Семинар 4.
ММП, осень 2012–2013
9 октября

Темы семинара:

- смеси нормальных распределений;
- скрытые переменные, совместное распределение скрытых и наблюдаемых переменных;
- KL-дивергенция, свойства;
- EM-алгоритм в общем случае.

1 Введение

Напомним вкратце часть материала, рассмотренного на лекции. Мы рассматриваем случайную величину, описываемую смесью нормальных распределений:

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_{i=1}^{K} \pi_i \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_i, \Sigma_i).$$

При этом мы хотим найти параметры $\boldsymbol{\theta} = \{\boldsymbol{\pi}_i, \boldsymbol{\mu}_i, \Sigma_i\}$, максимизируя правдоподобие наблюдаемой выборки:

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \sum_{i=1}^{\ell} \ln\Big[\textstyle\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \Sigma_j)\Big] \to \max_{\boldsymbol{\theta}}.$$

В отличие от случая нормально распределенных случайных величин, где логарифмирование плотности позволяло выписать решение задачи в аналитическом виде, здесь сумма под логарифмом кардинально меняет картину. Лекция была посвящена EM алгоритму, который (в частности) позволяет решить эту проблему, вводя в модель *скрытые переменные* $\boldsymbol{z} \in \{0,1\}^K,\ \sum_{i=1}^K \boldsymbol{z}_i = 1$, отвечающие за компоненту смеси, породившую наблюдаемую случайную величину.

Рассмотрим еще раз роль скрытых переменных в описанной процедуре.

Задача 1. *Рассмотрим совместное распределение* $p(\boldsymbol{x}, \boldsymbol{z})$ *на* $\boldsymbol{x} \in \mathbb{R}^n$ *и* $\boldsymbol{z} \in \{0,1\}^K,\ \sum_{i=1}^K \boldsymbol{z}_i = 1$, *такое что выполнены следующие тождества для маргинального распределения* $\boldsymbol{z}$

$$p(\boldsymbol{z}) = \prod_{i=1}^{K} \pi_i^{\boldsymbol{z}_i}, \quad 0 \leqslant \pi_i \leqslant 1,\ \textstyle\sum_{i=1}^K \pi_i = 1,$$

и условного распределения

$$p(\boldsymbol{x}|\boldsymbol{z}) = \prod_{i=1}^{K} \left[\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_i, \Sigma_i)\right]^{Z_i}, \quad \mu_i \in \mathbb{R}^n, \ \Sigma_i \in \mathbb{R}^{n\times n}.$$

Найдите маргинальное распределение переменной $\boldsymbol{x}$.

Итак, нам удалось получить смесь нормальных распределений в качестве маргинального распределения, вводя в модель скрытые переменные. Что же мы при этом получаем?

Алгоритм начинается с инициализации вектора параметров $\boldsymbol{\theta} \equiv \boldsymbol{\theta}_0$ и заключается в последующем применении двух шагов:

Е-шаг: вычисление апостериорных вероятностей $p(\boldsymbol{z}_k = 1|\boldsymbol{x})$ с помощью формулы Байеса:

$$p(\boldsymbol{z}_k = 1|\boldsymbol{x}) = \frac{p(\boldsymbol{z}_k = 1)p(\boldsymbol{x}|z_k = 1)}{p(\boldsymbol{x})} = \frac{\pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \Sigma_k)}{\sum_{i=1}^{K} \pi_i \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_i, \Sigma_i)};$$

М-шаг: решение задачи максимизация правдоподобия наблюдаемых переменных $\ln p(\boldsymbol{X}|\boldsymbol{\theta}) \to \max_{\boldsymbol{\theta}}$ и выражение параметров модели ($\boldsymbol{\theta}$) через апостериорные вероятности $g_{i,k} \equiv p(\boldsymbol{z}_k = 1|\boldsymbol{x}_i)$:

$$\frac{\partial \ln(\boldsymbol{X}|\boldsymbol{\theta})}{\partial \boldsymbol{\pi}_k} = 0 \Leftrightarrow \boldsymbol{\pi}_k = \frac{\sum_{j=1}^{\ell} g_{j,k}}{\ell};$$

$$\frac{\partial \ln(\boldsymbol{X}|\boldsymbol{\theta})}{\partial \boldsymbol{\mu}_k} = 0 \Leftrightarrow \boldsymbol{\mu}_k = \frac{\sum_{j=1}^{\ell} g_{j,k}\boldsymbol{x}_j}{\sum_{i=1}^{\ell} g_{i,k}};$$

$$\frac{\partial \ln(\boldsymbol{X}|\boldsymbol{\theta})}{\partial \Sigma_k} = 0 \Leftrightarrow \Sigma_k = \frac{\sum_{j=1}^{\ell} g_{j,k}(x_j - \boldsymbol{\mu}_k)(x_j j - \boldsymbol{\mu}_k)^{\top}}{\sum_{i=1}^{\ell} g_{i,k}}.$$

Описанная процедура не дает нам аналитического решения интересующей нас задачи (поскольку апостериорные вероятности $g_{i,k}$ зависят от параметров модели $\boldsymbol{\theta}$), но позволяет решать её эффективно с помощью итерационной процедуры.

2 ЕМ-алгоритм в общем случае

Теперь введем ЕМ алгоритм, как общую процедуру максимизации правдоподобия в моделях со скрытыми переменными. Пусть имеется модель, в которую входят наблюдаемые переменные X и скрытые переменные Z с совместным распределением $p(X, Z|\boldsymbol{\theta})$, где $\boldsymbol{\theta}$ — параметры модели. Мы будем решать задачу

$$\ln p(X|\boldsymbol{\theta}) \to \max_{\boldsymbol{\theta}}.$$

Заметим, что справедлива следующая цепочка неравенств:

$$\ln p(X|\boldsymbol{\theta}) = \int q(Z) \ln p(X|\boldsymbol{\theta}) dZ = \int q(Z) \ln \frac{p(X, Z|\boldsymbol{\theta})}{p(Z|X, \boldsymbol{\theta})} dZ = \int q(Z) \ln \left[\frac{p(X, Z|\boldsymbol{\theta})}{q(Z)} \frac{q(Z)}{p(Z|X, \boldsymbol{\theta})}\right] dZ =$$
$$= \underbrace{\int q(Z) \ln p(X, Z|\boldsymbol{\theta}) dZ - \int q(Z) \ln q(Z) dZ}_{\mathcal{L}(q,\boldsymbol{\theta})} + \underbrace{\int q(Z) \ln \frac{q(Z)}{p(Z|X, \boldsymbol{\theta})} dZ}_{KL\left(q\|p(Z|X,\boldsymbol{\theta})\right)}, \qquad (1)$$

где $KL\big(q\|p(Z|X,\boldsymbol{\theta})\big)$ — *дивергенция Кульбака-Лейблера* между распределениями q и апостериорным распределением $p(Z|X,\boldsymbol{\theta})$.

Задача 2. *Докажите, что*

а) $KL(p\|q) \geqslant 0$*;*

Решение*:* $-KL(p\|q) = \displaystyle\int q(x)\ln\frac{p(x)}{q(x)}dx \leqslant \ln\int q(x)\frac{p(x)}{q(x)}dx = 0$ *(неравенство Йенсена).*

б) в общем случае $KL(p\|q) = 0$ *тогда и только тогда, когда* $p \equiv q$.

Задача 3. *Покажите, что максимизация функции правдоподобия* $p(X|\boldsymbol{\theta})$ *эквивалентна минимизации KL-дивергенции* $\mathrm{KL}\big(\text{эмпирическое распределение}\|p(x|\boldsymbol{\theta})\big)$.

Таким образом, мы нашли нижнюю оценку интересующей нас функции правдоподобия:

$$\ln p(X|\boldsymbol{\theta}) \geqslant \mathcal{L}(q,\boldsymbol{\theta}) = \int q(Z)\ln p(X,Z|\boldsymbol{\theta})dZ - \int q(Z)\ln q(Z)dZ. \qquad (2)$$

Сейчас мы опишем EM-алгоритм с помощью неравенств (1) и (2), попутно показав, что эта процедура сходится к локальному максимуму функции правдоподобия $p(X|\boldsymbol{\theta})$.

Фиксируем некоторое начальное значение $\boldsymbol{\theta} = \boldsymbol{\theta}^0$. **(Е-шаг:)** Попробуем максимизировать нижнюю оценку (2) $\mathcal{L}(q,\boldsymbol{\theta})$ по распределению q при фиксированном $\boldsymbol{\theta} = \boldsymbol{\theta}^0$:

$$\mathcal{L}(q,\boldsymbol{\theta}^0) \to \max_q.$$

Задача 4. *Найдите распределение q, максимизирующее оценку (2) $\mathcal{L}(q,\boldsymbol{\theta})$ при фиксированном значении $\boldsymbol{\theta} = \boldsymbol{\theta}^0$.*

Решение: *поскольку левая часть (1) не зависит от q, то несложно сделать вывод, что оценка максимизируется при $q(Z) = p(Z|X,\boldsymbol{\theta}^0)$.*

В этом случае KL-дивергенция обнуляется и нижняя оценка $\mathcal{L}(q,\boldsymbol{\theta})$ превращается в точную оценку логарифма функции правдоподобия $\ln p(X|\boldsymbol{\theta})$ в точке $\boldsymbol{\theta} = \boldsymbol{\theta}^0$.

(М-шаг:) Теперь при фиксированном распределении q, найденном на Е-шаге, максимизируем нижнюю оценку $\mathcal{L}(q,\boldsymbol{\theta})$ по параметрам модели $\boldsymbol{\theta}$, что эквивалентно:

$$\int q(Z)\ln p(X,Z|\boldsymbol{\theta})dZ \to \max_{\boldsymbol{\theta}}, \qquad (3)$$

поскольку q зависит лишь от первичной инициализации вектора параметров. Решение этой задачи назовем $\boldsymbol{\theta}^{\text{new}}$ В левой части (3) записано ни что иное, как матожидание логарифма правдоподобия всех переменных $\ln p(X,Z|\boldsymbol{\theta})$ (скрытых и наблюдаемых) относительно апостериорной вероятности скрытых переменных $q(Z) = p(Z|X,\boldsymbol{\theta}^0)$.

Заметим, что М-шаг непременно увеличит значение нижней оценки $\mathcal{L}(q,\boldsymbol{\theta})$, а значит и правдоподобия $p(X|\boldsymbol{\theta})$, благодаря особому выбору распределения q на Е-шаге (сделавшего нижнюю оценку точной). Поскольку правдоподобие наблюдаемых данных ограничено сверху и в результате описанных Е–М итераций монотонно возрастает, то описанная процедура непременно сойдется (правда всего лишь к локальному максимуму).

ВСТАВИТЬ КАРТИНКУ, ИЛЛЮСТРИРУЮЩУЮ ИТЕРАЦИИ.

Задача 5. *(задачка к картинке) Докажите, что производные нижней оценки (2) (для* $q(Z) = p(Z|X, \boldsymbol{\theta}^0)$*) и логарифма правдоподобия* $\ln p(X|\boldsymbol{\theta})$ *в точке* $\boldsymbol{\theta} = \boldsymbol{\theta}^0$ *совпадают.*

Решение: *поскольку* $KL(q\|p)$ *зависит от* $\boldsymbol{\theta}$ *и достигает своего минимума (значения 0) в точке* $\boldsymbol{\theta} = \boldsymbol{\theta}^0$*, то*

$$\frac{\partial}{\partial \boldsymbol{\theta}} KL(q\|p) = 0 \text{ в точке } \boldsymbol{\theta} = \boldsymbol{\theta}^0.$$

Таким образом, продифференцировав обе части неравенства (1) по вектору параметров и подставив в полученное тождество $\boldsymbol{\theta} = \boldsymbol{\theta}^0$*, мы получим искомое утверждение.*

Список литературы

[1] *Д. П. Ветров, Д. А. Кропотов.* Байесовские методы машинного обучения. — Курс лекций, ММП ВМиК МГУ.
www.machinelearning.ru

[2] *C. Bishop.* Pattern Recognition and Machine Learning. 9-я глава. — Springer, 2006.