

Машинное обучение, ФКН ВШЭ

Теоретическое домашнее задание №6

Задача 1. Для двух одномерных нормальных распределений $\mathcal{N}(x | \mu_1, \sigma_1)$, $\mathcal{N}(x | \mu_2, \sigma_2)$ найдите дивергенцию Кульбака-Лейблера.

$$\text{KL}(\mathcal{N}(x | \mu_1, \sigma_1) \| \mathcal{N}(x | \mu_2, \sigma_2)) - ?$$

Задача 2. Рассмотрим метод восстановления плотности распределения с помощью гистограмм. Разобьем все пространство на непересекающиеся области δ_i . Каждому δ_i ставится в соответствие вероятность h_i . По заданной выборке $\{x_i\}_{i=1}^\ell$, найдите оптимальные значения h_i с помощью метода максимального правдоподобия.

Задача 3. Рассмотрим общую схему ЕМ-алгоритма, выводимую через разложение

$$\log p(X | \Theta) = \mathcal{L}(q, \Theta) + \text{KL}(q \| p).$$

На Е-шаге ищется распределение q , доставляющее максимум нижней оценке $\mathcal{L}(q, \Theta^{\text{old}})$ при фиксированном Θ^{old} .

Модифицируем Е-шаг: будем теперь искать максимум не среди всех возможных распределений, а лишь среди вырожденных, то есть присваивающих единичную вероятность одной точке и нулевую вероятность всем остальным. Как будут выглядеть Е- и М-шаги в этом случае?

Задача 4. Наблюдается выборка бинарных значений $y = (y_1, \dots, y_n)$, $y_i \in \{0, 1\}$. Все элементы выборки генерируются независимо, но известно, что в некоторый момент z меняется частота генерации единиц. Т.е., для всех $i < z$ $P(y_i = 1) = \theta_1$, а для всех $i \geq z$ $P(y_i = 1) = \theta_2$. Необходимо вывести формулы для ЕМ-алгоритма, где z – скрытая переменная, а θ_1, θ_2 – параметры распределений.

Задача 5. Аналогично рассмотренному на семинаре релаксированному методу минимизации функционала $\text{RatioCut}(A_1, \dots, A_k)$, сформулируйте и решите релаксированную задачу для минимизации функционала при $k = 2$:

$$\text{Ncut}(A_1, \dots, A_k) = \frac{1}{2} \sum_{i=1}^k \frac{W(A_i, \bar{A}_i)}{\text{vol}(A_i)}, \quad \text{vol}(A) = \sum_{i \in A} d_i.$$