Лабораторные работы по курсам:
"Интеллектуальный анализ данных",
“Методы машинного обучения”,
"Математические методы анализа данных",
“Теория статистических решений”

Для допуска к экзамену (зачёту) минимальное требование (на удовлетворительную оценку) - выполнить три лабораторных работы.

Получение оценки за экзамен “автоматом” не предусмотрено.

Возможна сдача экзамена досрочно. Для этого необходимо выполнить курсовую работу и от трёх (для оценки “удовлетворительно”) до пяти или шести (для оценки “отлично”) лабораторных работ.

Каждая работа оценивается дифференцированно, поэтому итоговая оценка зависит как от числа работ, так и от качества их выполнения.

Досрочный экзамен принимается одновременно с защитой курсовой работы и состоит в ответе на вопросы по программе лекций. Итоговая оценка определяется как минимум из оценок за практику и за теоретические вопросы.

Для сдачи экзамена в обычном формате выполнение курсовой работы не обязательно.

Сроки сдачи (защиты): 01.10, 22.10, 12.11, 02.12, 16.12. Если число сданных работ меньше числа прошедших дедлайнов, то оценка за (очередную) работу снижается.

Лабораторная работа 1.

(Квази) линейные методы классификации.

Задание.

1. Загрузить таблицу Iris из репозитория UCI (можно из любого доступного источника: UCI, sklearn, kaggle). Поместить данные в объект DataFrame библиотеки pandas. Вычислить корреляции между признаками на всей таблице и отдельно по классам (использовать pd.groupby).
Визуализировать распределения классов на всех парах переменных.

2. Выбрать две переменные. Построить и визуализировать (показать разделяющую кривую) решения методами линейный и квадратичный дискриминант, логистическая регрессия, SVM (линейное и квадратичное ядро). Пример подходящей визуализации есть в sklearn.

В следующих заданиях можно оставить только два (наименее разделимых) класса.

3. Построить линейный дискриминант на всех переменных. Визуализировать ответы алгоритма (выделить, например, цветом) и классы объектов (выделить, например, формой маркеров) во всех двумерных подпространствах.

4. На двух переменных из п2 вычислить квадратичную разделяющую функцию непосредственно по оценкам ковариационных матриц и средних (самостоятельно

реализовать метод, не используя готовый). Визуализировать её и сравнить с решением из п2.

Лабораторная работа 2.

«Наивный байесовский» классификатор.

Задание.

1. Загрузить таблицу Mushroom из репозитория UCI (любого источника) в объект DataFrame.
Вычислить распределение значений категориальных признаков по классам (использовать pd.groupby).
Визуализировать распределения.

2. Построить решающую функцию по каждой переменной на основе частот. Вычислить точность каждого решения (на той же обучающей выборке). Найти наиболее информативную переменную (с минимальным числом ошибок).

3. Построить «наивный» байесовский классификатор из sklearn. Оценить точность.

4. Самостоятельно реализовать метод, не используя готовый. Сравнить полученное решение с библиотечным. Добавить регуляризатор в оценки частот.

5. Применить метод логистической регрессии, использовав в качестве переменных оценки вероятностей, подвергнутые обратному логистическому преобразованию.

Лабораторная работа 3.

Деревья решений. Ансамбли решающих деревьев.

1. Выбрать подходящую таблицу данных. Построить и визуализировать дерево решений.

2. Применить метод градиентного бустинга. Вычислить значимость переменных. Выдать список построенных деревьев.

3. Построить зависимость качества решения (на обучении и скользящем контроле) от числа вершин дерева.

4. Для метода градиентного бустинга построить зависимость качества решения (на обучении и скользящем контроле) от числа деревьев.
Для разной глубины дерева нужно построить несколько зависимостей качества от числа деревьев, чтобы найти оптимальную комбинацию этих параметров.

5. Выполнить предыдущий пункт для случайного леса.

Лабораторная работа 4.

Задача восстановления зависимостей. Манипулирование признаками. Сокращение размерности.

1. Выбрать подходящую таблицу данных (должна содержать числовые и категориальные переменные). Временно убрать категориальные признаки. Построить линейную регрессию.

2. Построить решение методом бустинга. Сравнить с линейной регрессией.

3. Применить one hot и target encoding для категориальных признаков (взять данные, где такие признаки есть). Сравнить точность.

4. Визуализировать объекты (не обязательно для той же таблицы) в пространстве главных компонент.

Лабораторная работа 5.

Критерии качества. Кривая ошибок. Оценивание качества.

1. Подобрать таблицу данных с несбалансированными классами. Решить задачу классификации любым подходящим методом.

2. Вычислить точность, полноту, специфичность.

3. Построить кривую ошибок и найти площадь под ней. Построить для сравнения кривую "точность-полнота".

4. Разбить данные на обучающую и контрольную выборки. Построить ROC- кривую для каждой из подвыборок. Построить ROC- кривую на основе кроссвалидации.

5. Исследовать влияние выбора критерия обучения на AUC. Один из критериев — log loss, ещё один или два — на выбор. Как вариант: сравнить AdaBoost с градиентным бустингом.

Лабораторная работа 6.

Исследование эффективности методов классификации с помощью статистического моделирования.

Цель: исследовать статистические свойства эмпирического риска, оценки скользящего экзамена и вероятности ошибочной классификации на синтезированных данных.

Ход работы.

1. Придумать двумерную вероятностную модель для двух классов. Это может быть смесь нормальных распределений (с различными параметрами) с числом компонент, большим чем число классов.

2. Задать параметры: размер обучающей выборки (порядка 100 объектов), число разбиений кроссвалидации. Выбрать метод классификации и задать его параметры.

3. Повторять шаги 4–7 заданное число раз (50–100).

4. Сгенерировать обучающую выборку заданного размера.

5. Построить решающую функцию. Вычислить эмпирический риск (число ошибок на обучении).

6. Сгенерировать контрольную выборку достаточно большого размера (больше 10000 объектов). Вычислить оценку вероятности ошибочной классификации.

7. Вычислить оценку вероятности ошибочной классификации методом скользящего экзамена (на исходной обучающей выборке).

8. Результаты свести в таблицу

номер выборки	эмпирический риск	скользящий экзамен	контрольн. выб.

9. Вычислить средние и стандартные отклонения по каждому столбцу.

10. Провести аналогичное моделирование, изменив вероятностную модель, или метод классификации, или параметры метода.

11. Сделать выводы, насколько выбранный метод классификации и его параметры соответствуют сложности модели и объёму выборки.

12. Факультативно: вычислить смещение и разброс (bias-variance decomposition).

Лабораторная работа 7.

Ансамблевые методы: stacking, blending. Оценка out-of-fold.

1. Подобрать подходящую таблицу данных для задачи классификации. Изучить статью А.Г. Дьяконова про стэкинг.

2. Выбрать три различных метода классификации. Применить их к задаче по отдельности. Оценить качество.

3. Реализовать и применить усреднение (*blending*), в т.ч. взвешенное. Оценить качество.

4. Реализовать и применить стэкинг. Оценить качество.

5. Исследовать влияние смещённости ответов базовых методов на обучающей выборке (если базовые методы и верхнеуровневый метод обучать на одной выборке).

Лабораторная работа 8.

Нейронные сети на табличных данных.

1. Подобрать подходящую задачу (таблицу данных) классификации или регрессии.

2. Построить решение на основе полносвязной нейросети с несколькими слоями.

3. Провести подбор параметров архитектуры (число и размер слоёв), функций активации и количества эпох.

4. Провести эксперимент с изменением функции потерь (применить MSE для классификации или logloss для регрессии).

5. Построить решение методом boosting. Сравнить качество с решением нейросети.

Курсовая работа.

Решение реальной задачи анализа данных (kaggle, UCI). Альтернатива: расширенный вариант работы 6.

Работа выполняется в бригадах (от 1 до 3 человек), на каждую бригаду задание индивидуальное, согласовывается с преподавателем.

1. Выбрать исходные данные (kaggle, UCI или любые другие реальные данные), сформулировать задачу.

2. Решить задачу подходящим методом. Обратить внимание на подбор параметров (должно быть экспериментальное обоснование выбора параметров).

3. Обосновать выбор метода. Если в бригаде более 1 человека, то решить задачу разными методами.

4. Оценить качество решения. Оценить точность оценки качества.

5. Представить отчёт (можно в форме notebook с комментариями).