Прикладная статистика в машинном обучении

Лекция №5

Адыгамов Ильяс 181, Болотин Арсений 182, Агаев Фархат 188

План лекции

1. Опишем EM алгоритм на другом языке используя дивергенцию Кульбака-Лейблера
2. Изучим Bootstrap: способ построения доверительных интервалов

Обозначения

- $\theta = (\theta_1,\ \theta_2,\ \ldots,\ \theta_n)$ - вектор неизвестных параметров
- $x = (x_1, x_2,\ \ldots, x_n)$ - наблюдения
- $z = (z_1,\ z_2,\ \ldots,\ z_n)$ - латентная переменная

Постановка задачи:

Мы хотим максимизировать лографим функции правдоподобия, чтобы найти оценку вектора параметров θ

$$\max_{\theta} \ln p(x|\theta)$$

Однако бывает так, что данная функция имеет такой вид, что максимизировать её сложно. Поэтому мы заменим данную процедуру на **EM** алгоритм.

EM-алгоритм в общем виде

- **Init**. Задать начальные условя для $\theta_{old} := \theta_{init}$
- **E-step**. Найти условное распределения латентной переменной $p(z|x,\theta)$

$$Q(\theta, \theta_{old}) = \mathbb{E}[\ln p(x, z|\theta)]$$

- **M-step**. Максимизируем функцию Q по θ. Получаем θ_{new}. Обновляяем $\theta_{old} := \theta_{new}$
- *Повторяем E-шаг и M-шаг до сходимости*

С точки зрения программирования **M-шаг** довольный легкий, так как есть куча оптимизаторов и максимизировать функцию будет несложно. На **E-шаге** нужно думать, поэтому заменим на что-то более простое, где нужно будет оптимизировать.

То есть глобльная наша цель поручить нахождение $p(z|x,\theta)$ компьютеру. В этом нам поможет дивергенция Кульбака-Лейблера.

$$KL(q||p_{z|x,\theta}) = CE(q||p_{z|x,\theta}) - H(q)$$

q - распределение кандидат, $p_{z|x,\theta}$ - распределение, которое хотим найти.

Напоминание: Дивергенция Кульбака-Лейблера на примере данетки

$$KL(a||b) = CE(a||b) - H(a)$$

- a - истинное распределение загадывающего данетку
- b - распределение, которое предполагает разгадывающий данетку
- CE(a||b) - среднее количество вопросов на разгадывание
- H(a) - среднее количество вопросов в идеальной ситуации (когда разгадывающий знает истинные вероятности, с которыми загадывающий загадывает данетку)
- CE(a||b) - H(a) - лишние вопросы, заданные разгадывающим (в среднем)

Также мы знаем, что $KL(a||b) \geqslant 0$

Первая идея (наивная) давайте попробуем минимизировать дивергенцию Кульбака-Лейблера:

$$\min_q KL(q||p_{z|x,\theta})$$

Мы знаем, что в таком случае минимум данной функции будет достигаться когда

$$q^* = p_{z|x,\theta}$$

То есть наивная идея не сработала так как, чтобы выписать KL, надо уже знать $p_{z|x,\theta}$. Но, к счастью, сработает лайфхак

Лайфхак

$$\ln p(x\theta) = KL(q||p_{z|x,\theta}) + LB(q,\theta)$$

- $\ln p(x\theta)$ не зависит от q
- $KL(q||p_{z|x),\theta} \geqslant 0$
- LB - lower bound

То есть мы попробуем заменить минимизацию KL на такой **Е-шаг:**

$$LB(q,\theta) \to \max_q, \text{ где ответом будет } q^* = p_{z|x,\theta}$$

И в этот моменте произойдет чудо: для записи LB не нужно заранее знать $p_{z|x,\theta}$

Еще одно маленькое напоминание

$$CE(q||p) = \int q(z) \log_{\frac{1}{2}} p(z)dz = -\int q(z) \log_2 p(z)dz \text{ (в битах)}$$

При переводе в наты появится константа

$$= -\int q(z) \ln p(dz) \cdot C$$

А теперь распишем LB

$$LB(q,\theta) = \ln p(x|\theta) - KL(q||p_{z|x,\theta}) =$$

$$= \ln p(x|\theta) - [CE(q||p_{z|x,\theta}) - H(q)] = \ln p(x|\theta) - (-\int q(z) \ln p(z|x,\theta)dz + \int q(z) \ln q(z)dz) =$$

$$= \int q(z) \ln p(x|\theta)dz + \int q(z) \ln p(z|x,\theta)dz - \int q(z) \ln q(z)dz$$

$$= \int q(z) \cdot \ln \frac{p(x|\theta) \cdot p(z|x,\theta)}{q(z)} dz = \int q(z) \cdot \ln \frac{p(x,z|\theta)}{q(z)} dz$$

В итоге как мы можем видеть наша функция LB не зависит от $p_{z|x,\theta}$. ☺