Санкт-Петербургский Политехнический Университет имени Петра Великого
Физико-Механический Институт
Высшая школа прикладной математики и вычислительной физики

Отчет по лабораторной работе №1

по дисциплине

"Методы Машинного обучения"

Выполнил студент:
Тырыкин Я. А.
группа 5040102/20101
Преподаватель:
Уткин Л. В.

Санкт-Петербург
2022

Содержание

1 Постановка задачи.

В рамках лабораторной работы необходимо применить наивный классификатор Байеса в следующих задачах классификации:

1. Имеются два набора данных: набор e-mail сообщений и . Первый набор данных собран Hewlett-Packard Labs, которая классифицировала 4601 e-mail сообщений как спам или не спам. 57 признаков, содержащих частоты определенных слов, соответствующих названию признака и букв в сообщениях. Данные содержат 2788 e-mail сообщений, классифицируемых как "не спам"(0) и 1813 сообщений, классифицируемых как "спам"(1). Часть признаков – частоты определенных слов, соответствующих названию признака. Часть признаков связана с числом заглавных букв в письме. Второй набор данных - результаты игры в "Крестики-Нолики"следующего вида:

   - левая верхняя ячейка: {x,o,b}.
   - центральная верхняя ячейка: {x,o,b}.
   - правая верхняя ячейка: {x,o,b}.
   - левая средняя ячейка: {x,o,b}.
   - центральная средняя ячейка: {x,o,b}.
   - правая средняя ячейка: {x,o,b}.
   - левая нижняя ячейка: {x,o,b}.
   - центральная нижняя ячейка: {x,o,b}.
   - правая нижняя ячейка: {x,o,b}.

   X начинает первым, цель - победа x.

   Для описанных данных необходимо определить зависимость качества классификации от размера тестовой и обучающей выборок.

2. Имеется набор из 100 точек с двумя признаками X1 и X2 в соответствии с нормальным распределением таких, что первые 50 точек (Type 1) имеют параметры: $\overline{X1} = 10, \overline{X2} = 14, \sigma(X1) = \sigma(X2) = 4$. Вторые 50 точек (Type 2) имеют параметры: $\overline{X1} = 20, \overline{X2} = 18, \sigma(X1) = \sigma(X2) = 3$. Построить соответствующие диаграммы, иллюстрирующие данные. Построить байесовский классификатор и оценить качество классификации.

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

2 Наивный Классификатор Байеса.

Перед проведением исследований следует описать используемый алгоритм классификации. В основе классификатора лежит теорема Байеса:

Теорема 1 (Теорема Байеса). $P(y \in C|x) = \dfrac{P(x|y \in C)P(y \in C)}{P(x)}$, *где:*

- $P(y \in C|x)-$ вероятность того, что объект **x** принадлежит классу **C** (апостериорная вероятность класса).
- $P(x|y \in C)-$ вероятность встретить объект **x** среди всех объектов класса **C**.
- $P(y \in C)-$ безусловная вероятность встретить объект класса **C** (априорная вероятность класса).
- $P(x)-$ безусловная вероятность объекта **x**.

С применением данной теоремы классификатор ищет *наиболее вероятный класс объекта* $\boldsymbol{x}$, то есть решается задача максимизации вероятности $P(y \in C|x)$:

$$C_{opt} = arg \max_{C \in C^*} P(y \in C|x) = arg \max_{C \in C^*} \frac{P(x|y \in C)P(y \in C)}{P(x)} \tag{1}$$

Так как вероятность $P(x)$ не зависит от класса, задачу можно упростить:

$$C_{opt} = arg \max_{C \in C^*} P(x|y \in C)P(y \in C) \tag{2}$$

Важно отметить, что наивный классификатор Байеса строится в предположении, что признаки условно не зависят друг от друга:

$$\prod_{i=1}^{m} P(f_i|y \in C) = P(f_1|y \in C)P(f_2|y \in C)... = P(x|y \in C) \tag{3}$$

где **x** $-m$-значный вектор. В этом случае задача оптимизации выражается следующим образом:

$$C_{opt} = arg \max_{C \in C^*} P(y \in C) \prod_{i=1}^{m} P(f_i|y \in C) \tag{4}$$

Формулы для вычисление априорных и апостериорных вероястностей:

$$P(y \in C) = \frac{N_C}{N} \tag{5}$$

где N_C- количество объектов класса C в обучающей выборке, $N-$ общий размер обучающей выборки.

$$P(f_i \big| y \in C) = \frac{M_i(C) + \alpha}{\sum_{j=1}^{m} \big(M_j(C) + \alpha\big)} \tag{6}$$

где $M_i(C)-$ общее количество элементов с заданным значением признака i в классе C, $\alpha > 0-$слагаемое, исключающее нулевых значений вероятности (например, $\alpha = 1$).

3 Классификация электронных писем.

При проверке зависимости качества классификации от размера обучающей выборки будем наблюдать заметриками от 2 величин - *ошибкой I рода* и *ошибкой II рода* (так как в нашем случае оба набора данных делятся на два класса, это сделать достаточно просто).

Определение (Ошибка I рода)**.** *Ошибка I рода - ситуация, когда отвергнута верная нулевая (основная) гипотеза* H_0.

Определение (Ошибка II рода)**.** *Ошибка II рода - ситуация, когда принята неверная нулевая гипотеза* H_0.

Ошибки I и II рода считаются по матрице ошибок (в случае 2 классов это четырехпольные таблицы сопряженности), по горизонтальной оси которых лежат реальные классы исследуемых объектов, а по вертикальной - результаты работы классификатора. Результаты классификации делятся на 4 группы (вне зависимости от количества различаемых классов) - **TP (True-Positive)**, **TN (True-Negative)**, **FP (False-Positive)** и **FN (False-Negative)**. Первые две группы показывают количество количество образцов которые верно отнесены к данному классу или не отнесены к данному классу. Оставшиеся две группы включают случаи классификации, когда объект ошибочно отнесен/не отнесен к классу. Например, так выглядит матрица ошибок классификатора на наборе данных для второго класса из присутствующих четырех среди имеющихся элементов:

	y = 1	y = 2	y = 3	y = 4
a(1)	TN	FP	TN	TN
a(2)	FN	TP	FN	FN
a(3)	TN	FP	TN	TN
a(4)	TN	FP	TN	TN

Таблица 1: Матрица ошибок (Confusion matrix) классификатора **a** при выявлении класса 2.

Будем оценивать **2** метрики - *ROC-кривые* и *Recall-Precision - кривые.*

3.1 Precision-Recall кривые

Определение («Precision» - Точность)**.** *Точность - доля правильных ответов модели в пределах класса или доля объектов, действительно принадлежащих данному классу относительно всех объектов, которые система отнесла к этому классу.*

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

Именно введение точности не позволяет нам записывать все объекты в один класс, так как в этом случае мы получаем рост уровня **FP**.

Определение («Recall» - Полнота)**.** *Полнота — это доля истинно положительных классификаций. Полнота показывает, какую долю объектов, реально относящихся к положительному классу, мы предсказали верно. Полнота демонстрирует способность алгоритма обнаруживать данный класс вообще.*

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

Имея матрицу ошибок, очень просто можно вычислить точность и полноту для каждого класса. Точность равняется отношению соответствующего диагонального элемента матрицы к сумме всей строки класса, а полнота — отношению диагонального элемента матрицы и суммы всего столбца класса:

$$Precision = \frac{A_{C,C}}{\sum_{i=1}^{n} A_{C,i}} \tag{9}$$

$$Recall = \frac{A_{C,C}}{\sum_{i=1}^{n} A_{i,C}} \tag{10}$$

Кривые «Precision-Recall» для сообщений, относящихся к классам «spam» и «non-spam», выглядят следующим образом:

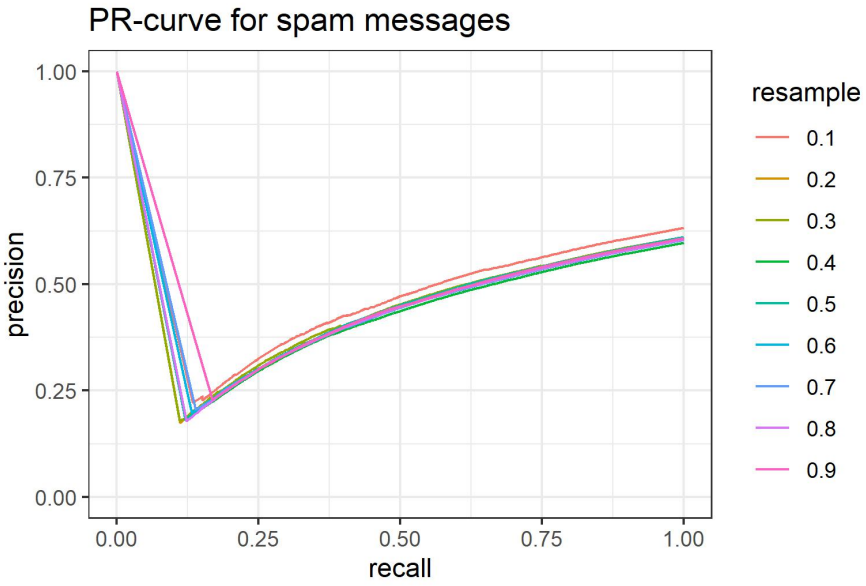

Рис. 1: PR-кривые качества выделения классификатором Байеса **спама** по размеру валидационной выборки как доли общего объема данных.

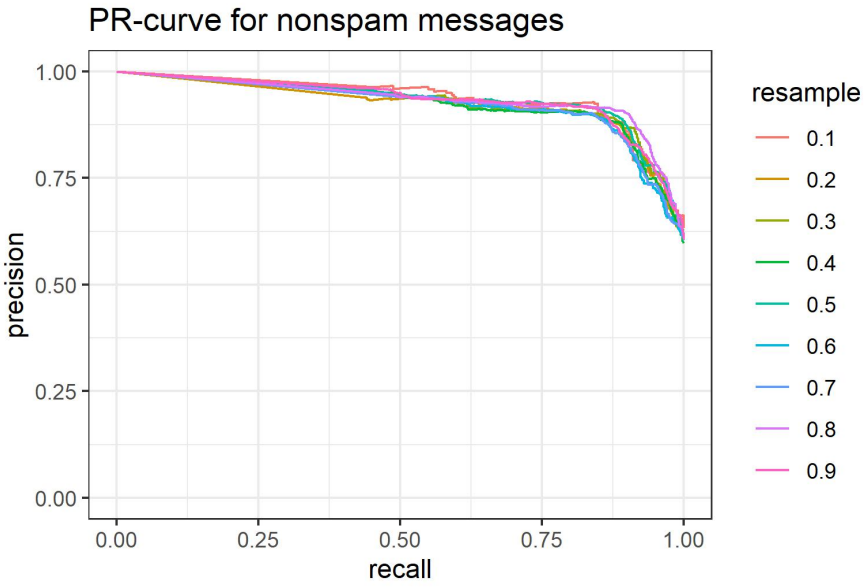

Рис. 2: PR-кривые качества выделения классификатором Байеса **не спама** по размеру валидационной выборки как доли общего объема данных.

Таблицы значений **AUC (Area Under Curve - площади под графиком)** для обоих графиков выглядит следующим образом:

Доля данных	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
AUC	0.495	0.462	0.466	0.459	0.468	0.47	0.473	0.464	0.489

Таблица 2: Таблица AUC значений для PR-кривых качества клиссификатора Байеса при индикации спама.

Доля данных	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
AUC	0.937	0.923	0.927	0.925	0.934	0.923	0.923	0.934	0.933

Таблица 3: Таблица AUC значений для PR-кривых качества клиссификатора Байеса при индикации не спама.

Оценки получились достаточно противоречивыми - классификатор отлично отделяет письма класса «не спам», но дает обратный или случайный прогноз для писем из класса «спам».

3.2 ROC-кривые

Определение (Кривая рабочих характеристик)**.** ***Кривая рабочих характеристик (Receiver Operating Characteristics curve)*** *Используется для анализа поведения классификаторов при различных пороговых значениях. Позволяет рассмотреть все пороговые значения для данного классификатора. Показывает долю* ***ложно положительных примеров (false positive rate, FPR)*** *в сравнении с долей* ***истинно положительных примеров (true positive rate, TPR)****, которые вычисляются по следующим формулам:*

$$TPR = \frac{TP}{TP + FN} \tag{11}$$

$$FPR = \frac{FP}{FP + TN} \tag{12}$$

ROC-кривые для сообщений, относящихся к классам «spam» и «non-spam», выглядят следующим образом:

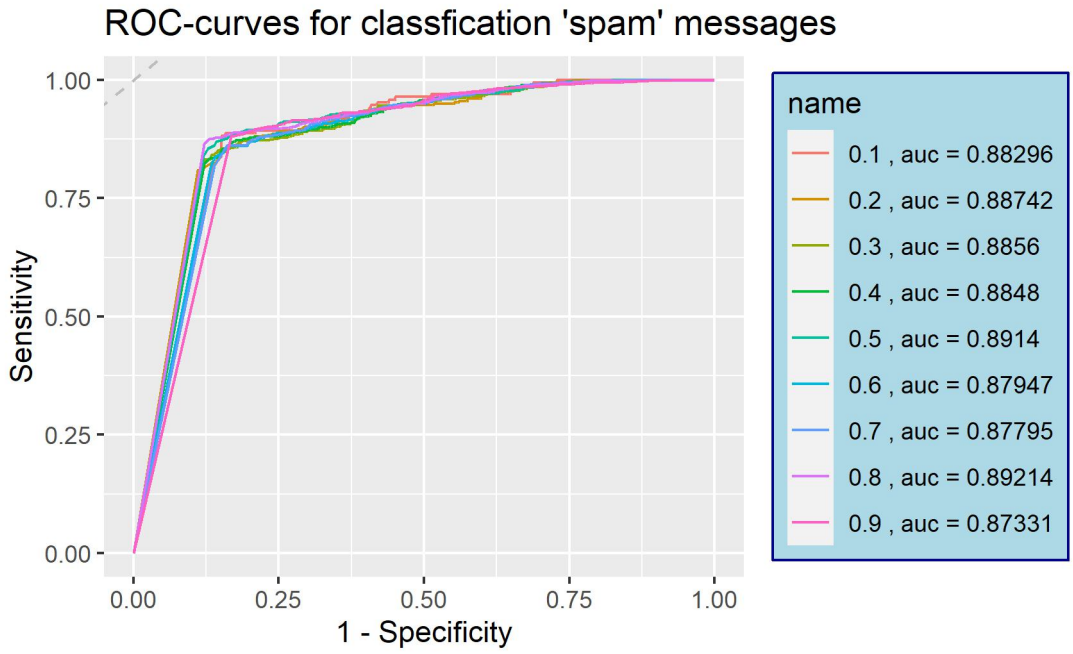

Рис. 3: ROC-кривые качества выделения классификатором Байеса **спама** по размеру валидационной выборки как доли общего объема данных.

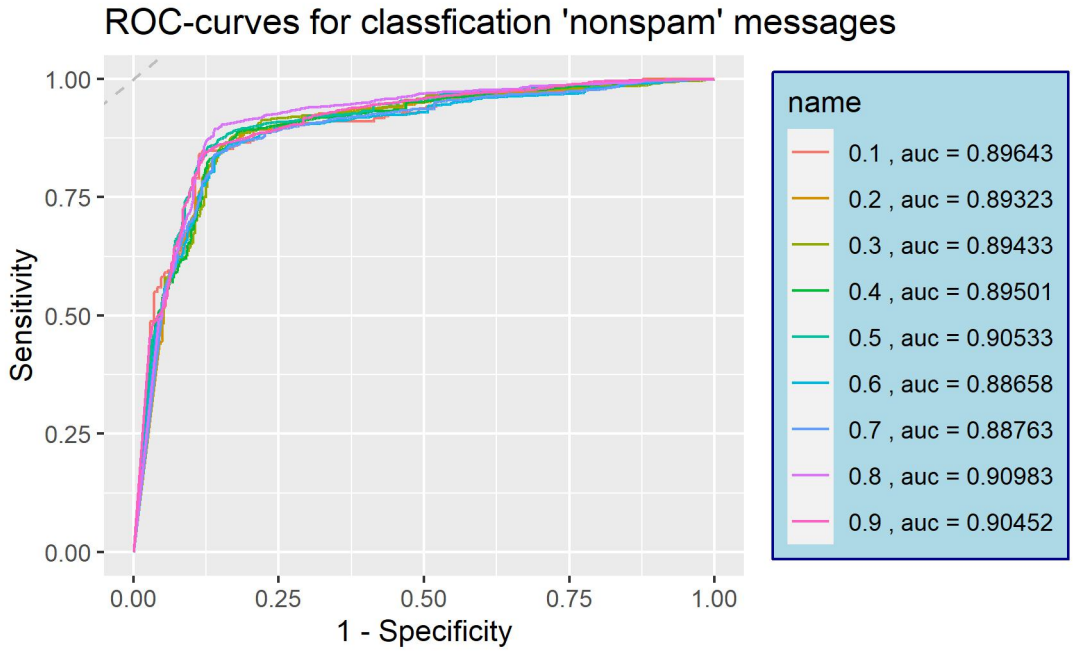

Рис. 4: ROC-кривые качества выделения классификатором Байеса **спама** по размеру валидационной выборки как доли общего объема данных.

Результаты бинарной классификации электронной почты, согласно оценке ROC, крайне хорошие, и в целом не сильно зависят от размера тестовой выборки по сравнению с обучающей. Это можно связать с большим размером исходного датасета (более 4500 писем).

Все результаты классификации можно признать удовлетворительными, но эталонные значения, по оценке AUC для каждой кривой, показывает классификатор на размере тестовой выборки в 30-50 % от имеющихся данных.

4 Классификация исходов игры «Крестики-Нолики».

Проанализируем качество работы классификатора Байеса на результатах игры в «Крестики-Нолики». Для этого оценим уже описанные метрики на имеющихся данных.

4.1 «Precision-Recall»-кривые

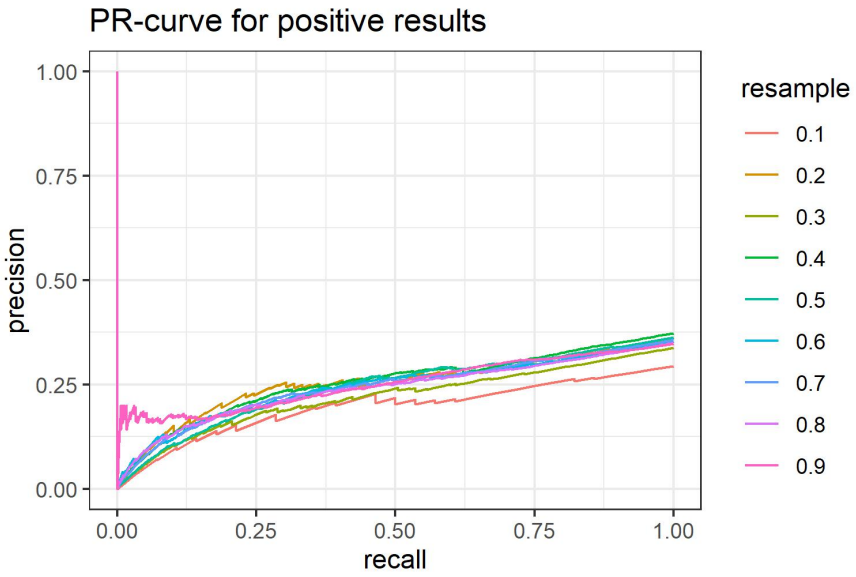

Рис. 5: PR-кривые качества выделения классификатором Байеса **положительных** результатов игры по размеру валидационной выборки как доли общего объема данных.

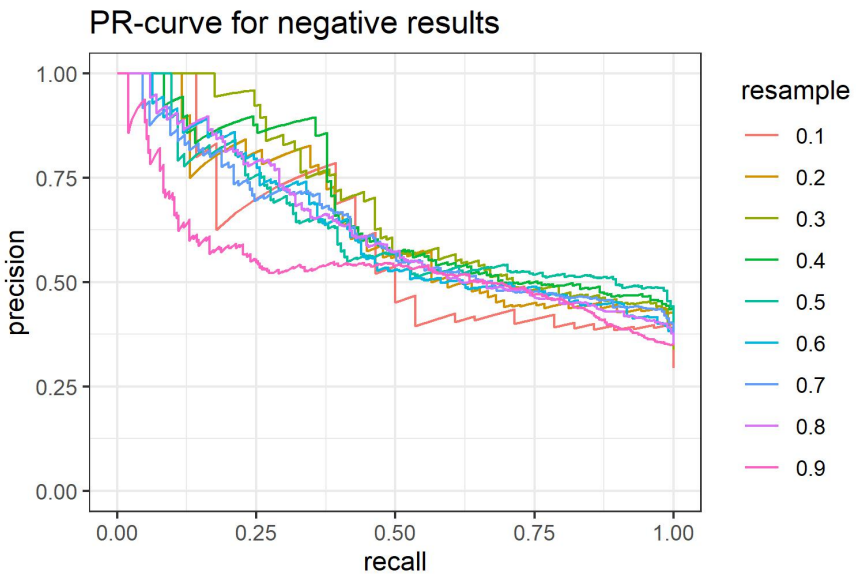

Рис. 6: PR-кривые качества выделения классификатором Байеса **отрицательных** результатов игры по размеру валидационной выборки как доли общего объема данных.

Доля данных	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
AUC	0.196	0.253	0.22	0.254	0.244	0.244	0.244	0.24	0.253

Таблица 4: Таблица AUC значений для PR-кривых качества клиссификатора Байеса индикации положительных результатов.

Доля данных	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
AUC	0.596	0.637	0.685	0.672	0.637	0.624	0.622	0.626	0.541

Таблица 5: Таблица AUC значений для PR-кривых качества клиссификатора Байеса индикации отрицательных результатов.

Кривые «Precision-Recall» показывают противоречивые результаты как для данного набора данных, так и для предыдущего (набора электронных писем). Класс отрицательных результатов игры выделяется плохо, но сравнимо со случайным бинарным классификатором. На классе положительных исходов классификатор Байеса дает обратный результат классификации, что сподвигает на более глубокий анализ методов и функциональных особенностей языка R.

Вряд ли это связано с данными, поскольку схожие результаты (с небольшими отклонениями) получены на двух различных датасетах.

4.2 ROC-кривые

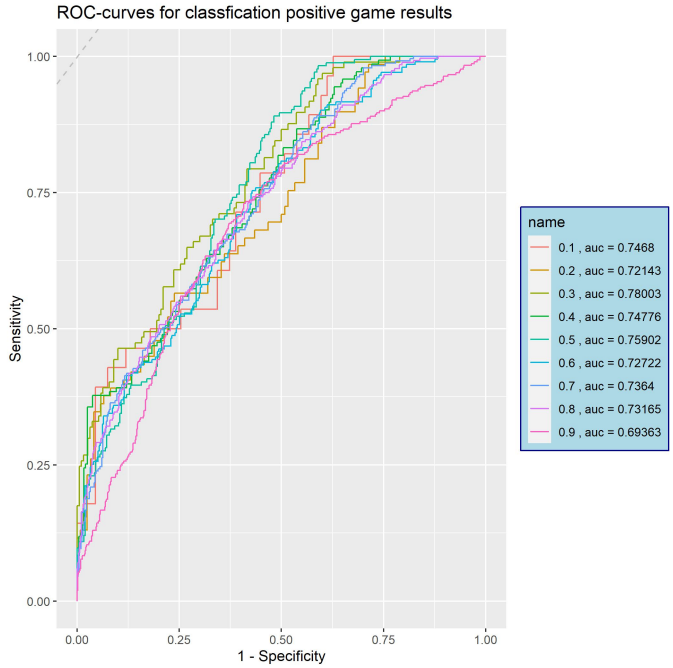

Рис. 7: ROC-кривые качества выделения классификатором Байеса **положительных** результатов игры по размеру валидационной выборки как доли общего объема данных.

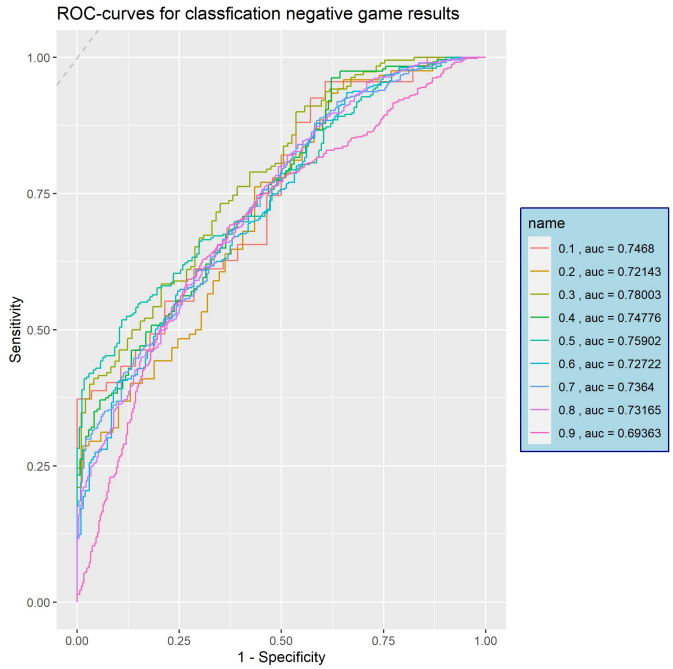

Рис. 8: ROC-кривые качества выделения классификатором Байеса **отрицательных** результатов игры по размеру валидационной выборки как доли общего объема данных.

Можно заметить общие очевидные закономерности: кривые меньших долей тестовой выборки от имеющихся данных находятся выше на большей части абсциссы, и площади под данными кривыми больше (пусть и незначительно). Но при этом можно отметить, что кривая, иллюстрирующая качество классификации при наименьшем размере тестовой выборки, не показывает эталонный результат, лучшее качестве - в случае тестирования классификатора на 30 % данных.

В целом минимальное значение AUC в 0.68 можно считать сравнительно неплохим результатом.

5 Классификация данных с нормально распределенными признаками.

Имеющиеся данные представим в виде гистограмм по каждому из нормально распреленных признаков:

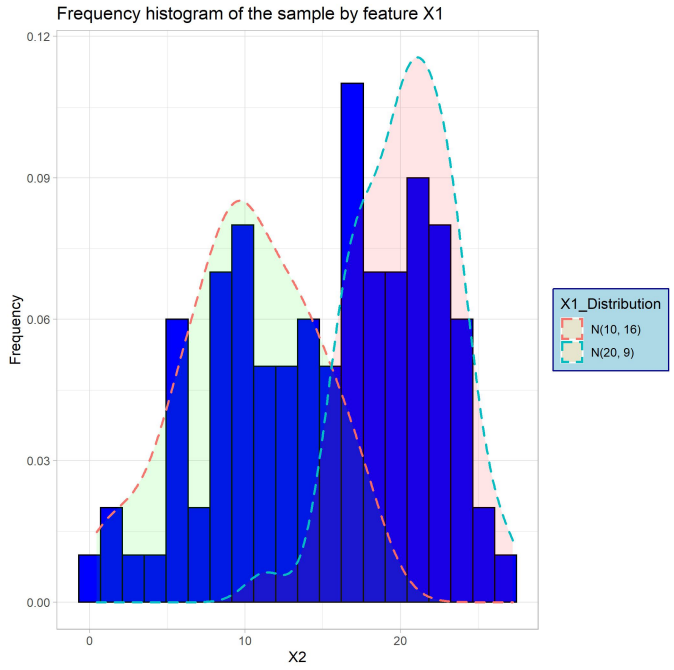

Рис. 9: Частотная гистограмма распределения значений признака X1 с плотностями распределения случайных величин, задающих этот признак.

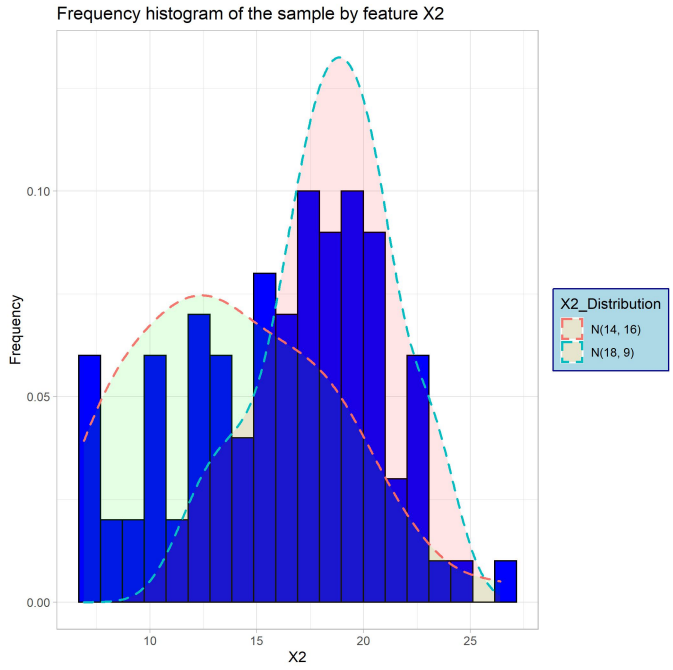

Рис. 10: Частотная гистограмма распределения значений признака X2 с плотностями распределения случайных величин, задающих этот признак.

Классификация изначально проводилась на тестовой и тренировочной выборках размеров 25 и 75 соответственно. В результате были получены следующие результаты:

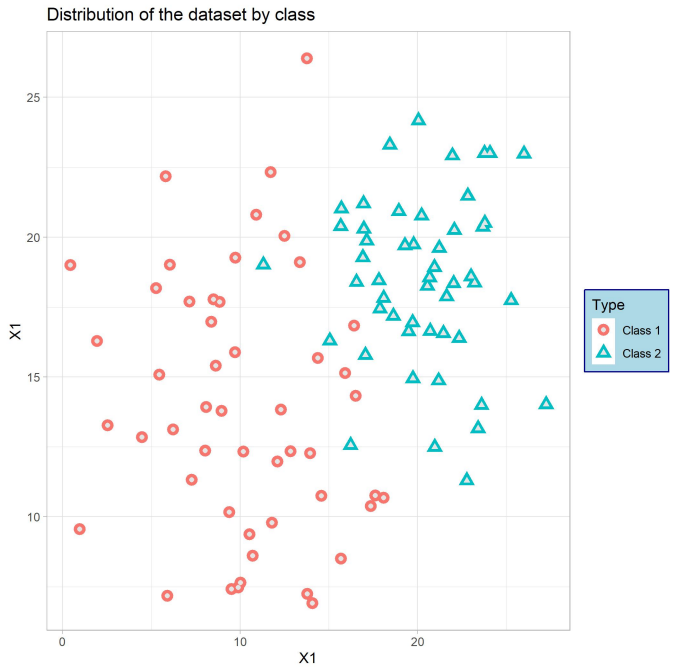

Рис. 11: Распределение классов точек по признакам.

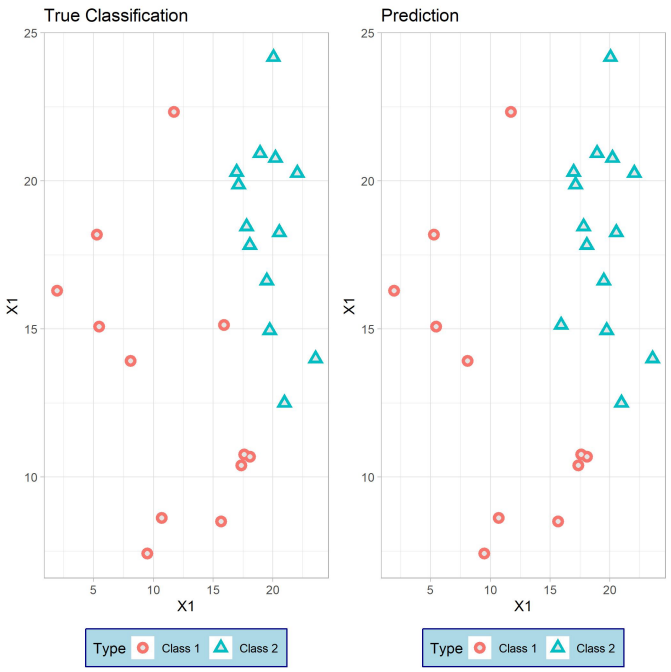

Рис. 12: Сравнение результата классификации с реальным разделением классов тестовой выборки.

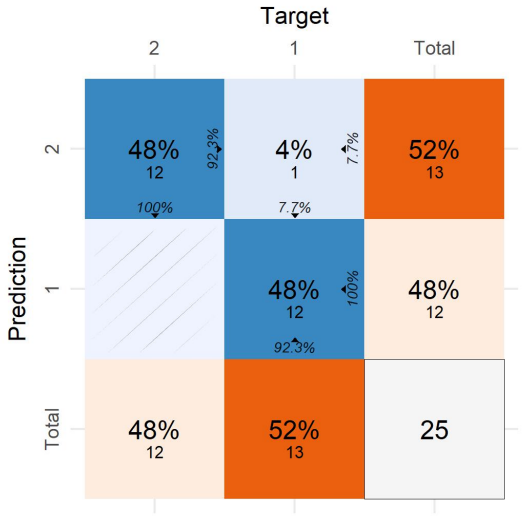

Рис. 13: Матрица ошибок классификации точек из тестовой выборки.

На полученных графиках и таблице видно, что с числовыми характеристиками и при значительном преобладании размера обучающей выборки над размером тестовой классификатор Байеса работает хорошо, дает малое (или вовсе нулевое) значение ошибок I и II рода. Следовательно, можно предположить, что в условии независимости количественных признаков, классификатор применим для классификации набора данных.

6 Классификация пассажиров Титаника.

Данные имеют начальное разделение на обучающую и тестовую выборки по файлам - «train.csv» и «test.csv». Оценим результаты классификации процентными соотношениями числа выживших/невыживших пассажиров, так как метки классов для валидачионного набора данных отсутствуют:

Класс / Выборка	Доля выживших	Доля погибших
Обучающие данные	0.6161616	0.3838384
Прогноз модели	0.7200957	0.2799043

Таблица 6: Таблица процентных оценок выживших и погибших пассажиров Титаника в размеченном и тестовом наборах данных.

Прогнозируемый процент выживших отклоняется в пределах 15 % от истинного значения в обучающем наборе, следовательно, классификатор Байеса показывает достаточно неплохие результаты на данных, содержащих как количественные, так и категориальные признаки.

7 Программная реализация вычислений.

Применение наивного классификатора Байеса осуществлялось при помощи программных средств языка программирования **R**, в частности таких пакетов, как **e1071**, **kernlab**, **ggplot2**, **tibble** и других. Более подробно ознакомиться с программным кодом и обучающими/тестовыми выборками можно в репозитории по следующей ссылке.

8 Заключение.

В целом, несмотря на то, что наивный классификатор Байеса является дотстаочно устаревшим методом, требующим смелых, упрощающих предположений, он остается достаточно эффективным средством как бинарной, так и множественной классификации, что подтвердилось в проведенных исследованиях.

Оценки ROC и PR дали противоречивые результаты, что дает почву для дальнейшего изучения их применимости к имеющимся данным и, особенно, программных реализаций классификатора и средств его визуализации на языке программирования R.

9 Список источников информации.

1. Сайт университета ИТМО - теоретическое описание и вычислительные формулы метрикк качества.

2. Сайт преподавателя СПБПУ Л. В. Уткина - основной лекционный и теоретический материал, разбирающий наивный байесовский классификатор.