Оглавление

ЛАБОРАТОРНЫЙ ПРАКТИКУМ (общие замечания)

Целью данного лабораторного практикума является изучение основ нейросетевого моделирования и обработки данных.

1. Требования к разрабатываемому программному обеспечению

1.1. Поставленная задача должна быть решена согласно выданного задания. Решение не своего варианта недопустимо.
1.2. Исходные данные должны быть представлены в виде xls файлов
1.3. Не должно быть ни каких ограничений на размеры файла с исходными данными
1.4. Все полученные результаты должны быть обоснованы
1.5. В обязательном порядке должна быть оценена степень адекватности решения (погрешность)

2. Требования к отчёту

2.1. Каждая лабораторная работа должна быть снабжена отчётом. Без наличия отчёта, удовлетворяющего ниже изложенным требованиям, работа к защите не допускается.
2.2. Отчёт по лабораторной работе должен быть загружен на учебный портал
2.3. Отчёт в обязательном порядке должен содержать:
  2.3.1. Условие задания
  2.3.2. Ф.И.О. выполнившего задание студента
  2.3.3. Описание процесса валидации (верификации), показывающее, что поставленная задача решена в полном объёме и верно
  2.3.4. Описание процесса вычислительного эксперимента, по результатам которого необходимо сделать выводы
  2.3.5. В результате выполнения работы должны быть получены графики, отображающие решение задачи (не менее 7 по каждой работе)

ВАРИАНТЫ ЗАДАНИЙ

1. Лабораторная работа №1. Предварительная обработка данных и их графическое отображение.

Общая часть: разработать программное обеспечение для решения задачи предварительной обработки данных и построения уравнений регрессии.

Исходные данные: набор данных в формате csv.

Необходимо: провести предварительную обработку и анализ данных с целью выявления значимых признаков. На основе выявленных признаков необходимо подобрать уравнение регрессии, которое с минимальной ошибкой выполняет заданный прогноз.

Ход выполнения работы:

1. Загрузить исходные данные согласно варианта из csv-файлов
2. На основании считанных данных сформировать DataFrame
3. Провести анализ на полноту данных
4. Построить гистограммы и графики плотности распределения по каждому признаку в исходных данных, при этом уметь изменять цвета, тип графика и строить графики на одном поле.
5. Согласно заданию, выбрать несколько параметров и построить для них распределения на одном графике
6. Вычислить математическое ожидание и дисперсию
7. Основываясь на решаемой задаче, визуализировать совместное распределение отдельных признаков нескольких переменных
8. Выполнить агрегирование данных по каждому из признаков (или по значению поля, например, по месяцу или году при анализе акций) и визуализировать с использованием различных функций (barplot, countplot, boxplot, violinplot, stripplot, swarmplot и др.**)**, при этом уметь пояснить, что изображено на графиках
9. Построить матрицу диаграмм рассеивания, тепловую карту и матрицу корреляции (scatterplotmatrix, heatmap, plotcorr)
10. Провести анализ на основе построенных графиков
11. Согласно индивидуального задания, построить уравнение регрессии (линейное, полиномиальное, логистическая, пробит-модель, «эластичная сеть»)
12. Решить задачу, согласно варианта задания

№	Индивидуальное задание	Данные для анализа
1	Сравнить вероятность выживание женщин и мужчин, имеющих на борту родственников. По заданному человеку (возможно с неполной картой признаков) на основе уравнения регрессии спрогнозировать вероятность его выживания	titanic.csv
2	Найти месяц, в котором наблюдался наибольший рост курса акций на конец месяца по сравнению с началом месяца. Восстановить значения в пропущенные дни за январь месяц 2017 года	apple.csv
3	Сравнить вероятность сердечно сосудистого заболевания у женщин и мужчин, в зависимости от уровня стенокардии. По заданному человеку (возможно с неполной картой признаков) на основе уравнения регрессии спрогнозировать вероятность наличия у него сердечно сосудистого заболевания	heart.csv
4	Сравнить вероятность выживание женщин и мужчин, разных возрастных категорий: до 18 лет, от 19 до 29, от 30 до 45, от 46 до 60 и старше 60 По заданному человеку (возможно с неполной картой признаков) на основе уравнения регрессии спрогнозировать вероятность его выживания	titanic.csv
5	Найти месяц, в котором наблюдалась наименьшая средняя разница между курсами акций на начало дня и на конец дня. Спрогнозировать курс акций на январь 2021 года.	apple.csv
6	Сравнить вероятность развития сердечно сосудистого заболевания у женщин и мужчин, разных возрастных категорий: до 18 лет, от 19 до 29, от 30 до 45, от 46 до 60 и старше 60 По заданному человеку (возможно с неполной картой признаков) на основе уравнения регрессии спрогнозировать вероятность высокого уровня холестерина (более 230)	heart.csv
7	Сравнить вероятность выживание женщин и мужчин, в зависимости от класса каюты. По пассажирам заданного класса восстановить значение в поле возраст.	titanic.csv
8		apple.csv
9		heart.csv
10		titanic.csv
11		apple.csv
12		heart.csv
13		titanic.csv
14		apple.csv
15		heart.csv

2. Лабораторная работа №2. Кластеризация данных

Общая часть: разработать программное обеспечение для решения задачи кластеризации (обучение без учителя)

Исходные данные: файл в формате csv

Необходимо: используя методы обучения без учителя, выполнить кластеризацию данных. Оценить качество классификации различными методами.

Ход работы:

1. Загрузить исходные данные согласно варианта из csv-файлов
2. На основании считанных данных сформировать DataFrame
3. Провести анализ на полноту данных
4. Построить матрицу диаграмм рассеивания, тепловую карту и матрицу корреляции (scatterplotmatrix, heatmap, plotcorr)
5. Провести анализ на основе построенных графиков
6. Согласно индивидуального задания, выполнить кластеризацию данных. Оценить качество классификации для каждого метода с помощью функционала качества
7. Построить дендрограммы
8. Выполнить отбор информативных признаков и уменьшить размерность пространства признаков
9. Построить матрицу диаграмм рассеивания, тепловую карту и матрицу корреляции (scatterplotmatrix, heatmap, plotcorr) для изменённого пространства признаков
10. Повторно выполнить кластеризацию данных. Оценить качество классификации для каждого метода с помощью функционала качества
11. Проверить, к какому классу принадлежит произвольный объект
12. Сделать выводы по результатам проделанной работы

№	Метод 1	Метод 2	Функционал качества	Данные для анализа
1	Выделение связных компонент	EM-алгоритм	$F_0 = \frac{\sum_{i<j}[y_i = y_j]\rho(x_i,x_j)}{\sum_{i<j}[y_i = y_j]} \to min$	winequality-red.csv
2	Минимальное оставное дерево	k-средних k = 5	$F_1 = \frac{\sum_{i<j}[y_i \neq y_j]\rho(x_i,x_j)}{\sum_{i<j}[y_i \neq y_j]} \to max$	responses.csv
3	Кратчайший незамкнутый путь	DBSCAN	$\Phi_0 = \sum_{y \in Y} \frac{1}{\lvert K_y \rvert} \sum_{i:y_i = y} \rho^2(x_i, \mu_y) \to min$	flavors_of_cacao.csv
4	Послойная кластеризация	k-средних k = 7	$\Phi_1 = \sum_{y \in Y} \rho^2(\mu_y, \mu) \to max$	winequality-red.csv
5	ForEL	EM-алгоритм	$\frac{F_0}{F_1} \to min$	responses.csv
6	ForEL-2 k=5	k-средних k = 5	$\frac{\Phi_0}{\Phi_1} \to min$	flavors_of_cacao.csv
7	SKAT	DBSCAN	$F_0 = \frac{\sum_{i<j}[y_i = y_j]\rho(x_i,x_j)}{\sum_{i<j}[y_i = y_j]} \to min$	winequality-red.csv
8	BigFor	DBSCAN	$F_1 = \frac{\sum_{i<j}[y_i \neq y_j]\rho(x_i,x_j)}{\sum_{i<j}[y_i \neq y_j]} \to max$	responses.csv
9	AGNES *алгоритм средней связи*	k-средних k = 9	$\Phi_0 = \sum_{y \in Y} \frac{1}{\lvert K_y \rvert} \sum_{i:y_i = y} \rho^2(x_i, \mu_y) \to min$	flavors_of_cacao.csv
10	AGNES *центроидный метод*	DBSCAN	$\Phi_1 = \sum_{y \in Y} \rho^2(\mu_y, \mu) \to max$	winequality-red.csv
11	AGNES *алгоритм одиночной связи*	k-средних k = 7	$\frac{F_0}{F_1} \to min$	responses.csv
12	AGNES *алгоритм полной связи*	EM-алгоритм	$\frac{\Phi_0}{\Phi_1} \to min$	flavors_of_cacao.csv
13	AGNES *метод минимума дисперсии Уорда*	k-средних k = 3	$F_1 = \frac{\sum_{i<j}[y_i \neq y_j]\rho(x_i,x_j)}{\sum_{i<j}[y_i \neq y_j]} \to max$	winequality-red.csv
14	ForEL-2 k=5	EM-алгоритм	$\Phi_0 = \sum_{y \in Y} \frac{1}{\lvert K_y \rvert} \sum_{i:y_i = y} \rho^2(x_i, \mu_y) \to min$	responses.csv
15	Кратчайший незамкнутый путь	DBSCAN	$\Phi_1 = \sum_{y \in Y} \rho^2(\mu_y, \mu) \to max$	flavors_of_cacao.csv

3. Лабораторная работа №3. Логические алгоритмы классификации

Общая часть: разработать программное обеспечение для решения задачи классификации (обучение с учителем) на основе деревьев решений

Исходные данные: файл в формате csv

Необходимо: используя дерево решений и случайный лес решить задачу бинарной классификации, согласно варианта

Ход работы:

1. Загрузить исходные данные согласно варианта из csv-файлов
2. На основании считанных данных сформировать DataFrame
3. Провести анализ на полноту данных
4. При необходимости выполнить нормировку, квантование и фильтрацию данных
5. При необходимости провести селекцию признаков и понижение размерности задачи
6. Согласно варианта построить дерево решений
7. Согласно варианта построить случайный лес
8. Оценить качество проведённой классификации с использованием кросс-валидации
9. Вычислить accuracy, precision, recall, f1-score
10. Провести варьирование параметрами классификатора с целью улучшения качества классификации
11. Провести анализ результатов работы классификатора на основе построенных графиков
12. Сделать выводы по результатам проделанной работы

№	Набор данных	Классы для классификации
1	titanic.csv	Выжившие/погибшие
2	bill_authentication.csv	class 0 / class 1
3	heart.csv	наличие/отсутствие заболевания
4	student-mat.csv	Наличие/отсутствие пристрастия к алкоголю
5	student-por.csv	Пол студента
6	titanic.csv	Пол пассажира
7	titanic.csv	Пассажир первого класса или нет
8	heart.csv	Пол пациента
9	student-mat.csv	Пол студента
10	student-por.csv	Наличие/отсутствие пристрастия к алкоголю
11		
12		

3. Лабораторная работа №4. Метрические методы классификации

Общая часть: разработать программное обеспечение для решения задачи классификации (обучение с учителем) на основе детерминированных методов классификации

Исходные данные: файл в формате csv

Необходимо: используя детерминированные методы классификации необходимо обучить модель для проведения классификации образов, оценить качество классификации.

Ход работы:

13. Загрузить исходные данные согласно варианта из csv-файлов
14. На основании считанных данных сформировать DataFrame
15. Провести анализ на полноту данных
16. При необходимости выполнить нормировку, квантование и фильтрацию данных
17. При необходимости провести селекцию признаков и понижение размерности задачи
18. Согласно варианта построить классификатор
19. Оценить качество проведённой классификации с использованием кросс-валидации
20. Построить матрицу ошибок
21. Вычислить аккуратность, точность и полноту
22. Оценить площадь под кривой ошибок
23. Построить логистическую функцию потерь
24. Провести варьирование параметрами классификатора с целью улучшения качества классификации
25. Провести анализ результатов работы классификатора на основе построенных графиков
26. Сделать выводы по результатам проделанной работы

№	Классификатор 1	Метрика	Набор данных	Предмет классификации
1	Метод дробящихся эталонов	евклидова	winequality-red.csv	Качество вина
2			flavors_of_cacao.csv	Качество батончика

3	Метод k-ближайших соседей	манхэттенская	winemag-data_first150k.csv	Качество вина
4	Метод парзерновского окна		winequality-red.csv	Качество вина
5	Метод потенциальных функци		flavors_of_cacao.csv	Процент какао
6	Полный метрический классификатор с классифицирующей функцией по МНК Тут существует больше 10 способа определения классификатора и плюс минимум 3 метрики, т.е. можно вариантов 30 дать	Чебышёва	flavors_of_cacao.csv	
7	СТОЛП	Минковского	winequality-red.csv	
8	FRiS-СТОЛП		flavors_of_cacao.csv	
9			winemag-data_first150k.csv	
10			flavors_of_cacao.csv	
11			flavors_of_cacao.csv	
12			winemag-data_first150k.csv	По стоимости продукта

5. Лабораторная работа №5. Статистические методы классификации.

Общая часть: разработать программное обеспечение для решения задачи классификации (обучение с учителем) на основе статистических методов классификации

Исходные данные: файл в формате csv или txt

Необходимо: используя статистические методы классификации необходимо обучить модель для проведения классификации образов, оценить качество классификации.

Ход работы:

1. Загрузить исходные данные согласно варианта из файла
2. На основании считанных данных сформировать DataFrame
3. Провести анализ на полноту данных
4. При необходимости выполнить нормировку, квантование и фильтрацию данных
5. Вычислить статистические характеристики данных (мат. Ожидание, дисперсию)
6. Построить графики распределения случайной величины
7. Графическим способом подобрать законы распределения случайных величин
8. С помощью заданного критерия, проверить гипотезу о законе распределения
9. При необходимости провести селекцию признаков и понижение размерности задачи
10. Согласно варианта построить классификатор
11. Оценить качество проведённой классификации с использованием указанного метода валидации
12. Построить матрицу ошибок
13. Вычислить аккуратность, точность и полноту
14. Оценить площадь под кривой ошибок
15. Построить логистическую функцию потерь
16. Выполнить робастную оценку среднего
17. Провести варьирование параметрами классификатора с целью улучшения качества классификации
18. Провести анализ результатов работы классификатора на основе построенных графиков
19. Сделать выводы по результатам проделанной работы

№	Классификатор	Набор данных	Предмет классификации	Критерий	Валидация
1	Наивный Байес	SMSSpamCollection.txt	Спам/не спам	Критерий Пирсона	Кросс-валидация
2	EM-алгоритм	flavors_of_cacao.csv	Качество батончика	Критерий Стьюдента	Последовательное случайное сэмплирование
3	Логистическая регрессия (МНК*)	winemag-data_first150k.csv	Качество вина	Критерий Фишера	Кросс-валидация по k-блокам
4	Логистическая регрессия (ММП**)	winequality-red.csv	Качество вина		Поэлементная кросс-валидация
5	Логистическая регрессия (BVD***)	bill_authentication.csv	Класс счёта		
6	Логистическая регрессия (L_2-регяляризация)				
7	Линейный дискриминантный анализ				
8					
9					
10					
11					
12					

*Метод наименьших квадратов
**Метод максимального правдоподобия
***Bias-variance decomposition
L2-регуляризация логистической функции потерь

СПИСОК ЛАБОРАТОРНЫХ РАБОТ

Номер лаб. работы	**Название темы**	**Объём в часах**
1	Предварительная обработка данных и их графическое отображение	8
2	Кластеризация данных	8
3	Логические алгоритмы классификации	6
4	Метрические методы классификации	6
5	Статистические методы классификации	6
Итого:		***34***