Восстановление плотности и EM-алгоритм

Викулин Всеволод

v.vikulin@corp.mail.ru

20 апреля 2019

Часть 1

Восстановление плотности. Математично и полезно.

Зачем восстановливать плотности?

Есть целое семейство алгоритмов, которые требует знание плотности распределений.

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)},$$

где $p(y|x)$– апостериорная вероятность принадлежность объекта x к классу y, $p(x|y)$– правдоподобие (likelihood), $p(y)$– априорная вероятность класса y, $p(x)$– правдоподие данных (evidence).

Вопрос

Как теперь предсказать метку класса?

Зачем восстановливать плотности?

Давайте разбираться, как считать:
$p(x|y)$ - берем все объекты класса и восстанавливаем плотность,
$p(y)$ - доля класса в выборке,
$p(x)-$ не очень то и нужна.
Осталось только восстановить плотность!

Пример

Подбрасываем монетку. Хотим узнать вероятность, что она выпадет орлом. Как использовать формулу Байеса?

Вопрос

А можно ли находить аномалии в данных с помощью плотности?

Методы восстановления плотности

Всего есть 3 метода восстановления плотности.

1. Непараметрическое
2. Параметрическое
3. Разделение смеси

Сегодня обсудим их все.

Непараметрическое восстановления плотности

Рассмотрим одномерный случай.
Если признак x категориальный - $p(x = k|y) =$ доля класса y, где признак x принимает значение k.
Если x – вещественный, то $p(x) = \lim_{h \to 0} P[x - h, x + h]$,
тогда $p(x|y) =$ доля точек класса y которые попали в окно $[x - h, x + h]$

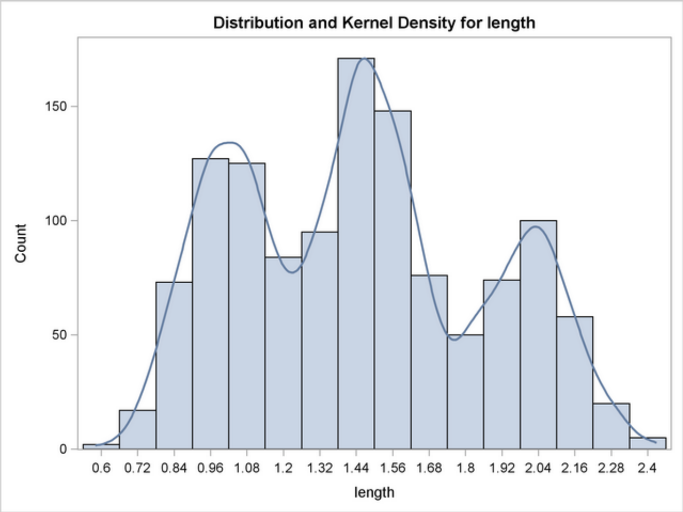
$$p(x|y) = \frac{1}{2Nh} \sum_{i=1}^{N} (|x - x_i| < h)[y_i = y]$$

Можно использовать метод парзеновского окна:

$$p(x|y) = \frac{1}{Nh} \sum_{i=1}^{N} K(\frac{x - x_i}{h})[y_i = y]$$

,
где K - нормированная функция, которую называют ядром.

Непараметрическое восстановления плотности



Параметрическое восстановления плотности

Считаем, что наши данные порождены параметрическим распределением, например, нормальным.

Вопрос

Почему всегда нормальным?

$$p(x|y,\theta) = \frac{1}{\sqrt{2\pi\sigma}} \exp(-\frac{(x-\mu_y)^2}{2\sigma_y^2})$$

,
осталось найти μ_y и σ_y.

Метод максимума правдоподобия

Правильные те параметры, при которых пронаблюдать такую выборку максимально правдоподобно!

Вы этот принцип применяете постоянно в жизни!

Пишем функцию правдоподобия, то есть вероятность пронаблюдать выборку:

$$L(\theta) = p(X|\theta) = \prod_{i=1}^{N} p(x_i|\theta)$$

$$\log L(\theta) = \sum_{i=1}^{N} \log p(x_i|\theta)$$

Правильные параметры те, которые максимизируют $L(\theta)$. Строим для каждого класса функцию правдоподобия, находим оптимальные параметры!

Пример

Подбрасываем монетку. Из n испытаний получили m орлов. Как оценить вероятность появления орла p?

$$L(\theta) = C_n^m p^m (1-p)^{n-m}$$

$$\log L(\theta) = \log C_n^m + m \log p + (n-m) \log(1-p)$$

$$\frac{\partial \log L(\theta)}{\partial p} = \frac{m}{p} - \frac{n-m}{1-p} = \frac{m - pm - np + pm}{p(1-p)} = 0$$

$$p = \frac{m}{n}$$

Просто процент выпадения орлов! А в чем тут проблема?

$$p = \frac{m + \lambda p_{prior}}{n + \lambda}$$

Пример

$$p(x|y,\theta) = \frac{1}{\sqrt{2\pi\sigma}} \exp(-\frac{(x-\mu_y)^2}{2\sigma_y^2})$$

$$\log L(\theta, y) = \sum_{i=1}^{N_y} \log p(x_i|y,\theta)$$

$$\log L(\theta, y) = -N_y \log \sigma_y - N_y \log \sqrt{2\pi} - \frac{1}{2\sigma_y^2} \sum_{i=1}^{N_y} (x_i - \mu_y)^2$$

$$\frac{\partial \log L(\theta|y)}{\partial \mu_y} = \frac{1}{\sigma^2} \sum_{i=1}^{N_y} (x_i - \mu_y) = 0$$

$$\mu_y = \frac{1}{N_y} \sum_{i=1}^{N_y} x_i$$

Пример

$$\log L(\theta, y) = -N_y \log \sigma_y - N_y \log \sqrt{2\pi} - \frac{1}{2\sigma_y^2} \sum_{i=1}^{N_y} (x_i - \mu_y)^2$$

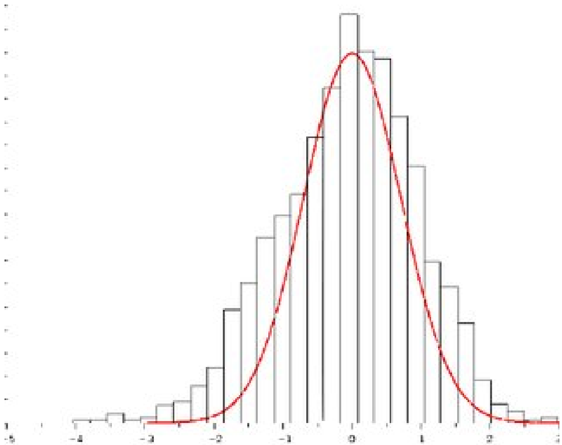
$$\mu_y = \frac{1}{N_y} \sum_{i=1}^{N_y} x_i$$

$$\frac{\partial \log L(\theta, y)}{\partial \sigma_y} = -\frac{N_y}{\sigma_y} + \frac{1}{\sigma^3} \sum_{i=1}^{N_y} (x_i - \mu_y)^2 = 0$$

$$\sigma^2 = \frac{1}{N_y} \sum_{i=1}^{N_y} (x_i - \mu_y)^2 = \frac{1}{N_y} \sum_{i=1}^{N_y} (x_i - \frac{1}{N_y} \sum_{i=1}^{N_y} x_i)^2$$

Решение по ММП - выборочное среднее и выборочная дисперсия!

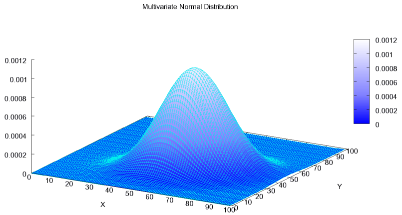
Пример



Многомерное нормальное распределение

$$N(x|\mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)},$$

μ - вектор размера n, Σ - ковариационная матрица $n \times n$, симметричная.



$$\mu_{\mathbf{y}} = \frac{1}{N_y}\sum_{i=1}^{N_y}\mathbf{x_i}, \Sigma = \frac{1}{N_y}\sum_{i=1}^{N_y}(\mathbf{x_i} - \mu_{\mathbf{y}})^T(\mathbf{x_i} - \mu_{\mathbf{y}})$$

Часть 2

EM алгоритм. Понять трудно, но Вы справитесь.

Модели со скрытыми переменными

Скрытые переменные – переменные, которые мы не наблюдаем, но которые влияют на внутреннее состояние модели.

- Кластеризация
- Машинный перевод
- Распознавание речи
- Тематическое моделирование
- Все, что сами сможете придумать

EM (Expectation-maximization) – алгоритм, позволяющий находить оценку максимального правдоподобия в задачах со скрытыми переменными.

Z – скрытые переменные. Правдоподобие:

- Неполное $\log P(X|\Theta)$
- Полное $\log P(X, Z|\Theta)$

Разумеется, $\log P(X|\Theta) = \log \int P(X, Z|\Theta) dZ$
$\log P(X|\Theta)$ в сложных задачах, как правило, тяжело максимизировать – не является выпуклой функцией. Скрытые переменные Z можем подобрать сами, чтобы упростить задачу.

Дивергенция Кульбака-Лейбера

Часто нужно мерить расстояние между двумя вероятностными распределенями.

$$KL(q||p) = \int q(x) \log \frac{q(x)}{p(x)} dx$$

$$KL(q||p) = -\int q(x) \log p(x) dx + \int q(x) \log q(x) dx$$

KL дивергенция **неотрицатальна**, она обращается в нуль тогда и только тогда, когда $q = p$, но при этом **не является метрикой** (Почему?).

Вывод EM алгоритма

Хотим максимизировать $\log P(X|\Theta) \to \max\limits_{\Theta}$.

Z – скрытые переменные, имеющие распределение $q(Z)$.

$$\log P(X|\Theta) = \int q(Z)\log P(X|\Theta)dZ = \int q(z)\log\frac{P(X,Z|\Theta)}{P(Z|X,\Theta)}dZ =$$

$$= \int q(z)\log\frac{P(X,Z|\Theta)q(Z)}{P(Z|X,\Theta)q(Z)}dZ =$$

$$= \int q(Z)\log\frac{P(X,Z|\Theta)}{q(Z)}dZ + \int q(Z)\log\frac{q(Z)}{P(Z|X,\Theta)}dZ =$$

$$= L(q,\Theta) + KL(q||P) \geq L(q,\Theta)$$

Вывод EM алгоритма

$$\log P(X|\Theta) = L(q,\Theta) + KL(q||P) \geq L(q,\Theta)$$

Будем максимизировать нижнюю оценку $L(q,\Theta)$ сначала по q, потом по Θ. Очевидно, что $L(q,\Theta)$ максимальна, когда $KL(q,\Theta) = 0$.

Е шаг:

$$q^*(Z) = arg\min_q \int q(Z) \log \frac{q(Z)}{P(Z|X,\Theta^{old})} dZ = P(Z|X,\Theta^{old})$$

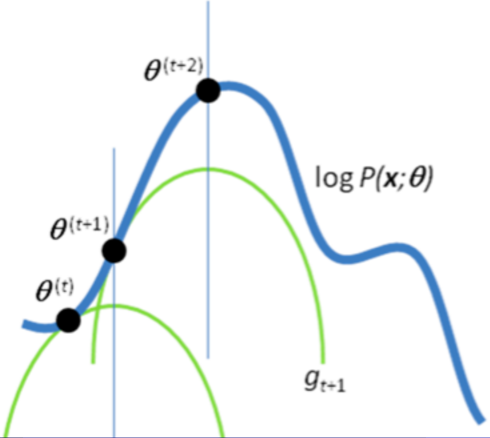
М шаг:

$$\Theta^{new} = arg\max_\Theta \int q^*(Z) \log \frac{P(X,Z|\Theta)}{q^*(Z)} dZ = arg\max_\Theta \int q^*(Z) \log P(X,Z|\Theta) dZ$$

То есть вместо $\log P(X|\Theta) \to \max\limits_\Theta$, на М шаге решаем $\mathbb{E}_Z \log P(X,Z|\Theta) \to \max\limits_\Theta$

Вывод EM алгоритма

На каждой итерации мы не уменьшаем правдоподобие – на E шаге нижняя оценка L равна правододобию, на M шаге мы ее максимизируем. Если правдоподие ограничено, то EM алгоритм **сходится к стационарной точке**.



Пример

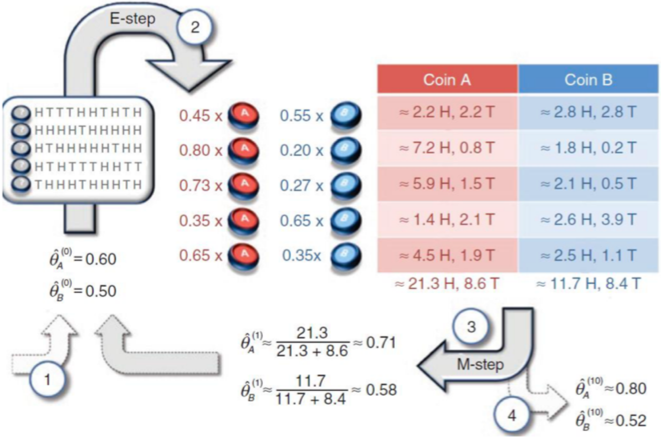
Coin A	Coin B
	5 H, 5 T
9 H, 1 T	
8 H, 2 T	
	4 H, 6 T
7 H, 3 T	
24 H, 6 T	9 H, 11 T

$$\theta_1 = \frac{24}{24+6} = 0.8$$

$$\theta_2 = \frac{9}{9+11} = 0.45$$

Источник: Adv. Statistical Machine Learning

Пример



Источник: Adv. Statistical Machine Learning

Пример

$\theta_A = 0.6, \theta_B = 0.5$

$$L(\theta) = C_{10}^{9} p^{9} (1-p)^{10-9}$$

$$L(\theta_A) = 0.004, L(\theta_B) = 0.001$$

. Для 2 броска вероятней, что это монетка A! А как оценить апостериорную вероятность?

$$P(Z = A | X, \Theta^{old}) = \frac{0.004}{0.004 + 0.001} = 0.8, P(Z = B | X, \Theta^{old}) = \frac{0.001}{0.004 + 0.001} = 0.2$$

Дальше считаем мат.ожидание. Для монетки A число орлов $9 * 0.8 = 7.2$, число решек $1 * 0.8 = 0.8$. Для монетки B число орлов $9 * 0.2 = 1.8$, число решек $1 * 0.2 = 0.2$. Дальше считаем θ_A, θ_B не через реальное число бросков, а через их мат.ожидания! Все как в формулах!

Смесь распределений. Кластеризуем мягко.

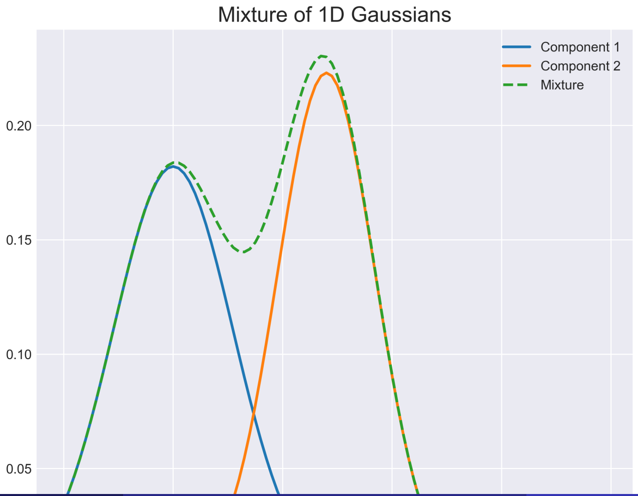
Смесь распределений

Говорят, что $p(x)$ – смесь распределений, если

$$p(x) = \sum_{k=1}^{K} \pi_k p_k(x), \sum_{k=1}^{K} \pi_k = 1, \pi_k \geq 0,$$

где K – число компонент смеси, $p_k(x)$ – распредление k компоненты, π_k – априорная вероятность k компоненты.

Смесь нормальных распределений



Правдоподобие

Пусть нам дана выборка размера N. Параметризуем $p_k(x) = \phi(x|\Theta_k)$

$$\log P(X|\Theta) = \log \prod_{i=1}^{N} p(x_i|\Theta) = \sum_{i=1}^{N} \log \sum_{k=1}^{K} \pi_k \phi(x_i|\Theta_k)$$

Введем скрытую переменную, которая будет отвечать за выбор компоненты.
z – k-мерный вектор, у которого одна компонента равна 1, а остальные равны 0.

$$\log P(X, Z|\Theta) = \log \prod_{i=1}^{N} p(x_i, Z|\Theta)$$

$$p(x_i, Z|\Theta) = \prod_{k=1}^{K} [\pi_k \phi(x_i|\Theta_k)]^{z_{i,k}}$$

$$\log P(X, Z|\Theta) = \sum_{i=1}^{N} \sum_{k=1}^{K} z_{i,k}(\log \pi_k + \log \phi(x_i|\Theta_k))$$

Смесь нормальных распределений

Наша смесь:

$$P(X|\Theta) = \prod_{i=1}^{N} p(x_i|\mu, \Sigma) = \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k N(x_i|\mu_k, \Sigma_k)$$

E шаг: считаем апостериорное распределение на скрытые переменные

$$p(z_{i,k} = 1|x_i, \Theta^{old}) = \frac{p(z_{i,k} = 1)p(x_i|z_{i,k} = 1, \Theta^{old})}{p(x_i|\Theta^{old})} = \frac{\pi_k^{old} N(x_i|\mu_k^{old}, \Sigma_k^{old})}{\sum_{j=1}^{K} \pi_j^{old} N(x_i|\mu_j^{old}, \Sigma_j^{old})} = g_{i,k}$$

Смесь нормальных распределений

Полное правдоподобие:

$$\log P(X, Z|\mu_k, \Sigma_k) = \sum_{i=1}^{N}\sum_{k=1}^{K} z_{i,k}(\log \pi_k + \log N(x_i|\mu_k, \Sigma_k))$$

М шаг: максимизируем мат. ожидание логарифма полного правдоподобия

$$E_z \log P(X, Z|\mu_k, \Sigma_k) = E_z \sum_{i=1}^{N}\sum_{k=1}^{K} z_{i,k}(\log \pi_k + \log N(x_i|\mu_k, \Sigma_k)) =$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} g_{i,k}(\log \pi_k + \log N(x_i|\mu_k, \Sigma_k)) \to \max_{\mu_k, \Sigma_k, \pi_k}$$

при условии $\sum_{k=1}^{K} \pi_k = 1$,

Смесь нормальных распределений

Можно аналитически найти максимум:

$$\pi_k = \frac{1}{N}\sum_{i}^{N} g_{i,k}$$

$$\mu_k = \frac{\sum_i^N g_{i,k} x_i}{\sum_i^N g_{i,k}}$$

$$\Sigma_k = \frac{\sum_i^N g_{i,k}(x_i - \mu_k)(x_i - \mu_k)^T}{\sum_i^N g_{i,k}}$$

$g_{i,k}$ – вес объекта i в компоненте k(насколько объект подходит под компоненту)
априорная вероятность компоненты π_k – средний вес компоненты по выборке
параметры нормального распределения считаются по тем же формулам, что и в принципе максимума правдоподобия, но взвешены с помощью $g_{i,k}$.

EM и K-means

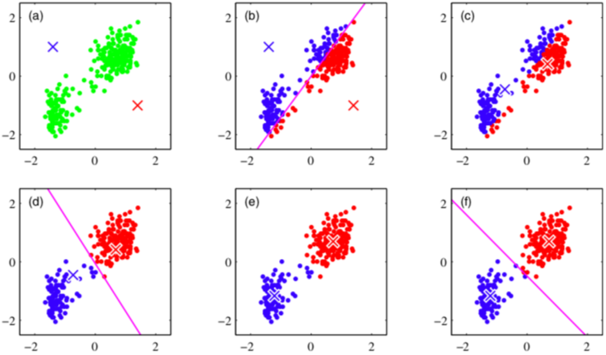
Разделяем смесь нормальных распределений. Пусть $\Sigma = \sigma^2 I$, единичная матрица, σ^2 стремится к нулю, априорные вероятности кластеров равны.

$$p(z_{i,k} = 1 | x_i, \Theta^{old}) = \frac{\pi_k^{old} N(x_i | \mu_k^{old}, \Sigma_k^{old})}{\sum\limits_{j=1}^{K} \pi_j^{old} N(x_i | \mu_j^{old}, \Sigma_j^{old})} = \frac{exp(-\frac{1}{2\sigma^2} ||x_i - \mu_k||^2)}{\sum_{j=1}^{K} exp(-\frac{1}{2\sigma^2} ||x_i - \mu_j||^2)} = g_{i,k}$$

Если σ^2 стремится к нулю, то $g_{i,k} = 1$ для самого близкого к объекту i кластеру и $g_{i,k} = 0$ для всех остальных кластеров.
Дальше пересчитали единственный параметр:

$$\mu_k^{new} = \frac{\sum_i^N g_{i,k} x_i}{\sum_i^N g_{i,k}}$$

k-means



Источник: Bishop

Заключение

Спасибо за внимание!

Дополнение. Пример на ММП. Показательное распределение

$$p(x|\theta) = \theta exp(-x\theta)$$

$$\log L(\theta) = \sum_{i=1}^{N} \log p(x_i \theta) = N \log \theta - \theta \sum_{i=1}^{N} x_i$$

$$\frac{\partial \log L(\theta)}{\partial \theta} = \frac{N}{\theta} - \sum_{i=1}^{N} x_i = 0$$

$$\theta = \frac{N}{\sum_{i=1}^{N} x_i}$$

Дополнение. Пример на ММП. Распределение Пуассона

$$p(x|\theta) = \frac{\theta^x exp(-\theta)}{x!}$$

$$\log L(\theta) = \sum_{i=1}^{N} \log p(x_i|\theta) = -N\theta + \sum_{i=1}^{N} (x_i \ln \theta - \ln x_i!)$$

$$\frac{\partial \log L(\theta)}{\partial \theta} = N - \frac{1}{\theta} \sum_{i=1}^{N} x_i = 0$$

$$\theta = \frac{1}{N} \sum_{i=1}^{N} x_i$$