

Семинар 4.
ММП, осень 2012–2013
9 октября

Темы семинара:

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

1 Введение

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

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

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

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

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* $\mathbf{z} \in \{0, 1\}^K$, $\sum_{i=1}^K z_i = 1$, отвечающие за компоненту смеси, породившую наблюдаемую случайную величину.

Рассмотрим еще раз роль скрытых переменных в описанной процедуре.

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

$$p(\mathbf{z}) = \prod_{i=1}^K \pi_i^{z_i}, \quad 0 \leq \pi_i \leq 1, \quad \sum_{i=1}^K \pi_i = 1,$$

и условного распределения

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

Найдите маргинальное распределение переменной \mathbf{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 $p(\mathbf{z}_k = 1|\mathbf{x})$ с помощью формулы Байеса:

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

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

$$\begin{aligned} \frac{\partial \ln(\mathbf{X}|\boldsymbol{\theta})}{\partial \boldsymbol{\pi}_k} = 0 &\Leftrightarrow \boldsymbol{\pi}_k = \frac{\sum_{j=1}^{\ell} g_{j,k}}{\ell}; \\ \frac{\partial \ln(\mathbf{X}|\boldsymbol{\theta})}{\partial \boldsymbol{\mu}_k} = 0 &\Leftrightarrow \boldsymbol{\mu}_k = \frac{\sum_{j=1}^{\ell} g_{j,k} \mathbf{x}_j}{\sum_{i=1}^{\ell} g_{i,k}}; \\ \frac{\partial \ln(\mathbf{X}|\boldsymbol{\theta})}{\partial \Sigma_k} = 0 &\Leftrightarrow \Sigma_k = \frac{\sum_{j=1}^{\ell} g_{j,k} (\mathbf{x}_j - \boldsymbol{\mu}_k)(\mathbf{x}_j - \boldsymbol{\mu}_k)^\top}{\sum_{i=1}^{\ell} g_{i,k}}. \end{aligned}$$

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 EM-алгоритм в общем случае

Теперь введем EM 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

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

Заметим, что справедлива следующая цепочка неравенств:

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

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

Задача 2. Докажите, что

а) $KL(p||q) \geq 0$;

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

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

Задача 3. обязательно покажи, что принцип максимума правдоподобия $\Rightarrow \min KL$ (эмпирическое распр||модельное распр) Покажите, что максимизация функции правдоподобия $p(X|\theta)$ эквивалентна минимизации KL -дивергенции $KL(\text{эмпирическое распределение}||p(x|\theta))$.

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

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

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

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

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

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

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

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

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

$$\int q(Z)p(X, Z|\theta)dZ \rightarrow \max_{\theta}, \quad (3)$$

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

Заметим, что М-шаг непременно увеличит значение нижней оценки $\mathcal{L}(q, \theta)$, а значит и правдоподобия $p(X|\theta)$, 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. (задача к картинке) Докажите, что производные нижней оценки (2) (для $q(Z) = p(Z|X, \theta^0)$) и логарифма правдоподобия $\ln p(X|\theta)$ в точке $\theta = \theta^0$ совпадают.

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

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

Таким образом, продифференцировав обе части неравенства (1) по вектору параметров и подставив в полученное тождество $\theta = \theta^0$, 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] С. Bishop. Pattern Recognition and Machine Learning. 9-я глава. — Springer, 2006.