_{X_i}[m];\, \boldsymbol{\theta}\right) = \prod_{i=1}^{n} \prod_{m=1}^{M} P\left(x_i[m] \mid Pa_{X_i}[m];\, \boldsymbol{\theta}\right) =$$

$$= \prod_{i=1}^{n} L_i\left(\boldsymbol{\theta}_{X_i \mid Pa_{X_i}}\right). \tag{2.6.4}$$

Апріорна щільність розподілу вектору параметрів $\boldsymbol{\theta}$ набуває вигляду добутку апріорних щільностей векторів параметрів $\boldsymbol{\theta}_{X_i|\,Pa_{X_i}}$ умовних імовірнісних розподілів вершин мережі:

$$p(\boldsymbol{\theta}) = \prod_{i=1}^{n} p\left(\boldsymbol{\theta}_{X_i|\,Pa_{X_i}}\right). \tag{2.6.5}$$

Отже, апостеріорна щільність розподілу вектору параметрів $\boldsymbol{\theta}$ запишеться у вигляді:

$$p(\boldsymbol{\theta}|D) = \frac{1}{p(D)} \prod_{i=1}^{n} L_i\left(\boldsymbol{\theta}_{X_i|\,Pa_{X_i}}\right) p\left(\boldsymbol{\theta}_{X_i|\,Pa_{X_i}}\right). \tag{2.6.6}$$

Оскільки множини параметрів $\boldsymbol{\theta}_{X_i|\,Pa_{X_i}}$ та $\boldsymbol{\theta}_{X_j|\,Pa_{X_j}}$, $i \neq j$, незалежні, то апостеріорна щільність $p(\boldsymbol{\theta}|D)$ запишеться у вигляді добутку апостеріорних щільностей параметрів $\boldsymbol{\theta}_{X_i|\,Pa_{X_i}}$ умовних імовірнісних розподілів вершин мережі:

$$p(\boldsymbol{\theta}|D) = \prod_{i=1}^{n} p\left(\boldsymbol{\theta}_{X_i|\,Pa_{X_i}}|D\right). \tag{2.6.7}$$

Спрогнозувати ймовірність того, що наступне вибіркове значення $(X_1[M+1], \dots, X_n[M+1])$ набуватиме певних значень можна за формулою:

$$P(X_1[M+1], \dots, X_n[M+1]|D) =$$

$$= \prod_{i=1}^{n} \int_{\boldsymbol{\theta}_{X_i|\,Pa_{X_i}}} P\left(X_i[M+1]|\,Pa_{X_i}[M+1], \boldsymbol{\theta}_{X_i|\,Pa_{X_i}}\right) P\left(\boldsymbol{\theta}_{X_i|\,Pa_{X_i}}|D\right) d\boldsymbol{\theta}_{X_i|\,Pa_{X_i}} \tag{2.6.8}$$

Згідно з (2.6.8) ми можемо знайти умовні ймовірності для кожної вершини мережі X_i окремо, а потім перемножити здобуті результати.

Про задання апріорного розподілу вектору параметрів $\boldsymbol{\theta}$ в байєсівській мережі. Кожна вершина X_i мережі має мультиноміальний розподіл з вектором параметрів $\boldsymbol{\theta}_{X_i|\,Pa_{X_i}}$, який має розподіл Діріхле з гіперпараметрами $\boldsymbol{\alpha}_{X_i|\,pa_{X_i}} = (\alpha_{x_i^1|\,pa_{X_i}}, \dots, \alpha_{x_i^{K_i}|\,pa_{X_i}})$, де K_i – число станів

вершини X_i. Ми можемо скористатися результатами пілотного експерименту, а саме обчислити

$$\alpha_{x_i|\,Pa_{X_i}} = \alpha[x_i, pa_{X_i}],$$

$\alpha[x_i, pa_{X_i}]$ −число спостережень, для яких $X_i = x_i$, $Pa_{X_i} = pa_{X_i}$ у вибірці D'. В цьому випадку байєсівське оцінювання параметрів $\boldsymbol{\theta}_{X_i|\,Pa_{X_i}}$ еквівалентно оцінювання параметрів $\boldsymbol{\theta}_{X_i|\,Pa_{X_i}}$ згідно з методом максимальної правдоподібності за об'єднанням вибірок $D' \cup D$.

Отже, байєсівські точкові оцінки параметрів дорівнюють

$$P(x_i|pa_{X_i}, D) = \frac{\alpha_{x_i|\,pa_{X_i}} + M[x_i, pa_{X_i}]}{\alpha_{pa_{X_i}} + M[pa_{X_i}]}, \quad \alpha_{pa_{X_i}} = \sum_{x_i} \alpha_{x_i|\,pa_{X_i}},$$

де величини $M[x_i,\ pa_{X_i}]$ обчислюються для кожної комбінації значень вершини X_i та значень її батьківських вершин Pa_{X_i}, а суми $M[pa_{X_i}] = \sum_{x_i} M[x_i, pa_{X_i}]$ знаходяться по всіх можливих значеннях вершини X_i.

2.7. Байєсівська мережа Credit

Проведемо дослідження байєсівських оцінок параметрів умовних імовірнісних розподілів вершин байєсівської мережі порівняно з експертними оцінками.

1. Змоделюємо навчальну вибірку даних з експертними оцінками параметрів умовних імовірнісних розподілів вершин мережі та застосуємо байєсівський метод статистичного оцінювання для знаходження байєсівських оцінок. Кожна вершина X_i має мультиноміальний розподіл з вектором параметрів $\boldsymbol{\theta}_{X_i|\,Pa_{X_i}}$, який розподілений згідно з розподілом Діріхле з гіперпараметрами $\boldsymbol{\alpha}_{X_i|\,pa_{X_i}}$. Байєсівські оцінки обчислюються так:

$$P(x_i|pa_{X_i}, D) = \frac{\alpha_{x_i|\,pa_{X_i}} + M[x_i, pa_{X_i}]}{\alpha_{pa_{X_i}} + M[pa_{X_i}]},$$

при цьому гіперпараметри $\alpha_{x_i|\,pa_{X_i}}$ знаходяться за апріорною інформацією, здобутою за результатами пілотного експерименту так:

$$\alpha_{x_i|\,pa_{X_i}} = \alpha \cdot P'(x_i, pa_{X_i}),$$

де α – обсяг вибірки в експерименті, $P'(x_i, pa_{X_i})$ – імовірність набуття вершиною X_i і її батьками Pa_{X_i} станів x_i та pa_{X_i} відповідно.

2. Обчислимо дивергенцію Кульбака-Лейблера між експертними та емпіричними умовними ймовірнісними розподілами вершин мережі.

3. Порівняємо експертні оцінки та байєсівські оцінки параметрів умовних імовірнісних розподілів вершин.

Залежність відстані Кульбака-Лейблера $D(P \parallel \hat{P})$ від обсягу M навчальної вибірки наведена на рис. 2.7.1, 2.7.2.

Неперервною лінією синього кольору зображено відстань Кульбака-Лейблера між експертними та емпіричними умовними ймовірнісними розподілами з параметрами, здобутими *методом максимальної правдоподібності*.

Неперервними лініями червоного та зеленого кольору зображено відстань Кульбака-Лейблера між експертними та емпіричними умовними ймовірнісними розподілами з параметрами, здобутими *байєсівським методом* за умови наявної апріорної інформації для низки вершин: *Income, Assets, Debts to Income Ratio, Age, Payment History, Reliability, Future Income* (див. рис. 2.7.1) і за умови наявної апріорної інформації для всіх вершин: *Income, Assets, Debts to Income Ratio, Age, Payment History, Reliability, Future Income, Credit Worthiness* (див. рис. 2.7.2), при цьому обсяги вибірок в пілотних експериментах $\alpha = 500$ та $\alpha = 1000$ відповідно.

Зі збільшенням обсягу навчальної вибірки вплив апріорної інформації зменшується й байєсівські оцінки параметрів умовних імовірнісних розподілів вершин мережі наближаються до оцінок максимальної правдоподібності.

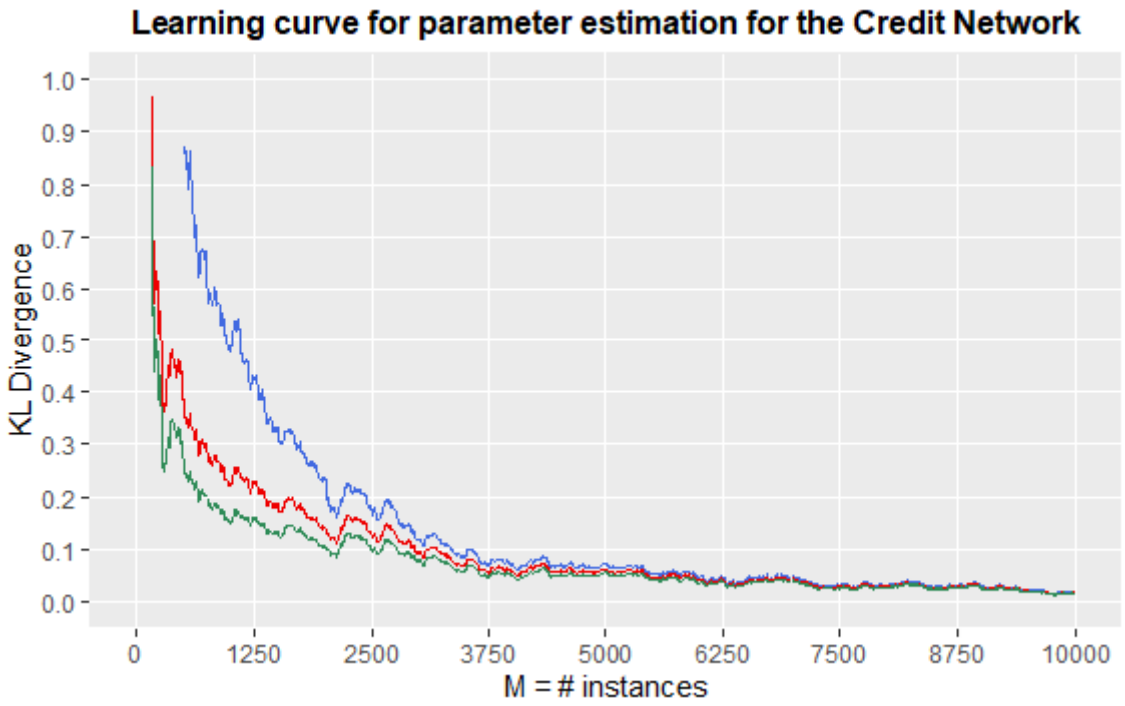


Рис. 2.7.1. Крива навчання байєсівської мережі Credit

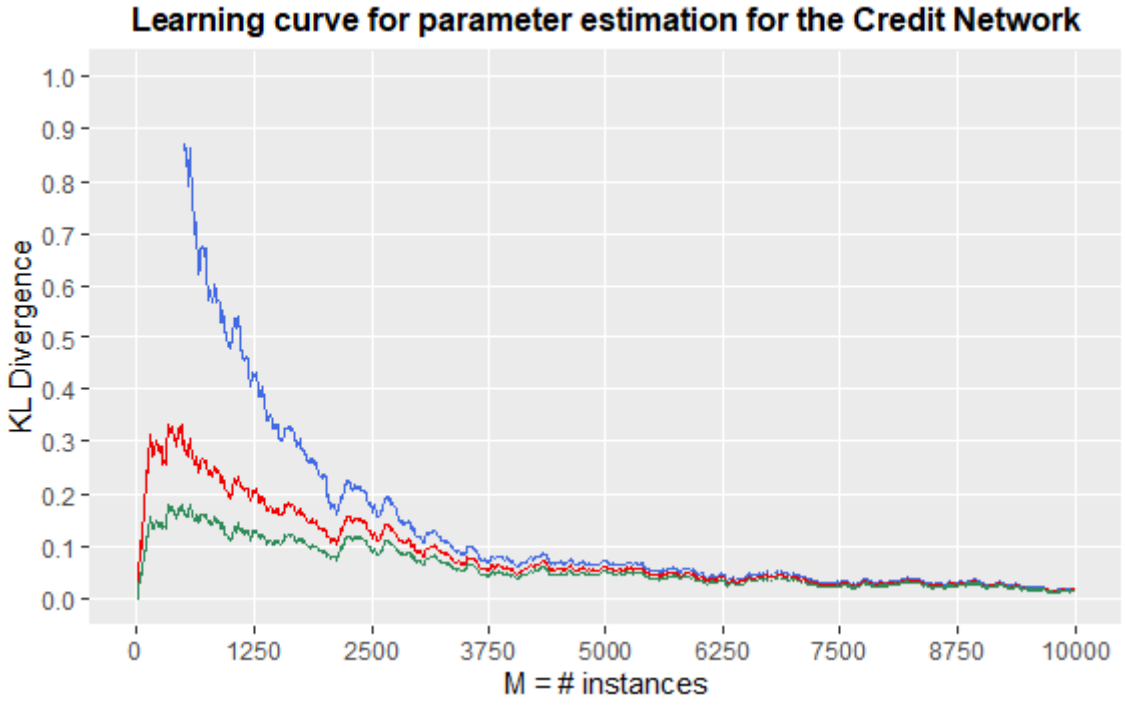


Рис. 2.7.2. Крива навчання байєсівської мережі Credit

Наявність апріорної інформації дозволяє набагато швидше здобути байєсівські оцінки параметрів близькі до експертних оцінок і розпочати застосовувати алгоритми формування ймовірнісного висновку для побудови моделей міркувань на основі байєсівської мережі.

2.8. Якість навчання байєсівської мережі

Для оцінювання якості навчання моделі як функції обсягу вибірки необхідно відповісти на питання: яким повинен бути мінімальний обсяг вибірки для здобуття результатів із заданою точністю ε та надійністю $1-\delta$.

Нехай $D=\{X[1],\dots,X[M]\}$ – вибірка, утворена незалежними, однаково розподіленими випадковими величинами, кожна з яких має мультиноміальний розподіл $P(X)$ з параметрами $\theta_1,\dots,\theta_K$. Щільність параметрів $\theta_1,\dots,\theta_K$ задається щільністю розподілу Діріхле з гіперпараметрами $\alpha_1,\dots,\alpha_K$, а отже, математичне сподівання та дисперсія обчислюються за формулами

$$M\theta_k=\frac{\alpha_k}{\alpha}, D\theta_k=\frac{\alpha_k(\alpha-\alpha_k)}{\alpha^2(\alpha+1)}, k=1,\dots,K,$$

$$\alpha=\alpha_1+\ \dots+\alpha_K,$$

звідки гіперпараметри дорівнюють

$$\alpha_k=\alpha M\theta_k,, k=1,\dots,K.$$

Перепишемо байєсівські точкові оцінки параметрів $\theta_1,\dots,\theta_K$

$$P(x[M+1]=x^k\big|x[1],\cdots,x[M])=\frac{\alpha_k+M[k]}{\alpha+M},$$

у вигляді:

$$P(x[M+1]=x^k\big|x[1],\cdots,x[M])=\frac{\alpha}{\alpha+M}M\theta_k+\frac{M}{\alpha+M}\cdot\frac{M[k]}{M}.$$

Ми здобули зважене середнє апріорного середнього значення та оцінки максимальної правдоподібності, при цьому ваги визначаються величиною еквівалентного обсягу вибірки α та обсягом вибірки M.

Тоді емпіричний розподіл $\tilde{P}(X)$ випадкової величини X з параметрами, здобутими байєсівським методом статистичного оцінювання є зваженим середнім апріорного розподілу Діріхле $P_0(X)$ з гіперпараметрами $\alpha_1,\dots,\alpha_K$ та емпіричного розподілу $\hat{P}(x)$ випадкової величини X з параметрами, здобутими методом максимальної правдоподібності:

$$\tilde{P}(X)=\frac{\alpha}{\alpha+M}P_0(X)+\frac{M}{\alpha+M}\hat{P}(X).$$

Теорема 1 [1]. Нехай $P(X)$ – мультиноміальний розподіл з параметрами $(\theta_1, \dots, \theta_K)$ випадкової величини X такий, що $P(x) \geq \lambda$ для всіх можливих значень $x \in Val(X)$. Тоді для довільних $\varepsilon > 0$, $\delta > 0$ справедлива нерівність:

$$P\{D(P(X) \parallel \tilde{P}(X)) > \varepsilon\}$$
$$\leq |Val(X)| exp\left\{-2M\left(\frac{M}{M+\alpha}\lambda + \frac{\alpha}{M+\alpha}\lambda_0\right)^2 \frac{\varepsilon^2}{(1+\varepsilon)^2}\right\}.$$

Наслідок [1]. Нехай виконуються умови теореми 1 і обсяг вибірки M задовольняє нерівності:

$$M\left(\frac{M}{M+\alpha}\lambda + \frac{\alpha}{M+\alpha}\lambda_0\right)^2 \geq \frac{1}{2}\frac{(1+\varepsilon)^2}{\varepsilon^2}\ln\frac{|Val(X)|}{\delta}.$$

Тоді

$$P\{D(P(X) \parallel \tilde{P}(X)) \leq \varepsilon\} \geq 1 - \delta.$$

Теорема 2 [1]. Нехай $P(X_i|Pa_{X_i})$ – мультиноміальний розподіл з параметрами $(\theta_1, \dots, \theta_K)$ випадкової величини X_i такий, що $P(x_i|pa_{X_i}) \geq \lambda$ для всіх можливих значень $x_i \in Val(X_i), pa_{X_i} \in Val(Pa_{X_i}), i = 1 \dots, n$. Тоді для довільних $\varepsilon > 0$, $\delta > 0$ справедлива нерівність:

$$P\left\{\sum_{i=1}^{n} D(P(X_i|Pa_{X_i}) \parallel \hat{P}(X_i|Pa_{X_i})) > n\varepsilon\right\}$$
$$\leq nK^{d+1} exp\left\{-2M\left(\frac{M}{M+\alpha}\lambda + \frac{\alpha}{M+\alpha}\lambda_0\right)^{2(d+1)} \frac{\varepsilon^2}{(1+\varepsilon)^2}\right\},$$

де K – максимальне значення числа можливих значень випадкових величин X_i, d – максимальне число батьківських вершин в байєсівській мережі.

Наслідок [1]. Нехай виконуються умови теореми 2 і обсяг вибірки M задовольняє нерівності:

$$M\left(\frac{M}{M+\alpha}\lambda + \frac{\alpha}{M+\alpha}\lambda_0\right)^{2(d+1)} \geq \frac{1}{2}\frac{(1+\varepsilon)^2}{\varepsilon^2}\ln\frac{nK^{d+1}}{\delta}.$$

Тоді

$$P\left\{\sum_{i=1}^{n} D(P(X_i|Pa_{X_i}) \parallel \hat{P}(X_i|Pa_{X_i})) < n\varepsilon\right\} > 1 - \delta.$$

Список рекомендованої літератури

1. D. Koller and N. Friedman, Probabilistic Graphical Models: Principles and Techniques. The MIT Press, 2009.
2. D. Koller, Probabilistic Graphical Models 1: Representation // Stanford University. URL: https://www.coursera.org/learn/probabilistic-graphical-models. Online. Accessed 1-Sep-2020.
3. D. Koller, Probabilistic Graphical Models 2: Inference // Stanford University. URL: https://www.coursera.org/learn/probabilistic-graphical-models. Online. Accessed 1-Sep-2020.
4. D. Koller, Probabilistic Graphical Models 3: Learning // Stanford University. URL: https://www.coursera.org/learn/probabilistic-graphical-models. Online. Accessed 1-Sep-2020.
5. Турчин В.Н. Теория вероятностей и математическая статистика. Учебник для студентов высших учебных заведений. – Днепр, Издательство «Лира». – 2018. – 752 с.
6. Бондаренко Я.С. Посібник до вивчення дисципліни “Байєсівський аналіз даних” [Текст]/ Бондаренко Я.С., Кравченко С.В., Сологуб К.М. – Дніпро: Ліра, 2018. – 40 с.
7. Айвазян С.А. Байесовский подход в эконометрическом анализе // Прикладная Эконометрика. – 2008. – №1. – С. 93–130.
8. J. Pearl, Probabilistic Inference in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmenn, 1988.
9. J. Pearl, Causality: Models, Reasoning and Inference. Cambridge University Press, 2000.
10. K.B. Korb, A. E. Nicholson, Bayesian artificial intelligence. Chapman & Hall/CRC Press LLC, 2004.
11. C.M. Bishop, Pattern Recognition and Machine Learning. Springer, 2007.
12. A. Darwiche, Modeling and Reasoning with Bayesian networks. Cambridge University Press, 2009.
13. D. Barber, Bayesian Reasoning and Machine Learning. Cambridge University Press, 2012.
14. K.P. Murphy, Machine Learning: A Probabilistic Perspective. The MIT Press, 2012.
15. SamIam: Sensitivity Analysis, Modeling, Inference and More – програмний продукт для моделювання та аналізу байєсівських мереж, розроблений Automated Reasoning Group prof. Adnan Darwiche in University of California, Los Angeles (UCLA). URL: http://reasoning.cs.ucla.edu/samiam/
16. Бондаренко Я.С. Посібник до вивчення дисципліни «Імовірнісні графічні моделі» [Текст]/ Бондаренко Я.С., Рачко Д.О., Розливан А.О. – Дніпро: Ліра, 2019. – 64 с.
17. Ya.S. Bondarenko, D.O. Rachko, A.O. Rozlyvan Probabilistic Inference in Bayesian Insurance Network // Питання прикладної математики та математичного моделювання. - Д.: Вид-во ДНУ, 2020. - с. 3-20.

Зміст