«Метрические методы и валидация моделей»

Задача 1

Классифицировать объекты Δ, представленные на (Рис. 1), методом k-ближайших соседей:

1. k = 1, метрика Евклидова
2. k = 3, метрика Евклидова
3. k = 1, метрика Манхэттона
4. k = 3, метрика Манхэттона

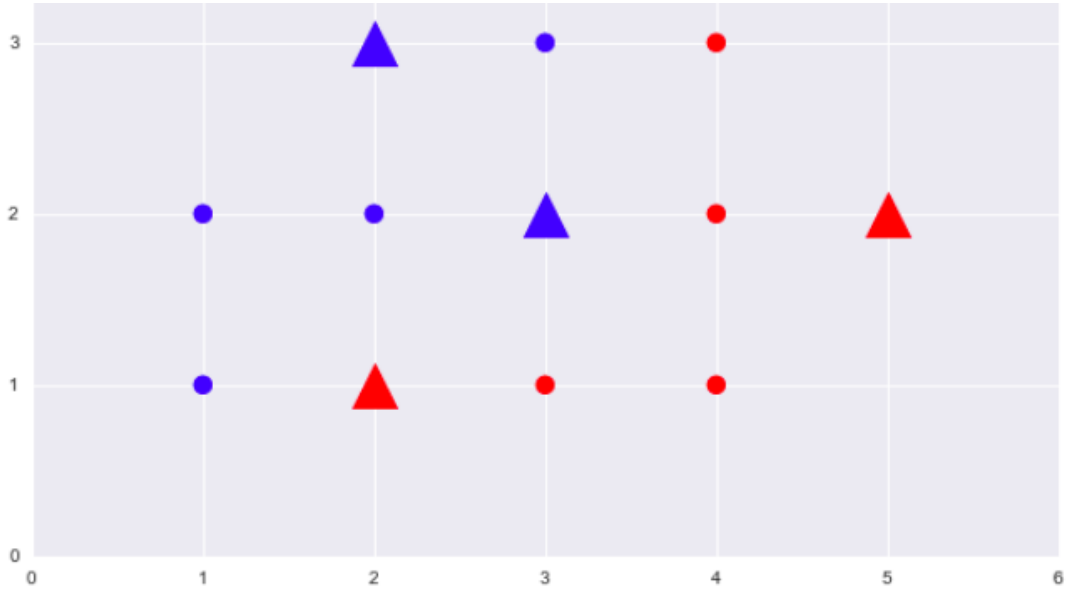

Рис. 1: Задание 1

При каких значениях параметров (метрика и количество соседей) достигается лучшая точность классификации?

Задача 2

Нарисовать решающую границу (1-nn) для следующего набора данных (Рис. 2) и провести классификацию объектов A и B методами 1-nn и 3-nn.

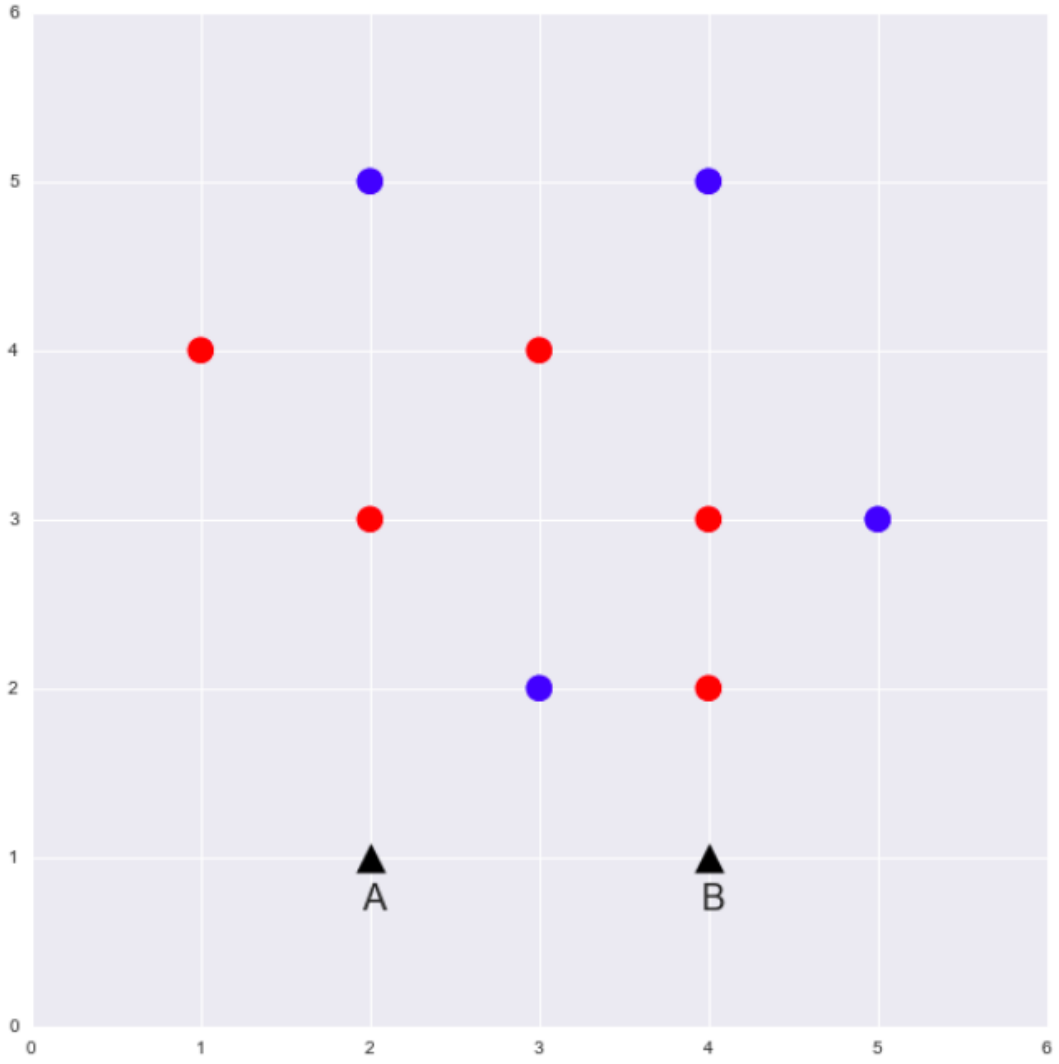

Рис. 2: Задание 2

✓ Задача 3 (шумовые признаки)

Пусть имеется два объекта: $x_1 = 0.1$ класса 1 и $x_2 = 0.5$ класса 2. Добавим к текущему признаку ещё один случайный: равномерно распределённый на отрезке $[0, 1]$. Посчитайте вероятность того, что новый объект $t(0, 0)$ будет отнесён ко второму классу.

Это же утверждение можно проверить эмпирически: сгенерируем двумерную равномерную выборку из [-2, 2] $\times$ [-2, 2] с метками $y = \begin{cases} 1, & x_1^3 - 2x_1 > 0 \\ 0, & x_1^3 - 2x_1 \leq 0 \end{cases}$. Найдём качество классификации по правилу

1-NN. Теперь сгенерируем ещё 15 случайных признаков и посмотрим на качество классификации метода 1-NN на новом признаковом пространстве.

Задача 4

1. Пусть ξ_i равномерно распределены на трёхмерном шаре радиуса r ($x^2 + y^2 + z^2 \leq r^2$). Найти средний квадрат расстояния от начала координат до точек шара

2. Найти матожидание квадрата расстояний между объектами, имеющими распределения $\mathcal{N}(a, \Sigma),\ \mathcal{N}(b, \Omega)$

✓ Задача 5 («проклятие размерности»)

Рассмотрим $d-$мерный куб объёма V.

1. Чему равна длина каждой из граней этого куба?

2. Мы хотим классифицировать новый объект с помощью метода ближайшего соседа. Предположим, известно, что в окрестности объёма $0.0018 \cdot V$ от заданного объекта находится тренировочный объект. На какую величину необходимо отойти от текущего объекта по одной из координат, чтобы дойти до тренировочного объекта. Привести рассчёты при условии, что $V = 1, d = 10$. А в случае $V = 1, d = 100$?

3. Посчитайте предел отношения объёма n-мерного шара радиуса R к объёму n-мерного куба с длиной грани $2 \cdot R$ при числе размерностей стремящемся к бесконечности. (Подсказка: Объём n-мерного шара равен $V_n = \frac{\pi^{\frac{n}{2}} R^n}{\Gamma(\frac{n}{2}+1)}$)

4. Сгенерируйте 2 набора по 10000 точек $(\{x_a\}_{i=1}^{1000}, \{x_b\}_{i=1}^{1000})$ из $\mathcal{N}(0, I_d)$. Изобразите плотности распределения расстояний $\|x_a - x_b\|_2$ для $d = 1, 2, 10, 100$. Какой вывод про евклидово расстояние в многомерных пространствах можно сделать?

Задача 6

1. Доказать, что Евклидова метрика и Манхэттонова метрика являются метриками.

2. Найти предел метрики Минковского при $p \to \infty$: $\lim\limits_{p \to +\infty} \left(\sum\limits_{i=1}^{d} |x_i - y_i|^p \right)^{\frac{1}{p}}$

Задача 7

Тренировочная выборка состоит из одномерных объектов первого класса: $0.5, 0.1, 0.2, 0.4, 0.3, 0.2, 0.2, 0.1, 0.35, 0.25$ и второго класса: $0.9, 0.8, 0.75, 1$.

1. Подогнать одномерные гауссианы для каждого из классов.
2. Оценить априорные вероятности первого и второго классов.
3. Оценить вероятность того, что объект $x_{test} = 0.6$ принадлежит классу 1.
4. Классифицировать объект $x_{test} = 0.6$ методами 1 и 3-х ближайших соседей.

✓ Задача 8

Для задачи 1 оценить точность распознавания с помощью LOO (leave-one-out) оценки методом:

1. 1-NN
2. 3-NN
3. Nearest centroid

✓ Задача 9

Для обучающей выборки $(x, y) : (1, 2), (3, 5), (5, 4), (7, 9)$:

1. Выполнить регрессию для точки $x = 3.5$ методом k-NN (расстояние - евклидово, k $= 3$) с весами $w_i = \frac{k-i+1}{k}$, где i - индекс i-ого ближайшего соседа
2. Изобразить прогнозные значения регрессии на промежутке $x \in [0, 8]$

✓ Задача 10

Загрузите набор текстов «20 News groups»

```
from sklearn.datasets import fetch_20newsgroups
data = fetch_20newsgroups(subset='all', remove=('headers', 'footers', '
    quotes'))
```

Сделайте Tf-idf преобразование текстов и примените метод K-NN (с косинусной метрикой) для классификации текстов. При каком количестве соседей достигается наибольшее качество на валидации? Как ещё можно улучшить качество классификации?

Задача 12

Примените алгоритм Neighbourhood Component Analysis (раз, два) к набору данных "wine"(from sklearn.datasets import load_wine). Чему равна точность предсказания на отложенной выборке методом KNN в исходном признаковом пространстве? А в пространстве, найденном методом NCA с числом компонент = 2?

Задача 13

Имеется выборка из позитивных и негативных отзывов о некотором заведении (для простоты - выборка размера 2 отзыва). Также имеется новый отзыв:

«Бургеры здесь вкусные, а атмосфера ужасная»

Требуется классифицировать этот отзыв (позитивный/негативный) методом ближайшего соседа с косинусной метрикой близости $\left(\rho_{\cos}(x,z) = 1 - \frac{\langle x,z\rangle}{\|x\|\cdot\|z\|}\right)$. В качестве признакового описания текста использовать среднее векторов слов, входящих в данный текст. Представление слов на рисунке 3 (удалены "стоп-слова": "на"и "а"):

Размеченная выборка отзывов:

Положительные:

«Котлетки на гриле - высший класс!»

Негативные:

«Ужасный ресторан, больше не придём сюда!»

ужасный	-0.21	-0.52	-0.12	-0.04
высший	0.14	0.14	0.39	-0.20
не	-0.11	-0.31	0.15	-0.17
гриле	-0.62	-0.37	0.30	-0.09
придём	0.44	0.33	-0.01	-0.03
здесь	-0.11	-0.31	0.15	0.16
ресторан	0.36	-0.57	-0.24	0.06
сюда	0.19	-0.27	-0.30	0.13
атмосфера	0.57	-0.03	0.21	-0.17
класс	0.36	-0.57	-0.24	-0.19
ужасная	-0.21	-0.52	-0.12	-0.34
бургеры	-0.36	-0.01	-0.02	0.26
больше	0.06	-0.42	-0.04	0.31
котлетки	-0.03	0.46	0.13	0.06
вкусные	-0.58	0.16	0.19	0.07

Рис. 3:

Задача 14

На основе датасета interactions_processed.csv построить рекомендации фильмов на основе implicit.nearest_neighbours.TFIDFRecommender. На кросс-валидации подобрать оптимальное число соседей.

✓Задача 15

Допустим, что признаковое описание для объектов состоит из векторов на сфере (то есть $\|x\|_2 = 1,\ \forall x \in \mathbb{X}$). Аркадий обучил k-NN на этих данных с косинусной метрикой и подобрал оптимальное число соседей (n_neighbors) которое даёт точность классификации порядка 0.9. Спустя время Аркадию нужно проскорить выборку этой моделью, однако он забыл, по какой метрике нужно искать соседей и использует евклидову метрику. Ожидается ли просадка качества такой модели?

(Решение)

«Решения»
(Colab notebook)

Задача 3

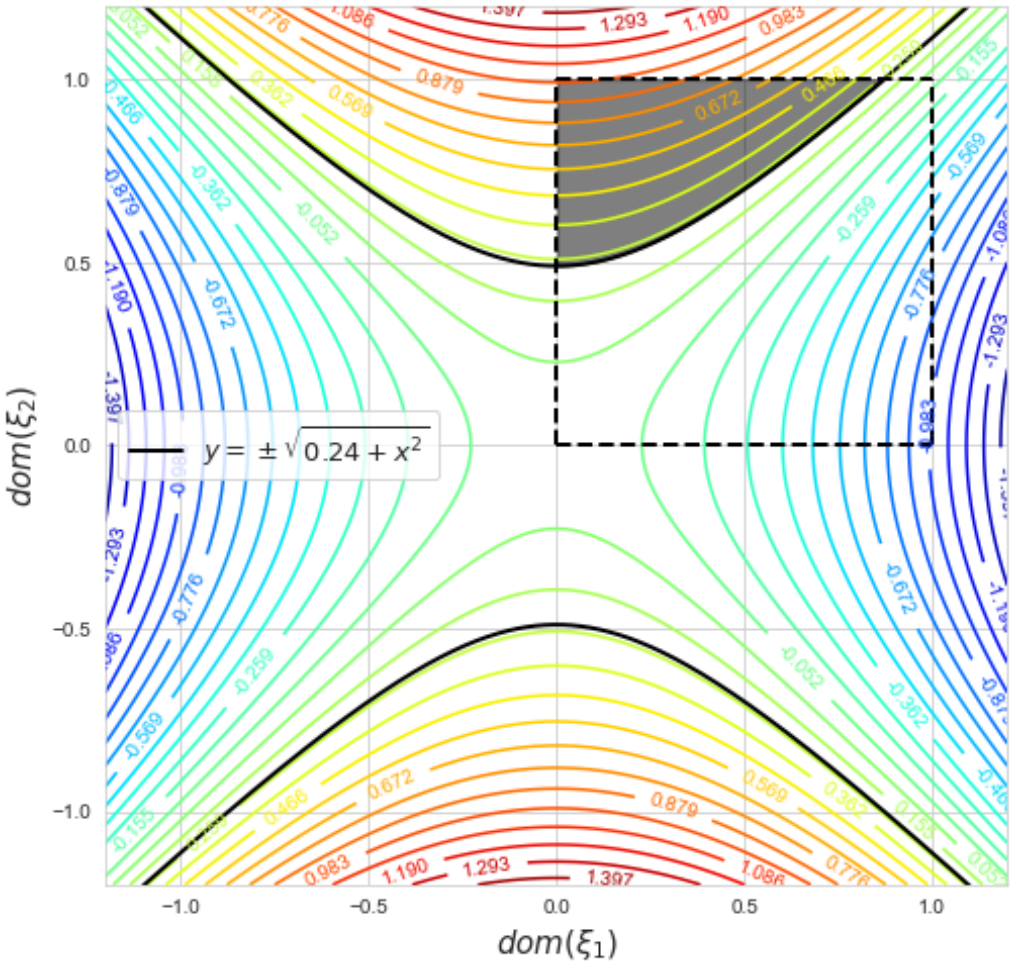

Рис. 4: Задание 3

$x_{1,\text{new}} = (0.1, \xi_1),\ x_{2,\text{new}} = (0.5, \xi_2)$

Необходимо найти $\mathbb{P}(\underbrace{0.1^2 + \xi_1^2}_{\rho^2(t, x_{1,\text{new}})} \geq \underbrace{0.5^2 + \xi_2^2}_{\rho^2(t, x_{2,\text{new}})}) = \mathbb{P}(\xi_1^2 - \xi_2^2 \geq 0.24) =$

$\int\limits_0^{\sqrt{0.76}} 1 - \sqrt{0.24 + x^2}\, dx \approx 0.275$

Задача 5

1. $V^{1/d}$

2. $0.0018^{1/d} \cdot V^{1/d}$

   $0.0018^{1/10} = 0.53,\ 0.0018^{1/10} = 0.94$

3.
```python
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
from scipy.spatial.distance import cdist

N = int(1e4)
for d in [1, 2, 10, 100]:
    x1, x2 = np.random.normal(size=(2, N, d))
    sns.distplot(cdist(XA=x1, XB=x2).ravel(), label=f'dim={d}');
plt.legend(fontsize=15);
plt.show();

```

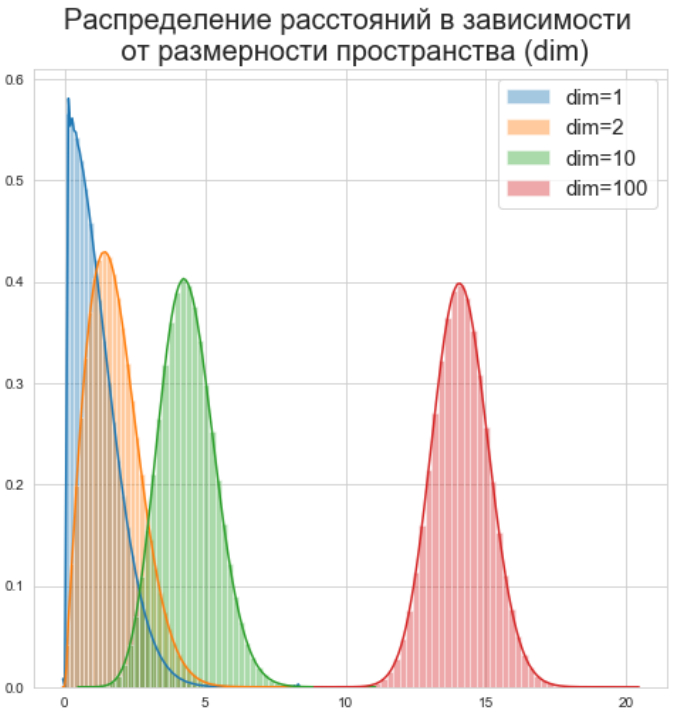

Рис. 5: Задание 2

Задача 9

3 ближайших соседа для объекта 3.5 есть: $[(3, 5), (5, 4), (1, 2)]$

$$a(x = 3.5) = \frac{5 \cdot \frac{3-1+1}{3} + 4 \cdot \frac{3-2+1}{3} + 2 \cdot \frac{3-3+1}{3}}{\frac{3-1+1}{3} + \frac{3-2+1}{3} + \frac{3-3+1}{3}} = \frac{5+4 \cdot \frac{2}{3} + 2 \cdot \frac{1}{3}}{1+\frac{2}{3}+\frac{1}{3}} = \frac{25}{6}$$

Задача 10

(Colab notebook)

Задача 13

```
positive_wv = tmp_df.loc[
    ['котлетки', "гриле", "высший", "класс"]
].mean(axis=0).values

negative_wv = tmp_df.loc[
    ['ужасный', "ресторан", "больше", "не", "придём", "сюда"]
].mean(axis=0).values

query_wv = tmp_df.loc[
    ['бургеры', "здесь", "вкусные", "атмосфера", "ужасная"]
].mean(axis=0).values

print('query_wv:', query_wv)
print('negative_wv:', negative_wv)
print('positive_wv:', positive_wv)
```

```
query_wv: [-0.138 -0.142  0.082 -0.004]
negative_wv: [ 0.12166667 -0.29333333 -0.09333333  0.04333333]
positive_wv: [-0.0375 -0.085   0.145  -0.105 ]
```

```
print('dist(query, positive):',
        round(scipy.spatial.distance.pdist(X=[query_wv, positive_wv], metric='cosine')[0], 2)
     )
print('dist(query, negative):',
        round(scipy.spatial.distance.pdist(X=[query_wv, negative_wv], metric='cosine')[0], 2)
     )
```

```
dist(query, positive): 0.32
dist(query, negative): 0.76
```

Рис. 6: Задание 13

Задача 15

(Условие)

$$\rho_{cos}(x,z) = 1 - \frac{\langle x,z\rangle}{\|x\|\cdot\|z\|} = \{\|x\| = \|z\| = 1\} = 1 - \langle x,z\rangle$$

$$\rho_{eucl}(x,z) = \|x-z\|_2^2 = (x-z)^T(x-z) = \underbrace{\langle x,x\rangle}_{\|z\|^2=1} - 2\langle x,z\rangle + \underbrace{\langle z,z\rangle}_{\|z\|^2=1} =$$

$$2\cdot(1-\langle x,z\rangle) = 2\cdot\rho_{cos}(x,z)$$

Значение Евклидовой метрики в данном случае будет однозначно соответствовать значению косинусной метрики, поэтому просадки качества не ожидается.