Санкт-Петербургский Политехнический Университет Петра Великого

Институт прикладной математики и механики

Кафедра Телематики при ЦНИИ РТК

Лабораторная работа

по теме

«Наивный Байесовский классификатор»

Преподаватель Уткин Л.В.

Студент гр.43607/2 Лисенкова А.А

Санкт-Петербург

2018 г.

# Постановка задачи

В ходе выполнения лабораторной работы необходимо 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).

# Ход работы

**1 задание**

Для того чтобы исследовать, как объём обучающей выборки и количество тестовых данных влияет на точность классификации выполнены следующие шаги:

1. Загрузка исходных данных из файла в структуру A\_raw.
2. Определение последовательности ‘x’: от 0.5 до 0.9 с шагом 0.05, которая отвечает за процентное соотношение обучающей и тестирующей выборок.
3. Для каждого разбиения обучающей и тестирующей выборки вызов функции calculate, в которой 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;
* построение таблицы для сравнения прогнозируемых значений с исходными;
* вычисление значения вероятности ошибочной классификации.

1. Построение графика зависимости значения вероятности ошибочной классификации от объема обучающей выборки.

Вычисления производились на двух выборках: крестики-нолики и спам e-mail сообщений. Соответственно, в результате получилось 2 графика (см. Рис.1-2). Следует также отметить, что для достижения более правдоподобной оценки действия из п.3 повторялись 5 раз, после чего находилось среднее арифметическое значения вероятности ошибочной классификации.

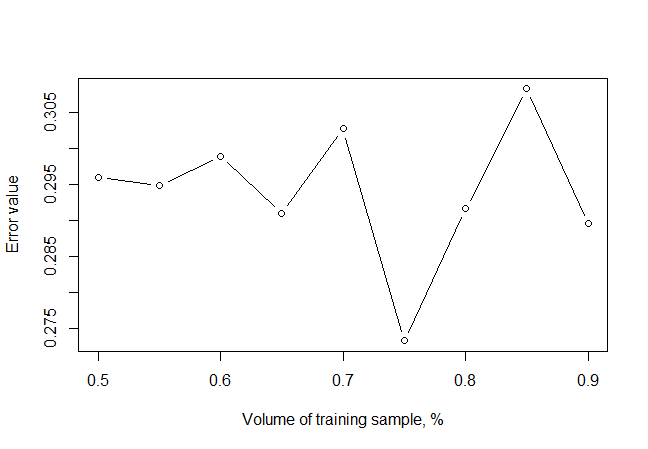


Рис. 1. График зависимости для примера «Крестики-нолики»

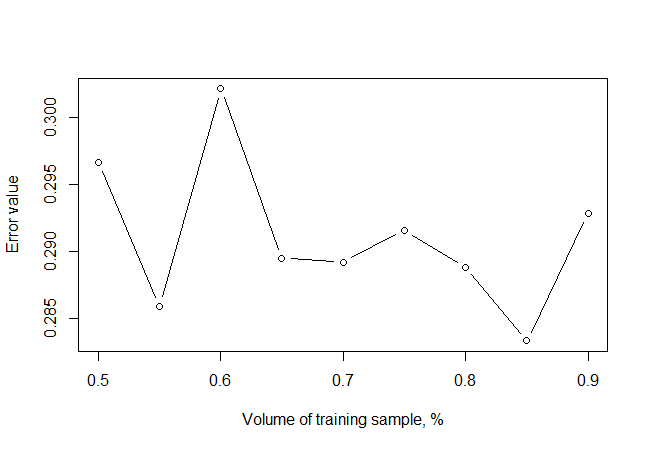


Рис. 2. График зависимости для примера «Спам e-mail сообщений»

**2 задание**

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.
4. Разбиение исходной выборки на обучающее и тестирующее множество.
5. Использование наивного Байесовского классификатора для вычисления условных апостериорных вероятностей категориальных переменных при условии независимости признаков.
6. Оценка полученной модели.
7. Построение таблицы для сравнения прогнозируемых значений с исходными (см. табл.1).
8. Построение графика принадлежности сгенерированных точек определённому классу.

|  |  |  |
| --- | --- | --- |
|  | **-1** | **1** |
| **-1** | 8 | 0 |
| **1** | 2 | 10 |

Табл. 1. Сравнение полученных результатов с исходными данными

Следует отметить, что тестирующая выборка составила всего 20% от исходной, поэтому для классификации было использовано 20 элементов. Анализируя табл.1, можно сделать вывод о том, что лишь 2 значения были классифицировано неверно. Таким образом, величина ошибки для данного примера составила 0,1.

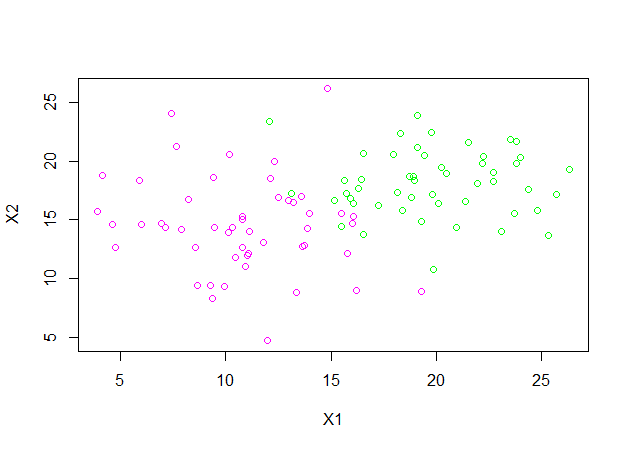
Для визуализации сгенерированных данных был построен график, представленный на Рис.3. Данные разделились на 2 класса: пурпурный для -1, зелёный для 1. 

Рис. 3. График распределения данных для двух классов

**3 задание**

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

1. Загрузка 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. табл.2).
5. Вычисление значения вероятности ошибочной классификации.

|  |  |  |
| --- | --- | --- |
|  | **0** | **1** |
| **0** | 211 | 44 |
| **1** | 55 | 108 |

Табл. 2. Сравнение полученных результатов с исходными данными для примера «Титаник»

Основываясь на полученных данных, величина ошибочной классификации составила 0,24.