Машинное обучение, ФКН ВШЭ
Теоретическое домашнее задание №6

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

$$\mathrm{KL}(\mathcal{N}(x \mid \mu_1, \sigma_1) \| \mathcal{N}(x \mid \mu_2, \sigma_2))$$

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 EM-алгоритма, выводимую через разложение

$$\log p(X \mid \Theta) = \mathcal{L}(q, \Theta) + \mathrm{KL}\left(q \,\|\, p\right).$$

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

Модифицируем E-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 E- и M-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 $i \geqslant z$ выполнено $P(y_i = 1) = \theta_2$. Необходимо вывести формулы для EM-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 $y_1, \dots, y_\ell$.

б) Семинаристы хотят оценить σ^2 с помощью EM-алгоритма. Выпишите E-шаг и $M-$шаг для нашей задачи. Найдите формулу пересчёта σ_t^2 в σ_{t+1}^2. Найдите предел $\lim_{t\to\infty} \sigma_t^2$.

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.