Правила

Строго запрещено:

- переговариваться (с любой целью);
- пользоваться устройствами связи (с любой целью — например, в качестве калькулятора);
- списывать (за исключением использования собственноручно написанного листочка формата A4).

Нарушение любого из этих пунктов влечет удаление с контрольной работы.

Можно использовать все факты, сформулированные на лекциях или в домашних работах, при наличии соответствующей ссылки и точной формулровки («по теореме такой-то верно то-то и то-то», «по задаче из домашнего задания известно, что» и т.д.).

Желаем удачи!

Задача 1. (5 баллов за каждый пункт) Рассеянный Александр решает задачу регрессии. В ходе предобработки данных один из признаков в датафрейме случайно сдублировался (то есть в датафрейме оказалось два совпадающих столбца), а остальные не изменились. Что изменится из-за этого дублирования (по сравнению с исходным датафреймом) с точки зрения процесса обучения и как это отразится на предсказаниях на обучающей выборке если Александр использует

a. метод k ближайших соседей (метрика — евклидова);
b. линейную регрессию без регуляризации, оценки для весов находит с помощью явной матричной формулы;
c. решающее дерево;
d. случайный лес?

Все ответы обосновать.

Задача 2. (20 баллов) Решается задача регрессии с одним признаком. Пусть истинный закон генерирования данных устроен следующим образом: для всякого x значение y определяется как случайная величина

$$y = 4x + 2 + \varepsilon, \tag{1}$$

где ε — случайная величина с нулевым матожиданием и дисперсией 16. У каждого наблюдения (объекта) в обучающей и тестовой выборках значения x фиксированы, а значения y определяются по указанной формуле, при этом ε для разных наблюдений независимы. В обучающей выборке четыре объекта, значения x для них равны -1, -1, 1 и 1. В тестовой выборке ровно один объект, значение x для него равно 5.

По обучающей выборке методом наименьших квадратов (без регуляризации) обучается линейная регрессия вида $y = w_0 + w_1 x$. Найти ожидаемую ошибку предсказания для квадратичной функции потерь на тестовой выборке и разложить её в сумму шума, смещения и разброса (найти явно значения шума, смещения и разброса, и указать, кто есть кто).

Иными словами, пусть $\mathcal{D} = \{(-1, y_1), (-1, y_2), (1, y_3), (1, y_4)\}$ — обучающая выборка (в которой значения y_i являются случайными величинами, заданными формулой (1)), $a(\mathcal{D})$ — это линейная регрессия, обученная на данных $\mathcal{D}$, то есть

$$a(\mathcal{D})\colon \mathbb{R}^1 \to \mathbb{R}^1,$$

$$a(\mathcal{D})(x) = w_0 + w_1 x,$$

где w_0 и w_1 — веса, находящиеся методом наименьших квадратов.

Пусть $x_{new} = 5$ и y_{new} определяется по формуле (1) для $x = x_{new}$. Найти

$$\mathbb{E}(a(\mathcal{D})(x_{new}) - y_{new})^2,$$

где матожидание берётся по $\mathcal{D}$ и y_{new}, и разложить его в сумму шума, смещения и разброса.

Задача 3. (15 баллов) Решается задача регрессии с двумерным пространством признаков. Обучающая выборка приведена в таблице.

x^1	x^2	y
3	1	-4
3	2	-5
5	1	-2
5	2	-8

Для получения предсказаний будем обучать решающий пень, то есть решающее дерево с одним предикатом. В качестве функции потерь возьмём квадратичную.

Найти среднее значение функции потерь на валидационном множестве при 4-фолдовой кросс-валидации.

Задача 4. (10 баллов) Рассмотрим задачу двухклассовой классификации. Пусть обученный метод машинного обучения для всякого значения признаков выдаёт свою степень уверенности в том, что объект относится к классу +1 (например, как это делает логистическая регрессия).

В таблице для каждого объекта обучающей выборки приведено значение уверенности алгоритма в том, что объект относится к классу +1 (первый столбец), и реальный класс этого объекта (второй столбец).

s_i	y_i
0.01	-1
0.14	-1
0.2	-1
0.29	-1
0.39	$+1$
0.42	$+1$
0.48	$+1$
0.74	$+1$
0.81	$+1$
0.97	$+1$

По данным в этой таблице можно построить ROC-кривую (но в задаче это не требуется; вопрос задачи см. ниже). Как известно, чем выше проходит ROC-кривая, тем лучше; если она проходит вдоль прямой $FPR = TPR$, такой классификатор не лучше, чем случайное подбрасывание монетки.

Можно ли перестановкой значений во втором столбце (при сохранении значений в первом) добиться того, чтобы ROC-кривая в некоторых точках проходила выше прямой $FPR = TPR$, а в других точках — ниже? Если нет — докажите. Если да, приведите пример такой перестановки.

Задача 5. (15 баллов) Рассмотрим задачу двухклассовой классификации с одномерным пространством признаков. Пусть обучающая выборка имеет следующий вид.

x	y
4	$+1$
4	$+1$
4	$+1$
4	$+1$
4	-1
2	$+1$
2	$+1$
2	-1

Будем решать её с помощью логистической регрессии со свободным членом:

$$\log \frac{p}{1-p} = w_0 + w_1 x,$$

где p — вероятность отнесения объекта к классу $+1$.

Найти предсказания логистической регрессии (значение p), обученной методом максимизации правдоподобия (без регуляризации), в точках $x = 4$ и $x = 2$.

Подсказка. Задачу можно решить, не находя весов w_0 и w_1.