Машинное обучение, ФКН ВШЭ
Теоретическое домашнее задание №16

Задача 1. Постройте двойственную задачу к задаче одноклассового SVM.

Задача 2. Покажите, что параметр ν, используемый в постановке задачи одноклассового SVM, является верхней оценкой на долю аномалий на обучающей выборке, т.е. что $\frac{1}{\ell}\sum_{i=1}^{\ell}[\langle w, x_i\rangle < \rho] \leqslant \nu$, где w — оптимальный вектор весов.

Задача 3. На семинаре было выведено байесовское решающее правило для задачи классификации. Напомним, что в прошлом семестре было показано, что для случая регрессии минимум среднего риска

$$R(a) = \mathbb{E}L(y, a(x)) = \int_{\mathbb{Y}}\int_{\mathbb{X}} L(y, a(x))p(x, y)dxdy$$

для квадратичной функции потерь $L(y, s) = (y-s)^2$ достигается на алгоритме $a^*(x) = \mathbb{E}(y \,|\, x) = \int_{\mathbb{Y}} yp(y|x)dy$. Для нахождения $p(y \,|\, x) \propto p(x \,|\, y)p(y)$ могут применяться методы восстановления плотности, однако иногда удобнее сразу задавать апостериорное распределение.

Пусть решается задача регрессии. Будем считать, что задан некоторый вектор весов w, и метка объекта $y(x)$ генерируется следующим образом:

$$y(x) = \langle w, x\rangle + \varepsilon,\ \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

Тогда апостериорное распределение примет следующий вид:

$$p(y \,|\, x, w) = \mathcal{N}(\langle w, x\rangle, \sigma^2).$$

Покажите, что метод максимального правдоподобия для полученной модели эквивалентен методу наименьших квадратов.

Задача 4. Пусть, как и в предыдущей задаче, рассматривается задача линейной регрессии. Введём априорное распределение на векторе весов:

$$p(w_j) = \mathcal{N}(0, \alpha^2),\ j = \overline{1, d}.$$

Покажите, что максимизация апостериорной вероятности $p(w \,|\, x, y)$ для полученной модели линейной эквивалентна решению задачи гребневой регрессии.

Задача 5. Пусть решается задача бинарной классификации текстов, $\mathbb{Y} = \{0, 1\}$. Будем считать, что имеется некоторый словарь V, и каждый текст описывается $|V| = d$ бинарных признаков, каждый из которых характеризует наличие или отсутствие соответствующего слова словаря в тексте: $X = \{(x_i, y_i)\}_{i=1}^{\ell}$, $x_i = (x_{i1}, \ldots, x_{id})$, $x_{ij} \in \{0, 1\}$, $i = \overline{1, \ell}$, $j = \overline{1, d}$. Пусть признаки независимы для объектов каждого класса, и j-ый признак объектов класса y генерируется из распределения Бернулли с параметрами p_{jy} и $q_{jy} \equiv 1 - p_{jy}$. Выведите оценки максимального правдоподобия для параметров p_{yj}, q_{yj}, $y \in \mathbb{Y}$, $j = \overline{1, d}$, распределений признаков внутри классов.

Задача 6. Пусть решается задача бинарной классификации текстов, $\mathbb{Y} = \{0, 1\}$. Будем считать, что каждый текст описывается последовательностью номеров слов из словаря V: $X = \{(x_i, y_i)\}_{i=1}^{\ell}$, $x_i = (x_{i1}, \ldots, x_{in_i})$, $x_{ij} \in \{1, \ldots, |V|\}$, $i = \overline{1, \ell}$, $j = \overline{1, n_i}$. Будем считать, что признаки независимы для объектов каждого класса и что признаки объектов класса y генерируются из одного и того же мультиномиального распределения с параметрами $\{p_{yk}\}_{k=1}^{|V|}$ (то есть распределение на номерах слов из словаря не зависит от позиции j слова в тексте). Выведите оценки максимального правдоподобия для параметров $\{p_{yk}\}_{k=1}^{|V|}$, $y \in \mathbb{Y}$, распределений признаков внутри классов.