Министерство образования и науки Российской Федерации
Санкт-Петербургский государственный технический университет
Институт прикладной математики и механики
Кафедра «Телематика»

ЛАБОРАТОРНАЯ РАБОТА

ПО ТЕМЕ

«Деревья решений»

по направлению 02.04.01.02 «Организация и управление суперкомпьютерными системами»

Выполнил:
Студент гр. 13643.1 Титов А.И.
Проверил: Уткин Л.В.

Санкт-Петербург
2019

Оглавление

Постановка задачи

Требуется выполнить следующие задачи:

1. Исследовать, как объем обучающей выборки и количество тестовых данных, влияет на точность классификации или на вероятность ошибочной классификации в примере крестики-нолики и примере о спаме e-mail сообщений.

2. Сгенерировать 100 точек с двумя признаками X1 и X2 в соответствии с нормальным распределением так, что первые 50 точек (class -1) имеют параметры: мат. ожидание X1 равно 10, мат. ожидание X2 равно 14, среднеквадратические отклонения для обеих переменных равны 4. Вторые 50 точек (class +1) имеют параметры: мат. ожидание X1 равно 20, мат. ожидание X2 равно 18, среднеквадратические отклонения для обеих переменных равны 3. Построить соответствующие диаграммы, иллюстрирующие данные. Построить байесовский классификатор и оценить качество классификации.

3. Разработать байесовский классификатор для данных Титаник (Titanic dataset).

1 Крестики-нолики и спам e-mail сообщений

Для того чтобы исследовать, как объем обучающей выборки влияет на точность классификации были проделаны следующие шаги:

1. Загрузка и препроцессинг исходных данных из файла в структуру **data_frame**. Для хранения и препроцессинга использован пакет *pandas*.

2. Инициализация классификатора *GaussianNB* из пакета *sklearn*.

3. Выполнение цикла по разбиению соотношения обучающей и тестовой выборки от 0.1 до 0.9 с количеством шагов - 10. Разбиение производится с помощью метода *train_ test_ split* из пакета *sklearn*.

4. Применение классификатора к каждому разбиению. На данном этапе происходит основная часть вычислений:

   - рандомизация исходных данных;
   - использование наивного Байесовского классификатора для вычисления условных апостериорных вероятностей категориальных переменных при условии независимости признаков;
   - оценка полученной модели;
   - сравнение прогнозируемых значений с исходными;
   - вычисление значения вероятности ошибочной классификации.

5. Далее происходит построение графика зависимости значения вероятности ошибочной классификации от объема обучающей выборки.

6. Также выводится таблица, отображающая ошибки кластеризации для разбиения исходной выборки с соотношением обучающей выборки к тестовой 0.8.

Вычисления производились на двух выборках: крестики-нолики и спам e-mail сообщений. Соответственно, в результате получилось 2 графика (см. Рис.1-2). А также получена таблица, отображающая ошибки кластеризации для разбиения исходной выборки с соотношением обучающей выборки к тестовой 0.8. (см. таблица 1-2)

	0	**1**
0	129	0
1	50	13

Таблица 1. Сравнение результатов с исходными данными («Крестики-нолики»)

	0	**1**
0	382	166
1	16	357

Таблица 2. Сравнение результатов с исходными данными («Спам e-mail сообщений»)

2 Сгенерированный dataset

Для построения Байесовского классификатора и оценки качества классификации выполнены следующие шаги:

1. Генерация двух векторов по 100 элементов согласно заданным параметрам.
2. Составление таблицы из полученных векторов, а также назначение меток класса получившимся элементам.
3. Рандомизация таблицы. Разбиение исходной выборки на обучающее и тестирующее множество.
4. Использование наивного Байесовского классификатора для вычисления условных апостериорных вероятностей категориальных переменных при условии независимости признаков.
5. Оценка полученной модели.
6. Построение таблицы для сравнения прогнозируемых значений с исходными (см. таблица 3).
7. Построение графика принадлежности сгенерированных точек определённому классу. (см. рис. 3)
8. Построение графика зависимости значения вероятности ошибочной классификации от объема обучающей выборки. (см. рис. 4) (Аналогично предыдущей задаче)

	-1	**1**
-1	7	1
1	0	12

Таблица 3. Сравнение результатов с исходными данными (Сгенерированный dataset))

Тестирующая выборка 20% от исходной, следовательно для классификации было использовано 20 элементов. Анализируя Таблицу 3, можно сделать вывод о том, что лишь 1 значение было классифицировано неверно. Таким образом, величина ошибки для данного примера составила 0.05.

3 Titanic dataset

Для разработки байесовского классификатора для данных «Титаник» выполнены следующие шаги:

1. Загрузка и препроцессинг обучающей и тестирующей выборок в соответствующие таблицы.
2. Использование наивного Байесовского классификатора для вычисления условных апостериорных вероятностей категориальных переменных при условии независимости признаков.
3. Оценка полученной модели.
4. Построение таблицы для сравнения прогнозируемых значений с исходными (см. таблица 4).
5. Вычисление значения вероятности ошибочной классификации.
6. Построение графика зависимости значения вероятности ошибочной классификации от объема обучающей выборки. (см. рис. 5) (Аналогично предыдущей задаче)

	0	**1**
0	221	45
1	39	113

Таблица 4. Сравнение результатов с исходными данными (Titanic dataset))

Таким образом величина ошибочной классификации в примере составила $\approx 0,21$. Стоит заметить, что на изначальная выборка тестовых данных была получена основываясь на гендерном признаке выживания (указано в источнике выборки https://www.kaggle.com/c/titanic/). В данном примере обучение производилось по всем возможным колонкам, за исключением колонки «Ticket», а также колонки «SibSp» и «Parch» были заменены на их сумму + 1 (колонка «FamilySize»), колонка «Cabin» была заменена на колонку с бинарными значениями «HasCabin».