Машинное обучение
ФКН ВШЭ
Теоретическое домашнее задание №4

Задача 1. Предположим, что целевая переменная y независима с признаками объекта. Докажите, что в таком случае дисперсия $\mathbb{D}[y]$ является нижней оценкой квадратичной ошибки любой модели.

Задача 2. Допустим, объекты описываются единственным признаком $x \in \mathbb{R}$, имеющим распределение p. Рассмотрим некоторую функцию $f(x)$, представимую рядом Тейлора в окрестности нуля. Запишем его: $f(x) = a(x) + \overline{o}(x^k)$, где $a(x)$ - многочлен степени не выше k. Пусть целевая переменная определена как $\mathbb{E}[y|x] = f(x)$. Возьмем многочлен $a(x)$ в качестве модели для регрессии y. Найдите смещение такой модели для следующих функций и распределений:

1. $f(x) = \sin(x), k = 1, p = U[-\frac{\pi}{2}, \frac{\pi}{2}]$
2. $(x) = e^{x/2}, k = 1, p = \mathcal{N}(0, 1)$
3. $f(x) = \sqrt{1-x}, k = 1, p = B(1, \frac{3}{2})$ (бета-распределение)

Указание: В последнем пункте воспользуйтесь табличными значениями гамма-функции.

Задача 3. Пусть $x \in \mathbb{R}^3$, и значения признаков равномерно распределены по шару радиуса R с центром в нуле. Пусть $\mathbb{E}[y|x] = \|x\|_2$. Найдите смещение константного алгоритма $\mu(X)(x) = C = \text{const}$. При каком значении C достигается минимум смещения?

Задача 4. (*) На семинаре выводилось разложение ошибки для одномерной линейной регрессии $\mu(X)(x) = k(X)x$. Вспомним модель порождения данных, которую мы использовали. Единственный признак генерировался из нормального распределения $x \sim \mathcal{N}(0, \sigma_1^2)$, а целевая переменная $y = f(x) + \varepsilon, \varepsilon \sim \mathcal{N}(0, \sigma_2^2)$. Рассмотрим теперь обучение модели с L_2-регуляризацией:

$$\sum_{i=1}^{\ell}(y_i - kx_i)^2 + \lambda k^2 \to \min_k$$

Как изменится шумовая компонента при использовании модели с регуляризацией? Найдите смещение и разброс модели для линейной $f(x) = ax$ и произвольной четной $f(x)$. Проанализируйте результаты при $\lambda \to \infty$.

Задача 5. (*) Предположим, что объекты описываются двумя независимыми признаками: $x = (x_1, x_2) \in \mathbb{R}^2$, каждый из которых имеет распределение Бернулли с параметром $p = \frac{1}{2}$. Пусть целевая переменная задана как $\mathbb{E}[y|x] = x_1 x_2$ и обучающая выборка X состоит из двух объектов. Будем строить решающее дерево по следующим правилам:

1. Разбиение объектов в вершине продолжается, пока они отличаются значением хотя бы одного признака.
2. Критерием информативности является дисперсия целевой переменной.
3. В случае равенства функционалов качества предпочтение отдается разбиению по первому признаку.

Найдите смещение и разброс такого решающего дерева.

Задача 6. (**) Рассмотрим пространство многочленов одной переменной степени не выше d: $p(x) = p_0 + p_1 x + ... + p_d x^d$. Пусть многочлены выступают в качестве объектов, а коэффициенты будут их признаками, распределенными нормально: $p_i \sim \mathcal{N}(0, 1)$. Допустим, что целевая переменная определена как $\mathbb{E}[y|p] = p(x_0)$, где x_0 - некоторое фиксированное (но нам неизвестное) число. Пусть обучающая выборка состоит из d многочленов: $p^1, ..., p^d$. Предложите алгоритм, минимизирующий сумму смещения и разброса.