Machine Learning
Home assignment 3: Unsupervised learning

Solutions are by Olexander Chepurnoi and Yaroslava Lochman.

Problem 2. MAP та ослаблення ваг. (15 балів)

Візьмемо логістичну регресію з сигмоїдою як функцією гіпотези: $h(x) = g(\theta^\top x)$ і навчальну вибірку: $\{(x^{(i)}, y^{(i)}) \,|i = \overline{1,m}\}$ Визначення ваг θ методом максимальної вірогідності (maximum likelihood) виглядає наступним чином:

$$\theta_{ML} = \arg\max_\theta \prod_{i=1}^{m} p(y^{(i)}|x^{(i)}, \theta)$$

Ми хочемо регуляризувати логістичну регресію, вводячи баєсову апріорну імовірність $p(\theta)$.

Нехай ми вибрали апріорний розподіл як гаусів (нормальний) розподіл: $\theta \sim N(0, \tau^2 I)$, де $\tau > 0$, а I – $n+1 \times n+1$ одинична матриця. Тоді MAP метод визначення вагів виглядатиме так:

$$\theta_{MAP} = \arg\max_\theta p(\theta) \prod_{i=1}^{m} p(y^{(i)}|x^{(i)}, \theta)$$

Доведіть, що:

$$\|\theta_{MAP}\|_2 \leq \|\theta_{ML}\|_2$$

Solution to the problem 2. Let's prove from contradiction: $\|\theta_{MAP}\|_2 > \|\theta_{ML}\|_2$
Let's denote:

$$f_1(\theta) = \log\left(\prod_{i=1}^{m} p(y^{(i)}|x^{(i)}, \theta)\right) = \sum_{i=1}^{m} \log p(y^{(i)}|x^{(i)}, \theta)$$

$$f_2(\theta) = \log\left(p(\theta) \prod_{i=1}^{m} p(y^{(i)}|x^{(i)}, \theta)\right) = f_1(\theta) + \log p(\theta)$$

$$= f_1(\theta) + \log\left(\frac{1}{2\pi^{n/2}\,\tau} \exp\left(-\frac{\theta^\top \theta}{2\tau^2}\right)\right) = f_1(\theta) - \log\left(2\pi^{n/2}\,\tau\right) - \frac{\|\theta\|_2^2}{2\tau^2}$$

$$\theta_{ML} = \arg\max_\theta f_1 \qquad \theta_{MAP} = \arg\max_\theta f_2$$

$$f_2(\theta_{MAP}) > f_2(\theta_{ML}) \quad \text{since } \theta_{MAP} = \arg\max_\theta f_2$$

$$\Leftrightarrow f_1(\theta_{MAP}) - \log\left(2\pi^{n/2}\,\tau\right) - \frac{\|\theta_{MAP}\|_2^2}{2\tau^2} > f_1(\theta_{ML}) - \log\left(2\pi^{n/2}\,\tau\right) - \frac{\|\theta_{ML}\|_2^2}{2\tau^2}$$

$$\Leftrightarrow f_1(\theta_{MAP}) - f_1(\theta_{ML}) > \frac{\|\theta_{MAP}\|_2^2 - \|\theta_{ML}\|_2^2}{2\tau^2} > 0 \quad \text{(by supposition)}$$

$$\text{but} \quad f_1(\theta_{MAP}) < f_1(\theta_{ML}) \quad \text{since } \theta_{ML} = \arg\max_\theta f_1$$

We have a contradiction $\Rightarrow \|\theta_{MAP}\|_2 \leq \|\theta_{ML}\|_2$

Problem 3. Упереджені викладачі. (30 балів)

Група студентів, що вивчає курс машинного навчання, в кінці навчання здала P курсових проектів. Курс веде T викладачів. Проекти оцінюються всіми викладачами колективно – кожен з них ставить свою оцінку $x^{(pt)}$.

Ми припускаємо, що кожен проект заслуговує на певну «істину» оцінку μ_p. Кожен викладач, читаючи фінальний звіт, намагається «вгадати» цю істину оцінку. Таким чином, $x^{(pt)}$ – «здогадка» викладача t про те, яким є справжнє значення μ_p.

Проте, викладачі – люди, і людський фактор має свій вплив на оцінку. Дехто з них вважає, що всі проекти хороші, і ставить всім високі бали. Інші можуть бути надто критичними і ставити загалом низькі бали. Також, оцінки разних викладачів можуть мати різну дисперсію, що робить одних більш надійними за інших.

Позначимо ν_t упередження викладача t. Іншими словами, викладач t в середньому оцінює роботи студентва на ν_t балів вище, ніж мав би.

На процес оцінювання робіт впливає величезна кількість випадкових факторів, тому ми будемо моделювати його таким чином.

$$y^{(pt)} \sim N(\mu_p, \sigma_p^2)$$

$$z^{(pt)} \sim N(\nu_t, \tau_t^2)$$

$$x^{(pt)}|y^{(pt)}, z^{(pt)} \sim N(y^{(pt)} + z^{(pt)}, \sigma^2)$$

Змінні $y^{(pt)}$ і $z^{(pt)}$ – незалежні. Змінні x, y, z для різних пар проект-викладач також є незалежними.

Маючи лише виставлені оцінки ($y^{(pt)}$), ми хочемо з'ясувати параметри μ_p , σ_p^2 , ν_t і τ_t^2 . Тоді ми можемо вважати значення μ_p«істинною» кількістю балів, на яку заслуговує курсовий проект.

Ми можемо визначити ці параметри, максимізувавши інтегровану правдоподібність$\{x^{(pt)}|p = 1 \dots P, t = 1 \dots T\}$. Таким чином, модель має латентні змінні $y^{(pt)}$ і $z^{(pt)}$, і проблема максимізації правдоподібності не може бути вирішена у явному вигляді. Тому ми використаємо ітеративний підхід і EM алгоритм. Ваша задача – визначити кроки E і M для цієї моделі.

Ваше рішення для E і M кроків має використовувати тільки такі операції.

Над скалярами: додавання, віднімання, множення, ділення, експонента, логарифм, корінь.

Над векторами і матрицями: інвертування, детермінант.

З метою спрошення задачі, нехай невідомими будуть тільки $\{\mu_p, \sigma_p^2; p = 1 \dots P\}$ і $\{\nu_t, \tau_t^2; t = 1 \dots T\}$. Будемо вважати σ^2 відомою константою.

a. (10 балів) Крок E алгоритму. Спільний розподіл імовірності $p(y^{(pt)}, z^{(pt)}, x^{(pt)})$ має форму спільного багатовимірного нормального розподілу. Виразіть середнє значення та матрицю коваріації цього розподілу через змінні μ_p , σ_p^2 , ν_t, τ_t^2 і σ^2 . Зверніть увагу, що може будти представлений, як $x^{(pt)} = y^{(pt)} + z^{(pt)}$

b. (10 балів) Крок E алгоритму. Виразіть $Q_{pt}(y^{(pt)}, z^{(pt)}) = p(y^{(pt)}, z^{(pt)}|x^{(pt)})$, використовуючи правило залежності від підмножин спільного багатовимірного нормального розподілу.

c. (10 балів) Крок M алгоритму. Сформулюйте крок M алгоритму для оновлення змінних μ_p , σ_p^2 , ν_t, τ_t^2 . Ви можете це зробити через нижню межу правдоподібності з застосуванням математичного очікування $(y^{(pt)}, z^{(pt)})$, взятого з розподілу з густиною $Q_{pt}(y^{(pt)}, z^{(pt)})$.

Solution to the problem 3. a.

$$Ey^{(pt)} = \mu_p \qquad Ez^{(pt)} = \nu_t \qquad Ex^{(pt)} = E\left(y^{(pt)} + z^{(pt)} + \varepsilon\right) = \mu_p + \nu_t$$

$$Dy^{(pt)} = \sigma_p^2 \qquad Dz^{(pt)} = \tau_t^2$$

$$Dx^{(pt)} = D\left(y^{(pt)} + z^{(pt)} + \varepsilon\right)$$

$$= E\left(y^{(pt)2} + z^{(pt)2} + \varepsilon^2 + 2z^{(pt)}y^{(pt)} + 2y^{(pt)}\varepsilon + 2z^{(pt)}\varepsilon\right) - \mu_p^2 - \nu_t^2 - 2\mu_p\nu_t$$

$$= \sigma_p^2 + \mu_p^2 + \tau_t^2 + \nu_t^2 + 2\mu_p\nu_t + \sigma^2 - \mu_p^2 - \nu_t^2 - 2\mu_p\nu_t = \sigma_p^2 + \tau_t^2 + \sigma^2$$

$$cov(y^{(pt)}, z^{(pt)}) = 0$$

$$cov(x^{(pt)}, y^{(pt)}) = Ex^{(pt)}y^{(pt)} - Ex^{(pt)}Ey^{(pt)} = E\left(y^{(pt)2} + y^{(pt)}z^{(pt)} + y^{(pt)}\varepsilon\right) - \mu_p^2 + \nu_t\mu_p =$$

$$= \sigma_p^2 + \mu_p^2 + 0 + \mu_p\nu_t + 0 - \mu_p^2 + \nu_t\mu_p = \sigma_p^2$$

$$cov(x^{(pt)}, z^{(pt)}) = \tau_t^2 \qquad \text{(analogously)}$$

Hence:

$$p(y^{(pt)}, z^{(pt)}, x^{(pt)}) \sim N(\mu, K)$$

where

$$\mu = \begin{pmatrix} \mu_p \\ \nu_t \\ \mu_p + \nu_t \end{pmatrix} \qquad K = \begin{pmatrix} \sigma_p^2 & 0 & \sigma_p^2 \\ 0 & \tau_t^2 & \tau_t^2 \\ \sigma_p^2 & \tau_t^2 & \sigma_p^2 + \tau_t^2 + \sigma^2 \end{pmatrix}$$