Задача кластеризации и EM-алгоритм

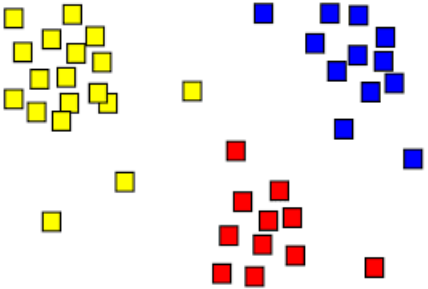
Викулин Всеволод

v.vikulin@corp.mail.ru

16 ноября 2018

Задача кластеризации

Разбиение исходного набора объектов на группы таким образом, чтобы объекты в группе были похожи друг на друга, а объекты из разных групп - отличались. Обучение **без учителя**.



Примеры применения

- Сегментация
- Суммаризация
- Обнаружение аномалий
- В помощь для классификации/регрессии

На самом деле намного больше!

Мы учимся без учителя

When we're learning to see, nobody's telling us what the right answers are — we just look. Every so often, your mother says "that's a dog", but that's very little information. You'd be lucky if you got a few bits of information — even one bit per second — that way. The brain's visual system has 10^{14} neural connections. And you only live for 10^9 seconds. So it's no use learning one bit per second. You need more like 10^5 bits per second. And there's only one place you can get that much information: from the input itself. — Geoffrey Hinton, 1996

Метрики качества кластеризации

Можно разделить на 2 типа:

1. Интутивные – свои близко, чужие подальше
2. По размеченным кластерам

Silhouette

Хотим, чтобы каждый объект к своему кластеру находился ближе, чем к соседнему.
Пусть C_i – кластер объекта i
a_i – среднее расстояние до объектов из кластера C_i,
b_i – среднее расстояние до объектов из ближайшего к C_i кластера

$$silhouette_i = \frac{b_i - a_i}{max(a_i, b_i)}$$

$$silhouette = \frac{1}{N} \sum_{i=1}^{N} \frac{b_i - a_i}{max(a_i, b_i)}$$

$silhouette \in [-1, 1]$

Rand Index

Пусть дана правильная кластеризация π^*, мы построили кластеризацию π.

Вопрос

Можно ли использовать метрики качества классификации?

Рассмотрим все пары объектов, проставим паре класс 1, если по π^* объекты из одного кластера и 0 – если из разных. Сделаем предскания для пар кластеризацией π.
Пусть a – число пар, где оба объекта принадлежат одному кластеру как в π^*, так и в π (True positive). b – число пар, где оба объекта принадлежат разным кластерам как в π^*, так и в π (True negative).

$$R = \frac{a + b}{C_N^2}$$

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.

Максимизация правдоподобия

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 **неотрицательна**, 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 $KL(q,\Theta) = 0$.

E шаг:

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

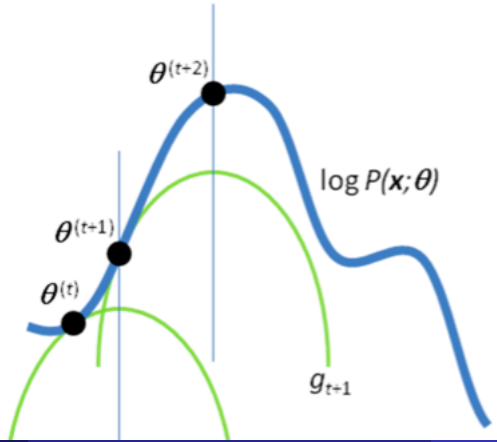
M шаг:

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

То есть вместо $\log P(X|\Theta) \to \max\limits_\Theta$, на M 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 праводододобию, на M шаге мы ее максимизируем. Если правдоподие ограничено, то EM алгоритм **сходится к стационарной точке**.



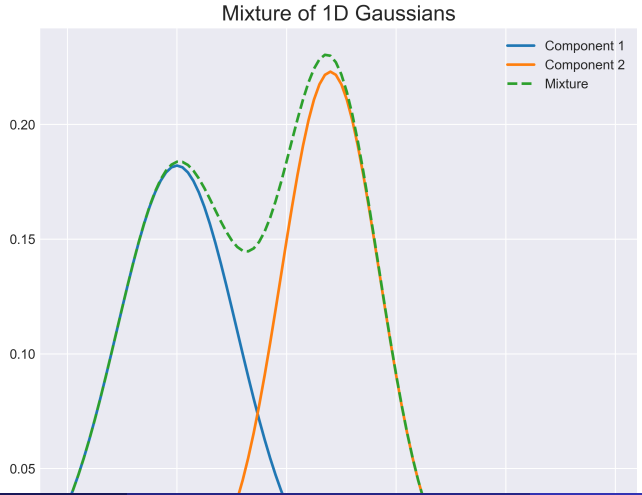
Смесь распределений

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.

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

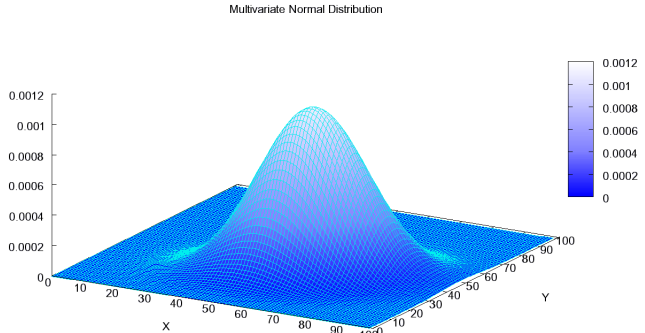
$$p(x_i, Z|\Theta) = \prod_{k=1}^{K} [\pi_k \phi(x_i|\Theta_k)]^{z_{i,k}}$$

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

Многомерное нормальное распределение

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

где μ - вектор средних значений размера n, Σ – ковариационная матрица размера $n \times n$, симметричная.



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

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

Смесь нормальных распределений

Полное правдоподобие:

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

M 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.

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

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:

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

Заключение

Спасибо за внимание!