«Машинное обучение и анализ данных». Тематическое содержание дисциплины (2023/2024)

Елена А. Харченко
elenakhaa@yandex.ru

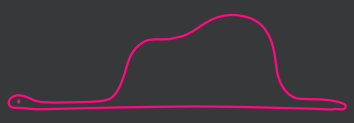
1. Методы машинного обучения для решения задач информационной безопасности: реляционная модель данных, векторное пространство, метрика (расстояние), большие данные, **примеры**.
2. Предварительное изучение одномерных данных: переменная (и случайная) величина, генеральная совокупность, выборка, математическое ожидание, дисперсия, среднее квадратическое отклонение, правило «трёх сигм» (неравенство Чебышёва), гистограмма (и плотность вероятности), нормальное распределение (кривая Гаусса), мода.
3. Предварительная обработка многомерных данных: количественные и качественные признаки, нормализация количественных признаков, преобразование качественных признаков, удаление выбросов, исключение неинформативных признаков, выравнивание объёмов классов.
4. **Классификация**, метод k ближайших соседей: машинное обучение с учителем, расстояние Евклида, алгоритм построения и порядок применения модели классификации.
5. **Классификация**, метод дерева решений: энтропия, алгоритм построения и порядок применения модели классификации.
6. Кросс-валидация модели классификации: метод Монте-Карло, алгоритмы и применение кросс-валидации.
7. **Классификация**, наивный метод Баеса: условная вероятность, теорема Баеса, алгоритм построения и порядок применения модели классификации.
8. **Классификация**, метод опорных векторов: скалярное произведение векторов, уравнение гиперплоскости, алгоритмы построения и порядок применения моделей классификации.
9. **Классификация**, многослойный персептрон (нейронная сеть прямого распространения): производная и градиент функции в точке, модель нейрона, функция активации, полновесная нейронная сеть, дельта-правило, алгоритм построения и порядок применения модели классификации.
10. Метод главных компонент: базис векторного пространства, корреляционная матрица, собственный вектор, алгоритм метода и **варианты** его использования.
11. **Кластеризация**, метод k средних: машинное обучение без учителя, центр масс, алгоритм построения и порядок применения модели кластеризации.
12. Подбор оптимального числа кластеров, методы локтя и силуэта: метод наименьших квадратов, алгоритмы методов подбора числа кластеров и их применение.
13. **Кластеризация**, метод DBSCAN: алгоритм построения и порядок применения модели кластеризации.
14. **Кластеризация**, метод агломеративной кластеризации: дендрограмма, алгоритм построения и порядок применения модели кластеризации.
15. **Кластеризация**, сеть Кохонена: алгоритм построения и порядок применения модели кластеризации.
16. **Кластеризация**, карта Кохонена: алгоритм построения и порядок применения модели кластеризации.
17. Удаление выбросов и выявление аномалий в многомерных данных: основные приёмы.

Рекомендуемая литература

[1]. Дайзенрот М.П., Альдо Ф.А., Чен С.О. *Математика в машинном обучении.*

[2]. *Учебник по машинному обучению.* – URL: education.yandex.ru/handbook/ml

[3]. *Machine Learning in Python.* – URL: scikit-learn.org/stable



«МАШИННОЕ ОБУЧЕНИЕ И АНАЛИЗ ДАННЫХ».
Лабораторный практикум. Работа 0

Задачи на обработку результатов эксперимента

1. Комиссия изучала состояние борьбы с преступностью в регионе. Случайным образом было выбрано 20 районов [не меньше 36, на самом деле]. Представленые данные о числе раскрытых убийств по ним: 11, 6, 12, 1, 3, 1, 6, 20, 10, 1, 1, 3, 3, 1, 23, 11, 3, 6, 10, 3. Охарактеризуйте по этим данным весь регион.

   - Найдите среднее значение и дисперсию [в электронной таблице, OpenOffice Calc или др.]:

x_i	1	3	6	10	11	12	20	23	Сумма
m_i	5	5	3	2	2	1	1	1	?
$p_i = m_i/\boldsymbol{n}$	?	?	?	?	?	?	?	?	?
$p_i x_i$	?	?	?	?	?	?	?	?	?
$x_i - \boldsymbol{\mu}$	?	?	?	?	?	?	?	?	?
$(x_i - \mu)^2$	?	?	?	?	?	?	?	?	?
$p_i(x_i - \mu)^2$	?	?	?	?	?	?	?	?	?

     Здесь m_i и p_i – соответственно частоты и относительные частоты значений переменной величины X (зарплаты).

   - Найдите среднее квадратическое отклонение и границы доверительного интервала.
   - Найдите выбросы.

2. Средняя месячная зарплата за год в некоторых условных единицах каждого из пятидесяти случайно отобранных работников хозяйства такова: 317, 304, 230, 285, 290, 320, 262, 274, 205, 180, 234, 221, 241, 270, 257, 290, 258, 296, 301, 150, 160, 210, 235, 308, 240, 370, 180, 244, 365, 130, 170, 250, 370, 267, 288, 231, 253, 315, 201, 256, 279, 285, 226, 367, 247, 252, 320, 160, 215, 350.

   - Составьте по этим данным интервальный ряд и постройте гистограмму [столбчатую диаграмму по выделенным значениям]:

Разряды	$[130, 170]$	$[170, 210]$	$[210, 250]$	$[250, 290]$	$[290, 330]$	$[330, 370]$	Сумма
x_i	150	190	230	270	310	350	–
m_i	4,5	5	12	14,5	9	5	?
p_i	?	?	?	?	?	?	?
$h_i = p_i/\Delta$	?	?	?	?	?	?	–

     Здесь Δ – длина разряда (40 в нашем случае), h_i – плотность частоты.

   - Определите долю значений величины X, принадлежащих отрезку $[200, 300]$.
   - Найдите интервал наиболее вероятных значений.

3. Хронологические данные о загруженности процессора компьютера (%CPU) за некоторый период следующие: 1.14, 1.10, 1.09, 1.08, 1.08, 1.08, 1.15, 1.13, 1.09, 1.06, 1.04, 1.03, 1.01, 1.00, 0.99, 0.98, 0.97, 0.95, 0.93, 0.92, 0.90, 0.88, 0.88, 0.87, 0.85, 0.89, 0.86, 0.84, 0.81, 0.79, 0.78, 0.76, 0.74, 0.72, 0.71, 0.71, 0.69, 0.68, 0.67, 0.65, 0.65, 0.64, 0.64, 0.64, 0.67, 0.65, 0.70, 0.66, 0.63, 0.61, 0.61, 0.64, 0.62, 0.62, 0.60, 0.57, 0.57, 0.58, 0.58, 0.67, 0.64, 0.62, 0.62, 0.60, 0.60, 0.64, 0.65, 0.65, 0.68, 0.68, 0.76, 0.76, 0.77, 0.78, 0.84, 0.85, 0.96, 0.97, 0.96, 0.99, 1.11, 1.22, 1.17, 1.16, 1.66, 1.60, 1.68, 2.03, 2.02, 1.79, 1.65, 1.52, 1.47, 1.74, 1.64, 1.86, 1.96, 2.11, 2.20, 2.27, 1.93, 1.73, 1.88, 1.75, 1.58, 1.57, 1.79, 1.69, 1.56, 1.47, 1.40, 1.38, 1.45, 1.37, 1.68, 1.54, 1.47, 1.39, 1.32, 1.72, 1.56, 1.42, 1.41, 1.32, 1.27, 1.24, 1.20, 1.18, 1.15, 1.14, 1.17, 1.15, 1.12, 1.11, 1.44, 1.73, 1.53, 1.37, 1.33, 1.23, 1.19, 1.16, 1.14, 1.11, 1.07, 1.05, 1.10, 1.08, 1.05, 1.12, 1.19, 1.13, 1.08, 1.04, 1.09, 1.04, 1.00, 0.97, 0.94, 0.93, 0.93, 0.93, 0.99, 0.96, 0.98, 0.94, 0.91, 0.91,

0.90, 0.86, 0.91, 0.86, 0.84, 0.81, 0.79, 0.79, 0.78, 0.76, 0.74, 0.73, 0.78, 0.74, 0.72, 0.70, 0.68, 0.65, 0.65, 0.64, 0.70, 0.74, 0.78, 0.79, 0.74, 0.72, 0.69, 0.71, 0.66, 0.63, 0.61, 0.61, 0.70, 0.67, 0.65, 0.65, 0.66, 0.65, 0.65, 0.65, 0.74, 0.73, 0.71, 0.69, 0.68, 0.71, 0.73, 0.76, 0.84, 0.82, 0.82, 0.83, 0.86, 0.88, 0.90, 0.92, 1.02, 1.03, 1.32, 1.41, 1.40, 1.36, 1.36, 1.30, 1.29, 1.27, 1.31, 1.32, 1.35, 1.80, 2.28, 2.01, 1.87, 1.72, 2.03, 1.80, 2.05, 2.18, 1.92, 1.78, 1.64, 1.54, 1.53, 1.58, 1.87, 1.74, 1.72, 1.70, 1.59, 1.85, 1.80, 1.65, 1.51, 1.47, 1.44, 1.42, 1.72, 1.59, 1.58, 1.50, 1.42, 1.43, 1.38, 1.32, 1.31, 1.33, 1.34, 1.39, 1.37, 1.31, 1.59, 1.93, 1.79, 1.62, 1.75, 1.60, 1.59, 1.46, 1.40, 1.35, 1.28, 1.23, 1.24, 1.29, 1.25, 1.28, 1.31, 1.27, 1.41, 1.41, 1.67, 1.48, 1.35, 1.27, 1.24, 1.20, 1.23, 1.22, 1.15, 1.14, 1.21, 1.16, 1.11, 1.09, 1.07, 1.04, 1.02, 1.08, 1.04, 0.99, 0.95, 0.92, 0.88, 0.85, 0.84, 0.84, 0.95, 0.88, 0.85, 0.94, 0.87, 0.81, 0.78, 0.75, 0.73, 0.70, 0.75, 0.71, 0.69, 0.68, 0.69, 0.68, 0.68, 0.68, 0.69, 0.75, 0.77, 0.72, 0.70, 0.70, 0.70, 0.79, 0.77, 0.77, 0.77, 0.79, 0.82, 0.84, 0.87, 0.99, 1.02, 1.02, 1.04, 1.06, 1.10, 1.48, 1.51, 1.53, 1.53, 1.53, 1.79, 1.66, 1.86, 2.04, 1.87, 1.84, 2.45, 2.48, 2.20, 1.92, 2.13, 1.87, 2.13, 2.23, 2.03, 2.14, 1.92, 2.15, 1.95, 1.75, 1.93, 2.13, 2.17, 2.24, 2.40, 2.11, 2.18, 1.97, 2.10, 2.53, 2.55, 2.45, 2.20, 2.26, 1.93, 1.77, 2.07, 1.88, 2.34, 2.44, 2.19, 2.04, 1.98, 2.07, 2.48, 2.45, 2.40, 2.41, 2.16, 1.93, 2.14, 1.96, 2.19, 1.97, 2.15, 1.94, 1.81, 1.67, 1.61, 1.58, 1.92, 2.06, 1.87, 1.81, 1.67, 1.56, 1.56, 1.57, 1.57, 1.56, 1.53, 1.56, 1.48, 1.48, 1.49, 1.46, 1.37, 1.32, 1.28, 1.22, 1.26, 1.20, 1.17, 1.23, 1.18, 1.14, 1.10, 1.05, 1.08, 1.03, 1.08, 1.01, 0.95, 0.91, 0.86, 1.00, 1.00, 0.93, 0.88, 0.84, 0.80, 0.88, 0.82, 0.78, 0.83, 0.78, 0.74, 0.73, 0.71, 0.69, 0.69, 0.68, 0.74, 0.72, 0.69, 0.73, 0.71, 0.71, 0.70, 0.69, 0.69, 0.69, 0.70, 0.78, 0.78, 0.76, 0.79, 0.90, 0.89, 0.93, 0.94, 1.02, 1.02, 1.03, 1.06, 1.08, 1.11, 1.15, 1.20, 1.28, 1.26, 1.30, 1.38, 1.39, 1.42, 1.41, 1.39, 1.45, 1.45, 1.65, 1.59, 1.51, 1.47, 1.73, 1.61, 1.54, 1.49, 1.50, 1.42, 1.46, 1.46, 1.49, 1.75, 1.89, 1.82, 2.06, 1.93, 2.14, 1.92, 2.03, 2.08, 2.15, 2.01, 1.83, 1.77, 1.68, 1.72, 1.70, 2.31, 2.03, 1.87, 1.74, 1.62, 1.54, 1.46, 1.46, 1.53, 1.77, 2.01, 1.84, 1.76, 1.62, 1.55, 1.51, 1.57, 1.53, 1.53, 1.55, 1.55, 1.52, 1.48, 1.49, 1.47, 1.42, 1.35, 1.31, 1.36, 1.33, 1.56, 1.56, 1.47, 1.68, 1.65, 1.88, 2.33, 2.16, 1.91, 1.71, 1.59, 1.48, 1.57, 1.59, 1.51, 1.53, 1.47, 1.45, 1.52, 1.45, 1.78, 1.77, 1.67, 1.67, 1.71, 1.58, 1.86, 1.66, 1.61, 1.52, 1.43, 1.34, 1.30, 1.62, 1.53, 0.04, 0.07, 0.03, 0.07, 0.03, 0.04, 0.06, 0.05, 0.07, 0.09, 0.04, 0.07, 0.03, 0.05, 0.05, 0.05, 0.03, 0.07, 0.06, 1.10, 2.13, 2.83, 2.44, 2.15, 2.05, 1.92, 2.17, 2.02, 1.60, 1.29, 1.31, 1.41, 2.22, 1.98, 1.85, 2.00, 2.12, 2.23, 2.50, 2.50, 2.39, 2.49, 2.18, 2.50, 2.55, 2.23, 2.11, 2.17, 2.15.

- Постройте гистограмму[1] [в Jupyter Notebook].
- Вычислите вероятность того, что загруженность процессора примет значение в диапазоне от до 2.0 до 2.5.
- Найдите границы доверительного интервала и выявите выбросы.
- Выделите области сгущения значений (по локальным минимумам) и математически опишите их (по правилу «трех сигм»). Выявите выбросы.
- Последовательно пронумеруйте значения загруженности процессора. По получившимся точкам вида (x, y), где x – номер значения и y – само значение, постройте график [представляется в виде сглаженной или ломаной линии, соединяющей точки].

[1] За оптимальное число интервалов рекомендуется принять нечетное число $k \in (0.55m^{0.4}, 1.25m^{0.4})$.

«МАШИННОЕ ОБУЧЕНИЕ И АНАЛИЗ ДАННЫХ».
Лабораторный практикум. Работа 1[1]

ВАЖНО. Результаты работы требуется документировать и представлять в формате PDF (лаконично, в свободной форме): группа, ФИО, источник данных, номер задания, фрагменты программного кода с пояснениями и комментариями, диаграммы, результаты и т.д. Рекомендуется использовать систему LaTeX: Overleaf или TeX Live. Работы следует направлять на адрес: elenakhaa@yandex.ru.

Задание 1a [до 14 февраля, коэффициент сложности равен 2] Сегментировать по яркости пиксели изображения, представленного файлом `plane.png` (или `plane_294x182.html`).

1. Преобразовать цвет каждого пикселя по правилу:

$$R'_i = G'_i = B'_i = Y_i,$$

   где яркость пикселя

$$Y_i = 0,299R_i + 0,587G_i + 0,114B_i.$$

2. По значениям яркости пикселей *программно* построить гистограмму и по ее характеристикам выделить области качественной однородности пикселей (их границами будут локальные минимумы гистограммы).

3. Заменить цвета пикселей, соответствующих первой по частоте области однородности [корпус самолета Су-57], на красный или обвести эту область минимальным прямоугольником. Результат (изображение) представить в виде PNG-файла (или HTML-таблицы).

4. В виде CSV-файла сформировать размеченный набор данных, описывающий классы (категории) пикселей изображения. За признаки (поля, атрибуты) следует принять R, G, B и *label*, где *label* – порядковый номер области сгущения значений яркости пикселей.

Задание 1b [14 февраля, коэффициент сложности равен 1]. С помощью последовательного применения гистограмм программно сегментировать каждое изображение `mask*.jpeg` так, чтобы область лица была выделена минимальным прямоугольником.

[1] Экзаменационная оценка определяется суммой баллов, набранных за семестр, включая экзаменационное задание (до 19 баллов): «удовлетворительно» – от 41 до 60 баллов; «хорошо» – от 61 до 80 баллов; «отлично» – от 81 до 100 баллов. Все работы подлежат защите. Каждая работа оценивается по шкале от 0 до 3 баллов, полученный результат умножается на коэффициент сложности и идет в общий зачет.

«МАШИННОЕ ОБУЧЕНИЕ И АНАЛИЗ ДАННЫХ».
Лабораторный практикум. Работа 2

Задание 2a [до 21 февраля, коэффициент сложности равен 1]. Подберите *размеченный* набор данных[1], отвечающий определению «большие» и пригодный для решения задач информационной безопасности, и произвести его предобработку:

1. Исключите строки с пропусками [иногда пропуски – это значения].
2. Исключите неинформативные признаки (по ним объекты классов неразличимы): у количественного признака значения однородны (одна мода/вершина у плотности распределения), качественный признак принимает тоько одно значение. Для этого для количественных признаков предварительно постройте гистограммы, для качественных – диаграммы частот [или сгенерируйте отчет с помощью библиотеки Pandas Profiling].
3. Преобразуйте качественные признаки к количественному виду.
4. Нормализуйте/стандартизируйте количественные признаки.
5. По каждому количественному признаку исключите строки с выбросами – значениями, выходящими за границы доверительных интервалов (лучше сначала отметить нежелательные значения каждого признака как `none` и только в конце удалить их) [проще после стандартизации].
6. Выровняйте классы по объему (объектов каждого класса должно быть примерно поровну).

Задание 2b [22 февраля, коэффициент сложности равен 1]. Осуществите типовую предобработку данных датасета `article.csv`.

[1]Примеры источников данных: Canadian Institute for Cybersecurity, UC Irvine, Kaggle. Пример хорошего датасета: DDOS attack SDN Dataset.

«МАШИННОЕ ОБУЧЕНИЕ И АНАЛИЗ ДАННЫХ».
Лабораторный практикум. Работа 3

Задание 3 [до 28 февраля, коэффициент сложности равен 3] По набору данных из задания 2 на основе метода k ближайших соседей разработайте классификатор:

1. Создайте, обучите и оцените качество классификатора [корректная предобработка данных подразумевается].

2. Реализуйте метод, проверяющий значения признаков классифицируемого объекта на соответствие областям допустимых значений признаков и выявляющую аномальные объекты [такие объекты не должны подаваться на вход классификатору].

3. Проиллюстрируйте варианты использования классификатора [с выводом вероятностей решений]. Подаваемые на вход классификатору данные следует предобрабатывать по той же схеме, что и обучающие/тренировочные данные.

«МАШИННОЕ ОБУЧЕНИЕ И АНАЛИЗ ДАННЫХ».
Лабораторный практикум. Работа 4

Задание 4 [до 13 марта, коэффициент сложности равен 2] По набору данных из задания 2 на основе метода дерева решений разработайте классификатор:

1. Создайте, обучите и оцените качество классификатора [корректная предобработка данных подразумевается]. Визуализировать дерево решений.

2. Реализуйте метод, проверяющий значения признаков классифицируемого объекта на соответствие областям допустимых значений признаков и выявляющую аномальные объекты [такие объекты не должны подаваться на вход классификатору].

3. Проиллюстрируйте варианты использования классификатора [с выводом вероятностей решений]. Подаваемые на вход классификатору данные следует предобрабатывать по той же схеме, что и обучающие/тренировочные данные.

«МАШИННОЕ ОБУЧЕНИЕ И АНАЛИЗ ДАННЫХ».
Лабораторный практикум. Работа 5

Задание 5 [до 27 марта, коэффициент сложности равен 3] По набору данных из задания 2 на основе наивного метода Байеса разработайте классификатор:

1. Создайте, обучите и оцените качество классификатора [предобработка данных подразумевается].

2. Реализуйте метод, проверяющий значения признаков классифицируемого объекта на соответствие областям допустимых значений признаков и выявляющую аномальные объекты [такие объекты не должны подаваться на вход классификатору].

3. Проиллюстрируйте варианты использования классификатора [с выводом вероятностей решений]. Подаваемые на вход классификатору данные следует предобрабатывать по той же схеме, что и обучающие/тренировочные данные.

«МАШИННОЕ ОБУЧЕНИЕ И АНАЛИЗ ДАННЫХ».
Лабораторный практикум. Работа 6

Задание 6 [до 10 апреля, коэффициент сложности равен 2] По набору данных из задания 2 на основе метода опорных векторов разработайте классификатор:

1. Создайте, обучите и оцените качество классификатора [предобработка данных подразумевается].

2. Реализуйте метод, проверяющий значения признаков классифицируемого объекта на соответствие областям допустимых значений признаков и выявляющую аномальные объекты [такие объекты не должны подаваться на вход классификатору].

3. Проиллюстрируйте варианты использования классификатора [с выводом вероятностей решений]. Подаваемые на вход классификатору данные следует предобрабатывать по той же схеме, что и обучачающие/тренировочные данные.

«МАШИННОЕ ОБУЧЕНИЕ И АНАЛИЗ ДАННЫХ».
Лабораторный практикум. Работа 7

Задание 7.1 [до 17 апреля, коэффициент сложности равен 1] С помощью кросс-валидации подберите наилучшее параметры классификаторов из работ 3-6 и выберите из них наилучшую модель классификации для рабочих данных.

Задание 7.2 [до 30 апреля, коэффициент сложности равен 2] Выполните задания 3-6, спроецировав данные в пространство главных компонент. Сравните результаты с предыдущими.

«МАШИННОЕ ОБУЧЕНИЕ И АНАЛИЗ ДАННЫХ».
Лабораторный практикум. Работа 8

Задание 8 [до 8-15 мая, коэффициент сложности равен 3] Примените метод главных компонент к данным из задания 3:

1. Визуализируйте данные в пространстве: 1) первых двух главных компонент; 3) первых трех главных компонент.

2. Постройте тепловую карту по координатам первых трех главных компонент.

3. По координатам первых трех главных компонент выявите наименее информативные исходные признаки[1].

[1]Малоинформативные признаки дадут нулевые или почти нулевые вклады во всех трех первых главных компонентах.

«МАШИННОЕ ОБУЧЕНИЕ И АНАЛИЗ ДАННЫХ».
Лабораторный практикум. Работа 9

Задание 9 [до 4 июня, коэффициент сложности равен 2] С помощью метода главных компонент оценить сходство документов по содержанию:

1. Выработать словарь терминов (не менее ста), характерных для литературы по информационной безопасности. Например: «криптосистема», «шифр», «ключ», «шифртекст» «шифрование», «хеширование», «Магма», «Кузнечик», «DES», «AES», «blockchain», «блокчейн», «СОРМ» и т.д. Если термин состоит из нескольких слов, в словарь их нужно внести по одному (без предлогов, сохраняя склонения по падежам), например: «персональные», «данные», «цифровая», «подпись», «центр», «обработки», «данных», «система», «технических», «средств», «обеспечения», «функций», «оперативно-розыскных», «мероприятий» и т.д.

2. Собрать для анализа банк тематических TXT-документов (не менее двухсот): они могут содержать тексты статей, правовых документов, web-страниц и др.

3. Составить датасет вида:

	word_1	word_2	...	word_n
document_1			...	
document_2			...	
...	...	...	...	...
document_m			...	

   На пересечении i-ой строки и j-ого столбца нужно указать, сколько раз j-ое слово встречается в i-ом документе.

4. Применить к данным метод главных компонент и визуализировать данные в пространстве первых двух главных компонент: рядом с каждой точкой указать номер документа, от начача координат до каждой точки провести отрезок (радиус-вектор).

5. Реализовать метод, вычисляющий расстояние между двумя произвольными документами (по номерам). За расстояние между документами следует принять косинус угла между соответствующими радиус-векторами в пространстве первых двух главных компонент.

6. Как изменится задача, если строки и столбцы датасета поменять местами?