Рубежный контроль 1

Схема РК

- РК проходит в устной форме и состоит из практической и теоретической части
- Студент берет билет с двумя практическими задачами на решение которых он может потратить не более 30 минут
- При решении задач можно пользоваться материалами курса и python для выполнения вычислений
- Далее студент защищает решенные задачи и отвечает на 2-3 теоретических вопроса
- Теоретические вопросы могут подразумевать работу с ручкой и бумагой

Примеры практических задач

Задача 1

1. Что такое мультиколлинеарность, как она влияет на линейные модели?
2. Пусть в датасете есть следующие признаки:
   - Пол (М, Ж)
   - Опыт работы (число)
   - Цвет глаз (Г, К, З, Другой)
   - Количество дипломов (число)

   Датасет без пропусков и каждая категория в датасете представлена.

   Запишите уравнение модели линейной регрессии со свободным членом (без конкретных значений коэффициентов) для предсказания уровня зарплаты в рублях, избежав при этом строгой мультиколлинераности (то есть уравнение вида $y = \beta_0 + \beta_1 x_{\text{опыт}} + \beta_2 x_{\text{пол=М}} + \dots$)

Задача 2

Для задачи линейной регрессии с одной переменной $\hat{y} = \beta_0 + \beta_1 x_1$ покажите, что

$$\beta_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i}{n}} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Задача 3

Дана следующая обучающая выборка для регрессии:

x_1	5.5	1.5	-3	-12	6.5	-3.5	7.5	-2	11	5
x_2	5	-7.5	8.5	-2	-4.5	-0.5	-3	-2	11	2
y	33	-63.5	66	-41	-25.5	-10.5	-14	-23	115	26

1. С помощью взвешенного метода k-NN выполните регрессию следующего датасета. Гиперпараметры алгоритма:

x_1	3.5	-4.5	-6.5	4.5
x_2	7.5	6	-3	-7

- $k = 3$
- метрика - Manhattan distance $d(a, b) = \sum_i |a_i - b_i|$
- вес i-го ближайшего соседа определяется как $w(i) = \frac{(k-i+1)}{k}$

2. Какова вычислительная сложность предсказания в наивной имплементации k-NN по выбоке с N объектами и D признаками?

Задача 4

1. Выпишите формулу F-меры. Почему метрика качества $\min(\text{precision}(a, X), \text{recall}(a, X))$ считается не очень хорошим способом объединения точности и полноты?

2. Пусть даны выборка X, состоящая из 8 объектов, и классификатор $a(x)$, предсказывающий оценку принадлежности объекта положительному классу. Предсказания $a(x)$ и реальные метки объектов приведены ниже:

$$\begin{aligned} a(x_1) &= 0.1, \quad y_1 = +1, \\ a(x_2) &= 0.8, \quad y_2 = +1, \\ a(x_3) &= 0.2, \quad y_3 = -1, \\ a(x_4) &= 0.25, \quad y_4 = -1, \\ a(x_5) &= 0.9, \quad y_5 = +1, \end{aligned}$$

Постройте ROC-кривую и вычислите AUC-ROC

Задача 5

С помощью алгоритма Naive Bayes предскажите значение целевой переменной *buy_computer* для объекта со следующими значениями признаков:

```
age > 40 & income = high & student = no & credit-rating = unknown
```

Для обучения модели используйте данные из таблицы и сглаживане Лапласа ($\alpha = 1$).

RID	age	income	student	credit_rating	Class: buys_computer
1	<=30	high	no	fair	no
2	<=30	high	no	excellent	no
3	31 . . . 40	high	no	fair	yes
4	>40	medium	no	fair	yes
5	>40	low	yes	fair	yes
6	>40	low	yes	excellent	no
7	31 . . . 40	low	yes	excellent	yes
8	<=30	medium	no	fair	no
9	<=30	low	yes	fair	yes
10	>40	medium	yes	fair	yes
11	<=30	medium	yes	excellent	yes
12	31 . . . 40	medium	no	excellent	yes
13	31 . . . 40	high	yes	fair	yes
14	>40	medium	no	excellent	no

Теоретические вопросы

1. Постановка задач классификации и регрессии.
2. Метрики качества классификации и регрессии.
3. Переобучение. Разделение выборки, скользящий контроль.
4. Задачи и особенности Text Mining. Этапы обработки текста.
5. Байесовский классификатор. Оптимальное байесовское правило. Naive Bayes.
6. Деревья решений. Постановка задачи. Рекурсивный алгоритм. Сложность алгоритма.
7. Деревья решений. Критерии разбения для классифкации и регрессии. Криетерии остановаю.
8. Деревья решений. Укорачивание деревьев. Работа с отстутствующими значениями.
9. Линейная регрессия. Модель линейной регрессии. Регуляризация.
10. Линейная регрессия. Байесовская интерпретация.
11. Линейная регрессия. Точное решние, выписать формулу обновления весов для SGD
12. Логистическая регрессия. Модель логистиечкой регрессии. Регуляризация
13. Логистическая регрессия. Байесовская интерпитация
14. Логистическая регрессия. Моделирование нескольких классов.
15. Метрические методы. KNN.
16. Методы оптимизация. Градиентная оптимизация.
17. Градиентный спуск. Стохастический градиентный спуск