Машинное обучение.
Домашнее задание №10

Задача 1. Рассмотрим задачу классификации на K классов: $Y = \{1, \dots, K\}$. Пусть на множестве пар номеров классов $Y \times Y$ задана функция потерь $L(y, s)$, значение которой на паре (y, s) мы обозначим через λ_{ys}. Покажите, что минимум функционала среднего риска $R(a)$ достигается на алгоритме

$$a_*(x) = \arg\min_{s \in Y} \sum_{y \in Y} \lambda_{ys} p(y) p(x \,|\, y).$$

Задача 2. Рассмотрим задачу классификации текстов $D = \{d_1, \dots, d_{|D|}\}$ на K классов $Y = \{1, \dots, K\}$. Каждый документ d_i представляет собой некоторое подмножество множества возможных слов $W = \{w_1, \dots, w_{|W|}\}$ (т.е. нас не интересует порядок слов и количество вхождений каждого слова). В качестве признаков для каждого документа выберем индикаторы вхождения слов в него. Матрица «объекты-признаки» задается как

$$x_{ij} = [w_j \in d_i], \quad i = 1, \dots, |D|,\ j = 1, \dots, |W|.$$

Для решения задачи воспользуемся наивным байесовским классификатором, который основывается на предположении, что признаки независимы:

$$p(x_{i1}, \dots, x_{i|W|} \,|\, y_i) = p(x_{i1} \,|\, y_i) \dots p(x_{i|W|} \,|\, y_i).$$

Будем считать, что при фиксированном классе каждый признак имеет распределение Бернулли. Таким образом, априорные распределения и функции правдоподобия задаются как

$$\begin{aligned} &p(k \,|\, \pi) = \pi_k, \quad k = 1, \dots, K; \\ &p(x_{ij} \,|\, k, \theta) = \theta_{jk}^{x_{ij}} (1 - \theta_{jk})^{1 - x_{ij}}, \quad i = 1, \dots, |D|,\ j = 1, \dots, |W|,\ k = 1, \dots, K. \end{aligned}$$

Распределение одного документа записывается следующим образом:

$$p(d_i, y_i \,|\, \pi, \theta) = p(y_i \,|\, \pi) \prod_{j=1}^{|W|} p(x_{ij} \,|\, y_i, \theta) = \prod_{k=1}^{K} \pi_k^{[y_i = k]} \prod_{j=1}^{|W|} \prod_{k=1}^{K} p(x_{ij} \,|\, k, \theta_{jk})^{[y_i = k]}.$$

Докажите, что оценки максимального правдоподобия на параметры π и θ имеют вид

$$\begin{aligned} &\hat{\pi}_k = \frac{\sum_i [y_i = k]}{|D|}, \\ &\hat{\theta}_{jk} = \frac{\sum_i [y_i = k][x_{ij} = 1]}{\sum_i [y_i = k]}, \end{aligned}$$

где все суммирования ведутся по документам от 1 до $|D|$.

Задача 3. Расширим модель из предыдущей задачи и введем априорные распределения на параметрах θ. В качестве априорного к распределению Бернулли удобно [1] брать бета-распределение

$$\mathrm{Beta}(x \mid \beta_1, \beta_2) = \frac{1}{B(\beta_1, \beta_2)} x^{\beta_1 - 1}(1-x)^{\beta_2 - 1},$$

где $\beta_1, \beta_2 \geqslant 0$, а $B(\beta_1, \beta_2)$ — бета-функция.

Выпишите апостериорные распределения

$$p(\theta_{jk} \mid D) \propto p(\theta_{jk}) \prod_{i=1}^{|D|} p(d_i, y_i \mid \theta_{jk}),$$

где $p(\theta_{jk}) = \mathrm{Beta}(\theta_{jk} \mid \beta_1, \beta_2)$.

В качестве оценок на θ_{jk} возьмем матожидания апостериорных распределений:

$$\hat{\theta}_{jk} = \int_0^1 \theta_{jk} p(\theta_{jk} \mid D) d\theta_{jk}.$$

Найдите их в явном виде. Какова роль параметров β_1 и β_2?

[1] Под «удобством» здесь понимается тот факт, что если функция правдоподобия имеет распределение Бернулли, а априорное распределение выбрано из класса бета-распределений, то апостериорное распределение тоже относится к классу бета-распределений.