

Семинар 5.
ММП, осень 2012–2013
16 октября

Темы семинара:

- Разбор домашки: EM-алгоритм;
- Проклятие размерности;
- Непараметрическая оценка плотности: связь парзеновского окна с алгоритмом ближайших соседей.

1 Разбор домашнего задания по EM-алгоритму

1.1 Задача 2

Пусть дана вероятностная модель с наблюдаемыми переменными $X = \{\mathbf{x}_1, \dots, \mathbf{x}_\ell\}$, $\mathbf{x}_i \in \mathbb{R}^n$, соответствующими им скрытыми переменными $Z = \{z_1, \dots, z_\ell\}$ и вектором параметров $\boldsymbol{\theta} = \{\boldsymbol{\mu}_k, \Sigma_k, \pi_k, k = 1, \dots, K\}$, рассмотренная на 4-ом семинаре. Скрытые переменные на этот раз будут принимать скалярные значения из множества $\{1, \dots, K\}$ (а не бинарные вектора с одной единицей, как на семинаре). По-прежнему,

$$p(z = k) = \pi_k \geq 0, \quad k = 1, \dots, K, \quad \sum_k \pi_k = 1;$$
$$p(\mathbf{x}|z = k) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k), \quad \mathbf{x} \in \mathbb{R}^n.$$

Убедитесь, что EM-алгоритм для смеси распределений, изложенный на лекции, получается применением общего EM-алгоритма для максимизации правдоподобия наблюдаемых переменных в этой модели. (Аккуратно выпишите E и M шаги общего EM-алгоритма).

Решение: Мы начинаем с фиксации начальных значений вектора параметров $\boldsymbol{\theta} = \boldsymbol{\theta}^0$. На E-шаге, как мы помним, мы берем распределение $q(Z)$ на скрытых переменных равным апостериорному распределению $p(Z|X, \boldsymbol{\theta} = \boldsymbol{\theta}^0)$, что максимизирует нижнюю оценку логарифма правдоподобия наблюдаемых данных (одновременно делая ее точной):

$$q(Z) = p(Z|X, \boldsymbol{\theta}^0) = \frac{p(X, Z|\boldsymbol{\theta}^0)}{\sum_Z p(X, Z|\boldsymbol{\theta}^0)} = \frac{\prod_{i=1}^N p(x_i, z_i|\boldsymbol{\theta}^0)}{\sum_Z \prod_{i=1}^N p(x_i, z_i|\boldsymbol{\theta}^0)} = \prod_{i=1}^N \frac{p(x_i, z_i|\boldsymbol{\theta}^0)}{\sum_{z_i} p(x_i, z_i|\boldsymbol{\theta}^0)} = \prod_{i=1}^N p(z_i|x_i, \boldsymbol{\theta}^0). \quad (1)$$

При этом

$$q(z_i = k) = p(z_i = k|x_i, \boldsymbol{\theta}) = \frac{\pi_k \mathcal{N}(x_i|\boldsymbol{\mu}_k, \Sigma_k)}{\sum_{j=1}^K \mathcal{N}(x_i|\boldsymbol{\mu}_j, \Sigma_j)}. \quad (2)$$

Кроме того

$$p(X, Z|\boldsymbol{\theta}^0) = \prod_{i=1}^N p(x_i, z_i|\boldsymbol{\theta}^0) = \prod_{i=1}^{\ell} \prod_{k=1}^K \{\pi_k^0 \mathcal{N}(x_i|\boldsymbol{\mu}_k^0, \Sigma_k^0)\}^{[z_i=k]}$$

и

$$\ln p(X, Z|\boldsymbol{\theta}^0) = \sum_{i=1}^{\ell} \sum_{k=1}^K [z_i = k] (\ln \pi_k^0 + \ln \mathcal{N}(x_i|\boldsymbol{\mu}_k^0, \Sigma_k^0)). \quad (3)$$

M-шаг заключается в максимизации мат.ожидания выражения (3) по распределению $q(Z)$, задаваемому выражением (1). Для этого выпишем мат.ожидание в явном виде:

$$\begin{aligned} \mathbb{E}_q \ln p(X, Z|\boldsymbol{\theta}) &= \sum_{i=1}^{\ell} \sum_{k=1}^K \mathbb{E}_q [z_i = k] (\ln \pi_k + \ln \mathcal{N}(x_i|\boldsymbol{\mu}_k, \Sigma_k)) = \\ &= \sum_{i=1}^{\ell} \sum_{k=1}^K q(z_i = k) (\ln \pi_k + \ln \mathcal{N}(x_i|\boldsymbol{\mu}_k, \Sigma_k)), \end{aligned} \quad (3)$$

где мы воспользовались очевидным фактом равенства мат.ожидания индикатора события вероятности этого события по рассматриваемой мере.

Подстановка (2) в (3) дает нам следующую оптимизационную задачу:

$$\begin{aligned} \mathbb{E}_q \ln p(X, Z|\boldsymbol{\theta}) &= \sum_{i=1}^{\ell} \sum_{k=1}^K g_{i,k} (\ln \pi_k + \ln \mathcal{N}(x_i|\boldsymbol{\mu}_k, \Sigma_k)) \rightarrow \max_{\boldsymbol{\theta}}; \\ \sum_{i=1}^K \pi_i &= 1, \quad \pi_i \geq 0, \quad i = 1, \dots, K. \end{aligned}$$

Эту задачу можно решать, используя метод множителей Лагранжа, как рассказывалось на лекции. Ее решение мы обозначим $\boldsymbol{\theta}^1$ и снова переходим на E-шаг.

1.2 Задача 3

Предположим, что в общем EM-алгоритме из прошлой задачи на E-шаге мы минимизируем дивергенцию $KL(q||p)$ по семейству вырожденных распределений q (распределение вырождено, если оно всю вероятностную меру сосредотачивает на одной точке). Также предположим, что ковариационные матрицы всех компонент совпадают и равны единичной. Как будут выглядеть в этом случае EM итерации?

Несложно заметить, что минимизация дивергенции по вырожденным распределениям $q(Z)$ равносильна выбору значения скрытых переменных Z , максимизирующего апостериорную вероятность $p(Z|X, \boldsymbol{\theta})$:

$$\begin{aligned} \arg \min_{q: q(Z_0)=1} KL(q(Z)||p(Z|X, \boldsymbol{\theta})) &= \arg \min_{q: q(Z_0)=1} \sum_Z q(Z) \ln \frac{q(Z)}{p(Z|X, \boldsymbol{\theta})} = \\ &= \arg \min_{q: q(Z_0)=1} -\ln p(Z_0|X, \boldsymbol{\theta}) = \arg \max_{q: q(Z_0)=1} \ln p(Z_0|X, \boldsymbol{\theta}). \end{aligned}$$

Таким образом, распределение q на E-шаге целиком сосредотачивается на значении вектора скрытых переменных $Z = Z_0$, максимизирующего апостериорную вероятность $p(Z|X, \theta)$. Заметим, что при этом нижняя оценка $\mathcal{L}(q, \theta)$ не становится точной при $\theta = \theta^0$. (Когда она будет точной? Когда сами апостериорные распределения являются вырожденными). Таким образом, увеличение нижней оценки на M-шаге не гарантирует увеличения правдоподобия наблюдаемых переменных.

M-шаг принимает простой вид, поскольку мат.ожидание логарифма совместного правдоподобия превращается в его значение в точке $Z = Z_0 = \{z_1^0, \dots, z_\ell^0\}$:

$$E_q \ln p(X, Z|\theta) = \ln p(X, Z = Z_0|\theta) = \prod_{i=1}^{\ell} \pi_{z_i^0} \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_{z_i^0}, I) \rightarrow \max_{\{\boldsymbol{\mu}_k, \pi_k\}}$$

Метод множителей Лагранжа дает нам следующее решение:

$$\pi_k = \frac{\sum_{i=1}^{\ell} [z_i^0 = k]}{\ell};$$

$$\boldsymbol{\mu}_k = \frac{\sum_{i=1}^{\ell} [z_i^0 = k] \mathbf{x}_i}{\sum_{i=1}^{\ell} [z_i^0 = k]}.$$

2 Вторая половина

На следующем примере изучим влияние размерности пространства объектов в случаях, когда мы используем непараметрические методы оценки плотностей распределений.

Задача. Представим, что N точек обучающей выборки равномерно раскиданы по шару единичного радиуса с центром в начале координат в p -мерном пространстве. Найдите медиану распределения расстояний от центра координат до ближайшей точки обучающей выборки.

Решение:

$$P(\min_i \rho(\mathbf{0}, \mathbf{x}_i) \leq d) = 1 - P(\forall i : \rho(\mathbf{0}, \mathbf{x}_i) > d) = 1 - \prod_{i=1}^N P(\rho(\mathbf{0}, \mathbf{x}_i) > d) = 1 - \left(1 - \frac{4/3\pi d^p}{4/3\pi}\right)^N.$$

Мы интересуемся решением уравнения

$$P(\min_i \rho(\mathbf{0}, \mathbf{x}_i) \leq d) = 1/2$$

относительно d . Получаем

$$1/2 = 1 - (1 - d^p)^N$$

$$d = \left(1 - \frac{1}{2}\right)^{1/N}.$$

Оказывается, при $N = 500$, $p = 10$ $d \approx 0.52$. Видно, что с ростом размерности точки сосредотачиваются у границ нашего пространства объектов, что ведет к сложностям при попытках оценивать плотность с помощью метода ближайших соседей или парзеновского окна. Этот эффект принято называть «проклятием размерности».

Разберем еще один очень полезный факт, описывающий связь метода парзеновского окна с алгоритмом ближайших соседей. Предположим, что наша выборка X^ℓ вытянута из распределения с плотностью $p(x)$ на \mathbb{R}^d . Возьмем небольшую окрестность \mathcal{R} вокруг фиксированной точки \mathbf{x} , в которой мы хотим оценить плотность. Вероятность попасть в эту окрестность равна

$$P = \int_{\mathcal{R}} p(\mathbf{x}) d\mathbf{x}.$$

Поскольку точки выборки вытянуты независимо из плотности p , то вероятность того, что ровно K из них попадут в окрестность \mathcal{R} задается биномиальным распределением:

$$B(K|N, P) = C_N^K P^K (1 - P)^{N-K}.$$

Рассмотрим случайную величину $\frac{K}{N}$:

$$\mathbb{E} \frac{K}{N} = P, \quad \mathbb{D} \frac{K}{N} = \frac{P(1 - P)}{N}.$$

Для больших значений N все распределение будет сосредоточено в узком интервале вокруг матожидания, поэтому мы можем грубо полагать

$$K \approx NP. \tag{1}$$

Теперь предположим, что одновременно окрестность \mathcal{R} настолько мала, что плотность в ней можно приблизить константой, так что

$$P \approx p(\mathbf{x})V, \tag{2}$$

где V — площадь окрестности. С учетом (1) и (2) мы получаем

$$p(\mathbf{x}) = \frac{K}{NV}. \tag{3}$$

Фиксируя в (3) число точек K , попавших в окрестность, и определяя затем уже объем области V , мы приходим к алгоритмам ближайших соседей. Наоборот, зафиксировав объем окрестности V и затем определяя число точек, попавших в окрестность, мы приходим к методу парзеновского окна.