

Машинное обучение, ФКН ВШЭ

Теоретическое домашнее задание №6

Задача 1. Найдите дивергенцию Кульбака-Лейблера, если она определена,

- а) из биномиального $\text{Bin}(n = 2, p = 1/3)$ в равновероятное на 0, 1, 2;
- б) из равновероятного на 0, 1, 2 в биномиальное $\text{Bin}(n = 2, p = 1/3)$;
- в) из $\mathcal{N}(\mu_1; \sigma_1^2)$ в $\mathcal{N}(\mu_2; \sigma_2^2)$;
- г) из $\mathcal{N}(0; 1)$ в экспоненциальное с $\lambda = 1$;
- д) из экспоненциального с $\lambda = 1$ в $\mathcal{N}(0; 1)$;

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

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

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 $i < z$ выполнено $P(y_i = 1) = \theta_1$, а для всех $i \geq z$ выполнено $P(y_i = 1) = \theta_2$. Необходимо вывести формулы для ЕМ-алгоритма, где z — скрытая переменная, а θ_1, θ_2 — параметры распределений.

Задача 5. Новогодние праздники подошли к концу. Все семинаристы курса по МО-2 хорошо кушали и теперь хотят похудеть. Вес семинариста имеет распределение $x_i \sim \mathcal{N}(0, \sigma^2)$. Весы работают с погрешностью $\varepsilon_i \sim \mathcal{N}(0, 1)$. После взвешивания каждый семинарист видит величину $y_i = x_i + \varepsilon_i$.

- а) Найдите оценку максимального правдоподобия для σ^2 . Выразите её через y_1, \dots, y_ℓ .
- б) Семинаристы хотят оценить σ^2 с помощью ЕМ-алгоритма. Выпишите E -шаг и M -шаг для нашей задачи. Найдите формулу пересчёта σ_t^2 в σ_{t+1}^2 . Найдите предел $\lim_{t \rightarrow \infty} \sigma_t^2$.
- в) Предложите семинаристам способ выяснить с помощью EM -алгоритма их настоящий вес.

Задача 6. Пусть мы пытаемся предсказать переменную-счётчик с аномальным значением в нуле. Например, это может быть количество рыб, пойманных на рыбалке. Чаще всего это ноль. Если это не ноль, то это счётчик, который распределён по Пуассону. Такую модель называют *моделью с нулевым вздутием (zero inflated model)*:

$$P(y_i = 0) = p(x_i) + (1 - p(x_i)) \cdot e^{-\lambda(x_i)}$$

$$P(y_i = k) = (1 - p(x_i)) \cdot \frac{\lambda(x_i)^k \cdot e^{-\lambda(x_i)}}{k!}.$$

Под $\lambda(x_i)$ и $p(x_i)$ имеются в виду какие-то зависимости от факторов. Например, может быть $\lambda(x_i) = \langle w, x_i \rangle$, а $p(x_i)$ — логистическая регрессия. Если $p(x_i) = 0$, получается пуассоновская регрессия.

Руководствуясь принципом максимизации правдоподобия, получите для такой модели функцию потерь для оптимизации.