

Правила

Работа выполняется самостоятельно. Вы можете пользоваться любыми источниками, но не можете общаться с другими людьми в процессе написания работы. Пожалуйста, пишите подробные решения!

Желаем удачи!

Задача 1. (20 баллов)

Рассмотрим задачу регрессии с одномерным пространством признаков. Пусть признак x может принимать только значения $-2, 0$ и 2 . Целевая переменная y является случайной и её распределение зависит от x следующим образом:

$$\begin{aligned} P(y = 6 \mid x = -2) &= \frac{4}{5}, & P(y = 0 \mid x = -2) &= \frac{1}{5}, \\ P(y = 3 \mid x = 2) &= \frac{1}{5}, & P(y = 0 \mid x = 2) &= \frac{4}{5}, \\ P(y = 0 \mid x = 0) &= 1. \end{aligned}$$

Мы решаем задачу с помощью решающего пня (решающего дерева с единственной нетерминальной вершиной). Функция потерь — квадратичная. Если в листе больше одного наблюдения, предсказание выбирается путём усреднения значений целевой переменной по всем наблюдениям, попавшим в лист.

Пусть обучающая выборка всегда состоит из трёх наблюдений и все значения x в выборке различны.

Найти ожидаемую квадратичную ошибку для предсказания в точке $x_{new} = 2$. Представить её в виде суммы шума, смещения и разброса.

Задача 2. (5+15+5 баллов)

Решаем задачу классификации с двумерным пространством признаков и двумя классами. Пусть обучающая выборка состоит всего из двух наблюдений:

x_1	x_2	y
-2	-7	0
-3	3	1

- Бандерлог из Лога решил использовать обычную логистическую регрессию (со свободным членом, без регуляризации), чтобы решить эту задачу. Что у него получилось?
- Бандерлог из Лога подумал-подумал и решил использовать логистическую регрессию с L_2 -регуляризацией. (Свободный член регрессии не входит в регуляризатор.) Найти уравнение разделяющей прямой. (Бандерлог относит объект к классу 1, если вероятность того, что он относится к классу 1, предсказанная логистической регрессией, больше $1/2$; разделяющая прямая соответствует вероятности $1/2$.)
- Бандерлог из Лога разочаровался в логистической регрессии и решил вместо неё использовать новый и модный инструмент: метод опорных векторов (SVM). Найти уравнение разделяющей прямой в этом случае.

Все утверждения аккуратно обосновать.

Задача 3. (15 баллов) Рассмотрим задачу регрессии с двумерным пространством признаков. Матрица объект-признак имеет следующий вид:

$$\begin{pmatrix} -8 & 0 \\ 0 & 2 \\ 0 & 2 \\ -8 & 0 \\ 0 & 2 \end{pmatrix}.$$

Значения y_i являются случайными величинами, определяемыми следующим образом:

$$y_i = \langle x_i, w \rangle + \varepsilon_i,$$

где ε_i — независимые случайные величины, принимающие значения 6 и -6 с равными вероятностями, а w — известный вектор истинных весов, равный $(-4, -1)$.

Пусть $\hat{w} = (\hat{w}_1, \hat{w}_2)$ — оценка для вектора весов, полученная с помощью метода наименьших квадратов. Найти совместное распределение \hat{w}_1 и \hat{w}_2 .

Задача 4. (12 баллов)

Джеймс Бонд решает задачу классификации текстов. Его выборка состоит из трёх документов (видимо, шифровок):

- a. haha haha haha one
- b. bad two bad
- c. three good one haha three

Бонд решает закодировать эти тексты с помощью своего собственного алгоритма TF²-IDF, похожего на TF-IDF. Разница с обычным TF-IDF состоит в том, что вместо частот слов (term frequency) берутся их квадраты. Логарифмы используются натуральные.

Построить таблицу, которая получится у Бонда после кодирования.