МИНОБРНАУКИ РФ

ФГБОУ Тверской государственный технический университет

Факультет информационных технологий

Кафедра «Программное обеспечение»

Лабораторная работа №3

Дисциплина: «Анализ больших данных»

Вариант №7

Работу выполнил: студент группы

ПИН. РИС.21.06

Олимов Авазбек.

Тверь, 2025

# Вопросы:

1. Что такое классификация?
2. Чем отличается бинарная классификация от многоклассовой?
3. Какие преимущества имеет логистическая регрессия?
4. Какую задачу решает логистическая регрессия?
5. Как исключить входные факторы (признаки) с низкой значимостью?
6. Как оценить точность классификации?
7. Как работает алгоритм KNN? Чем отличается алгоритм от логистической регрессии?
8. Что показывают критерии качества и ROC-кривая для задачи с определением пола?
9. Какие входные и выходные параметры используются в модели Логистической регрессии?
10. Что показывают критерии качества и ROC-кривая для задачи с определением доброкачественной опухоли?
11. Как построить таблицу сопряженности (confusion matrix)?

1. Что такое классификация?

Классификация — это задача машинного обучения, где алгоритм относит объекты к заранее определённым категориям (классам) на основе их признаков. Пример: определение спама в email (классы "спам"/"не спам").

2. Бинарная vs многоклассовая классификация

|  |  |  |
| --- | --- | --- |
| **Критерий** | **Бинарная** | **Многоклассовая** |
| Число классов | 2 | 3+ |
| Пример | Определение пола | Распознавание цифр 0-9 |
| Алгоритмы | LR, SVM | Softmax, RF |

3. Преимущества логистической регрессии

* Интерпретируемость: Можно анализировать вклад каждого признака
* Эффективность: Быстрое обучение даже на больших данных
* Вероятностный вывод: Даёт оценку P(y|x)
* Устойчивость к шумам в данных

4. Задача логистической регрессии

Решает задачу предсказания вероятности принадлежности к классу через сигмоидную функцию:

P(y=1|x) = 1 / (1 + e^(-wTx))

5. Исключение малозначимых признаков

Методы:

* ANOVA F-test (для линейных зависимостей)
* Mutual Information (для нелинейных)
* L1-регуляризация (LASSO) автоматически обнуляет слабые веса
* Feature Importance из деревьев решений

6. Оценка точности классификации

Основные метрики:

* Accuracy: (TP+TN)/(P+N)
* Precision: TP/(TP+FP)
* Recall: TP/(TP+FN)
* F1-score: 2\*(Precision\*Recall)/(Precision+Recall)
* AUC-ROC: Площадь под ROC-кривой

7. Алгоритм KNN vs Логистическая регрессия

KNN (k ближайших соседей):

* Непараметрический метод
* Классифицирует по большинству среди k соседей
* Чувствителен к шкале признаков
* Требует хранения всех данных

Отличия от LR:

* LR строит явную модель, KNN — instance-based
* LR быстрее на предсказании
* KNN лучше для сложных границ классов

8. ROC-кривая для определения пола

Для задачи бинарной классификации (М/Ж):

* Ось X: FPR = FP/(FP+TN)
* Ось Y: TPR = Recall
* AUC 0.9 означает: 90% chance model ranks random male higher than random female
* Оптимальный порог выбирается по точке близкой к (0,1)

9. Параметры логистической регрессии

Входные:

* Матрица признаков X (n\_samples × n\_features)
* Вектор меток y ∈ {0,1} для бинарной

Выходные:

* Веса w для каждого признака
* Смещение (bias) b
* Вероятности P(y=1|x)

10. ROC для опухолей (добро/зло)

В медицинских задачах:

* Recall важнее precision (меньше ложноотрицательных)
* FPR < 10% обычно требуется
* AUC > 0.85 — хорошая модель
* Кривая должна доминировать над бейзлайном (диагональю)

# Задача 1. Разбор данных из Kaggle

1. Загрузить один из наборов данных. Загружаются все файлы по ссылке.
2. Провести разведочный анализ данных, ответив на следующие вопросы:
3. Сколько строк в датафрейме, сколько столбцов

**117564 строк**

**10 столбцов**

1. Сколько места занимает датафрейм в оперативной памяти

**Память, занимаемая датафреймом: 9405252 байт**

memory\_usage = data.memory\_usage(deep=True).sum()

7print(f"Память, занимаемая датафреймом: {memory\_usage} байт")

1. Для каждой интервальной переменной подсчитать следующее - мин, медиана, среднее, макс и персентили 25, 75

A screen shot of a computer screen

AI-generated content may be incorrect.

1. Для каждой категориальной переменной рассчитать моду и сколько раз мода встречается в данных

A screen shot of a computer

AI-generated content may be incorrect.