Введение в анализ данных
Контрольная работа
Вариант 1

Задача 1 (1.5 балла). Ответьте на вопросы об обучении линейной модели регрессии:

1. Запишите модель линейной регрессии, поясните все обозначения.

2. Приведите пример выборки с одним признаком, на которой линейная модель со свободным членом w_0 может дать нулевую ошибку, а без свободного члена ошибка в любом случае будет больше нуля. Можно записать значения признака и ответа для объектов этой выборки, можно нарисовать её.

3. Мы обучаем модель линейной регрессии (с w_0) на выборке с 90 признаками, минимизируя среднеквадратичную ошибку. Сколько будет параметров у этой модели? А сколько будет параметров, если добавить в функционал L_2-регуляризатор?

Задача 2 (1.5 балла). Напомним, что при обучении метода опорных векторов решается следующая задача:

$$\begin{cases} \dfrac{1}{2}\|w\|^2 + C\sum\limits_{i=1}^{\ell}\xi_i \to \min\limits_{w,b,\xi} \\ y_i\left(\langle w, x_i\rangle + b\right) \geqslant 1 - \xi_i, \quad i = 1, \ldots, \ell, \\ \xi_i \geqslant 0, \quad i = 1, \ldots, \ell. \end{cases}$$

Ответьте на вопросы по линейным классификаторам и их обучению:

1. Объясните, на что влияет гиперпараметр C. Какую модель мы получим, если сделать его равным нулю? Ответ обоснуйте.

2. Допустим, мы обучили классификатор с помощью SVM и собираемся оценивать его качество на тестовой выборке. Почему обычно измеряют F-меру, а не среднее арифметическое?

3. Чтобы вычислить F-меру, надо подобрать порог t для классификатора $a(x) = \\ = \operatorname{sign}(\langle w, x\rangle - t)$. С помощью какой процедуры можно подбирать этот порог, если мы хотим получить максимальное значение F-меры на тестовой выборке?

4. Пусть теперь нам надо решить задачу многоклассовой классификации (с непересекающимися классами) на 10 классов. Опишите какой-нибудь способ, как сделать это, если мы умеем обучать только линейные классификаторы для двух классов с помощью метода опорных векторов.

Задача 3 (3 балла). Ответьте на вопросы по обучению решающих деревьев:

1. Опишите жадный алгоритм построения решающего дерева. Можно в виде псевдокода с пояснениями.

2. Допустим, мы решаем задачу регрессии и уже построили дерево. Какие значения хранятся в листьях этого дерева и как они вычисляются?

3. Говорят, что решающее дерево — это в некотором смысле линейная модель. Объясните, что это значит.

4. При построении решающего дерева можно задать минимальное число объектов в листе. Каким нужно выставлять это число, если мы хотим получить максимально переобученное дерево? Если хотим получить максимально простое дерево?

5. Нам нужно построить решающее дерево в задаче классификации, и мы решили использовать необычный критерий хаотичности (impurity) вершины

   $$H(p_1, \ldots, p_n) = p_1,$$

   где p_k — доля в текущей вершине объектов, относящихся к классу k. Напомним, что чем меньше значение этого критерия в обоих поддеревьях, тем лучше разбиение. Что не так с этим критерием? Почему он приведёт к тому, что будет построено плохое дерево?

Задача 4 (2 балла). Антон решил запрограммировать градиентный спуск для своей задачи. Для начала он записал общую схему того, что будет программировать:

- Инициализация: $w^0 = 0$;
- Градиентный шаг:

  $$w^t = 0.5w^{t-1} - \frac{1}{t^{0.5}}\nabla Q(w^{t-1} + w^{t-2});$$

- Останавливаемся, если $\|w^t/w^{t-1}\| < 0.01$.

После этого Антон уехал в отпуск и поручил запрограммировать этот алгоритм своему двойнику Тамерлану. Тамерлану крайне не хочется самому разбираться в идеях Антона. Помогите ему найти все эти ошибки, а заодно объясните, почему из-за них градиентный спуск будет работать не так, как надо. Подсказка: ошибок как минимум три.

Задача 5 (2 балла). Вам выдали классификатор $b(x)$, и вам предстоит разобраться, насколько она хороша. Для этого у вас есть тестовая выборка из 8 объектов. Ниже указаны правильные ответы и уверенности модели в положительном классе:

$b(x)$	0	2	1	4	3	10	11	-2
y	-1	-1	+1	+1	+1	-1	+1	-1

Выполните следующие шаги:

1. Нарисуйте ROC-кривую и посчитайте AUC-ROC.
2. Посчитайте точность и полноту этой модели при пороге $t = 5$.
3. Можно ли достичь полноты в хотя бы 25% при точности в 100%? Если да, укажите, при каком пороге. Если нет, поясните, почему.
4. Можно ли достичь полноты в 50% при точности в хотя бы 60%? Если да, укажите, при каком пороге. Если нет, поясните, почему.