

Санкт-Петербургский Политехнический Университет имени Петра Великого  
Физико-Механический Институт

**Высшая школа прикладной математики и вычислительной  
физики**

**Отчет по лабораторной работе №2  
по дисциплине  
"Методы Машинного обучения"**

Выполнил студент:

Тырыкин Я. А.

группа 5040102/20101

Преподаватель:

Уткин Л. В.

Санкт-Петербург  
2022

# Содержание

<b>1 Постановка задачи.</b> . . . . .	<b>2</b>
<b>2 Метод k ближайших соседей</b> . . . . .	<b>3</b>
2.1 Основной алгоритм. . . . .	3
2.2 Ядро сглаживания и метрика расстояния. . . . .	4
<b>3 Классификация электронных писем.</b> . . . . .	<b>5</b>
3.1 Precision-Recall кривые. . . . .	5
3.2 ROC-кривые. . . . .	8
<b>4 Классификация исходов игры в «Крестики-Нолики».</b> . . . . .	<b>10</b>
4.1 Precision-Recall кривые. . . . .	10
4.2 ROC-кривые. . . . .	13
<b>5 Классификация образцов стекла.</b> . . . . .	<b>16</b>
<b>6 Классификация цветных точек.</b> . . . . .	<b>20</b>
<b>7 Классификация пассажиров Титаника.</b> . . . . .	<b>22</b>
<b>8 Программная реализация вычислений.</b> . . . . .	<b>23</b>
<b>9 Заключение.</b> . . . . .	<b>23</b>
<b>10 Список использованных источников информации.</b> . . . . .	<b>23</b>

# 1 Постановка задачи.

В рамках лабораторной работы необходимо применить метод ближайших соседей (K Nearest Neighbours, Knn) 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 данных **Glass**, где каждый элемент характеризуется следующими 9 признаками:

- RI - показатель преломления.
- Na - процент содержания соды в соответствующем образце.
- Mg - процент содержания магния в соответствующем образце.
- Al - процент содержания алюминия в соответствующем образце.
- Si - процент содержания кремния в соответствующем образце.

- K - процент содержания калия в соответствующем образце.
- Ca - процент содержания кальция в соответствующем образце.
- Ba - процент содержания бария в соответствующем образце.
- Fe - процент содержания железа в соответствующем образце.

Образцы стекла разделяются на **7** классов:

- (a) **1** - Окна зданий, плавильная обработка.
- (b) **2** - Окна зданий, не плавильная обработка
- (c) **3** - Автомобильные окна, плавильная обработка.
- (d) **4** - Автомобильные окна, не плавильная обработка (нет в базе).
- (e) **5** - Контейнеры.
- (f) **6** - Посуда.
- (g) **7** - Фары.

Необходимо построить графики зависимости ошибки классификации от значения **k** и от типа ядра. Также нужно определить зависимость ошибки классификации от параметра расстояния Миньковского и от каждого из признаков. Также построенным классификатором отнести к одной из полученных групп образец с параметрами: RI = 1.516, Na = 11.7, Mg = 1.01, Al = 1.19, Si = 72.59, K = 0.43, Ca = 11.44, Ba = 0.02, Fe = 0.1.

3. Классифицировать с помощью рассматриваемого метода данные о точках на плоскости, характеризуемые двумя вещественными количественными компонентами и одной категориальной, задающей цвет. Обучающая выборка находится в файле «svmdata4.txt», валидационная выборка - в файле «svmdata4test.txt». Найти оптимальное значение **k**, обеспечивающее наименьшую ошибку классификации, а также визуализировать как исходные данные, так и результаты классификации.
4. Классифицировать данные пассажиров Титаника, представленных в Titanic dataset. Оценить качество классификации.

## 2 Метод k ближайших соседей

### 2.1 Основной алгоритм.

Для разбиения данных на классы в текущей лабораторной работе используется метод **k ближайших соседей (k nearest neighbors)**.

Суть алгоритма состоит в максимизации вероятности принадлежности элемента выборки одному из имеющихся классов. Для этого в ходе работы метода k ближайших соседей решается следующая оптимизационная задача:

$$a(x^*) = \arg \max_{y \in Y} \sum_{i=1}^n [y_i = y] w(i, x^*) = \arg \max_{y \in Y} \sum_{i=1}^n [y_i = y] K\left(\frac{\rho(x_i, x^*)}{h}\right) \quad (1)$$

где  $[y_i = y]$  – индикатор принадлежности  $x^*$  i-ому классу,  $w(i, x^*)$  – вес i-го соседа объекта  $x^*$ ,  $h$  – радиус ядра (или «ширина окна»),  $\rho(x_i, x_j)$  – метрика расстояния между элементами набора данных,  $K(u)$  – функция ядерного сглаживания.

Пошаговый алгоритм представляется следующим образом:

1. Для каждого из элементов (обучающей выборки) высчитываем расстояние до экземпляра (тестовой выборки), который необходимо классифицировать, по метрике  $\rho(x_i, x^*)$  с учетом сглаживания.
2. Выбираем  $k$  наибольших сглаженных (взвешенных) расстояний до элемента обучающей выборки.
3. Оцениаем частотную статистику классов в полученном наборе из  $k$  элементов: метка наиболее часто встретившегося класса присваивается первому.

## 2.2 Ядро сглаживания и метрика расстояния.

Простейший метод k соседей не очень эффективен, поэтому в реальных задачах используется усложненный метрический метод. В этом варианте используется взвешивание экземпляров при помощи весовой функции  $w(i, x)$ , в качестве которой обычно берут одну из следующих ядерных функций  $K(u)$ :

- **Прямоугольное ядро (rectangular):**  $K(u) = \frac{1}{2}$
- **Треугольное ядро (triangular):**  $K(u) = 1 - |u|$
- **Епанечниково ядро (epanechnikov):**  $K(u) = \frac{3}{4}(1 - |u|)$
- **Гауссово ядро (gaussian):**  $K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$

Ядерных функций, широко использующихся в машинном обучении и математической статистике, существует куда больше, но в данной работе использовались лишь 4 описанных варианта.

Близость определенной точки по отношению к соседям определяется при помощи метрики расстояния  $\rho(x_i, x_j)$ . Существуют различные варианты метрик расстояния между элементами данных в  $n$ -мерном пространстве признаков:

- **Косинусная метрика:**  $\rho(x_i, x_j) = 1 - \frac{x_i * x_j}{\|x_i\| * \|x_j\|}$
- **Расстояние Миньковского:**  $\rho(x_i, x_j) = \left( \sum_{k=1}^n |x_{i,k} - x_{j,k}|^p \right)^{\frac{1}{p}}$ . Различные вариации параметра  $p$  дают разные метрики расстояния: при  $p = 1$  используется **метрика Манхэттена (или метрика городских кварталов)**, при  $p = 2$  получается наиболее часто встречающаяся **Евклидова метрика**.
- **Расстояние Жаккара:**  $\rho(x_i, x_j) = 1 - \frac{|x_i \cap x_j|}{|x_i \cup x_j|}$ . Операнды данной метрики - некоторые множества, решение близости принимается по отношению мощностей пересечения и объединения операндов.

В данной работе исследуются лишь различные варианты расстояния Миньковского.

### 3 Классификация электронных писем.

Перед тем, как начинать оценивать зависимости качества классификации от размера тестовых выборок, были найдены наиболее оптимальные параметры числа ближайших соседей  $k$  и ядра:

Ядро	Треугольное	Прямоугольное	Гаусса	Епанчникова
$k$	15	1	20	18
Accuracy	0.929	0.918	0.925	0.924

Таблица 1: Таблица оптимальных значений  $k$  для каждого типа ядра.

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

#### 3.1 Precision-Recall кривые.

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

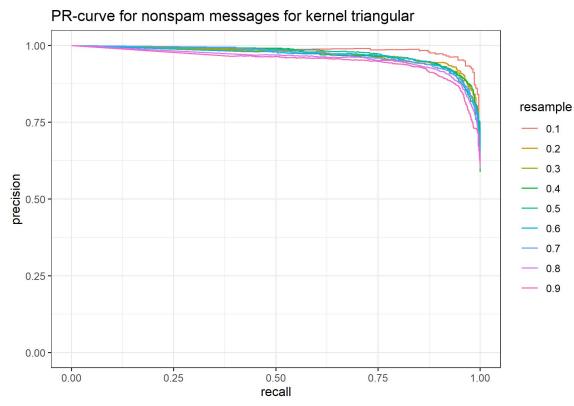


Рис. 1: PR-кривая классификации не спама с помощью Knn с треугольным ядром.

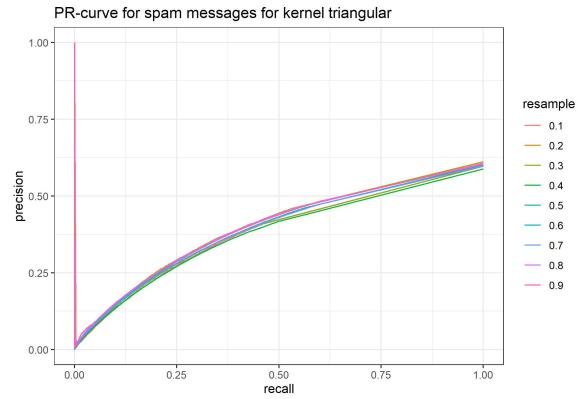


Рис. 2: PR-кривая классификации спама с помощью Knn с треугольным ядром.

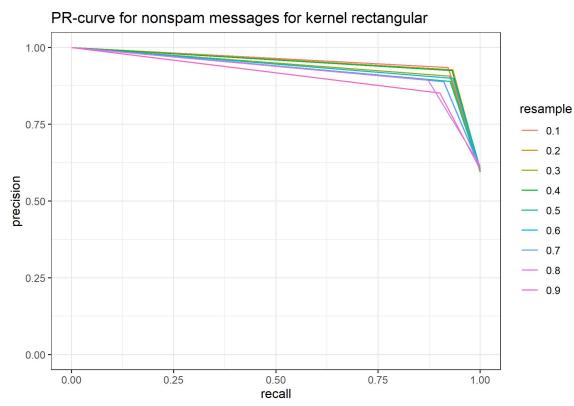


Рис. 3: PR-кривая классификации не спама с помощью Knn с прямоугольным ядром.

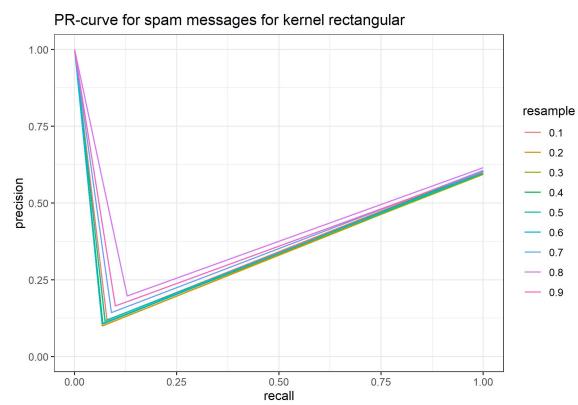


Рис. 4: PR-кривая классификации спама с помощью Knn с прямоугольным ядром.

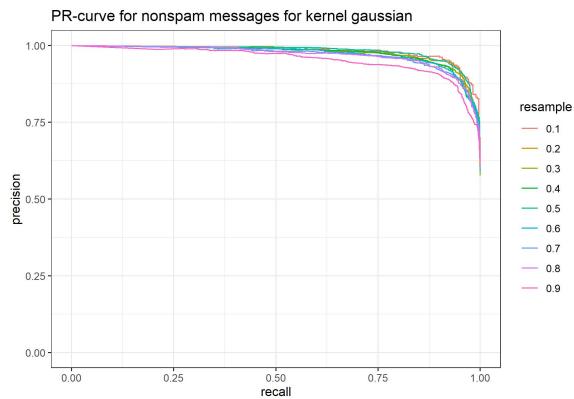


Рис. 5: PR-кривая классификации не спама с помощью Knn с ядром Гаусса.

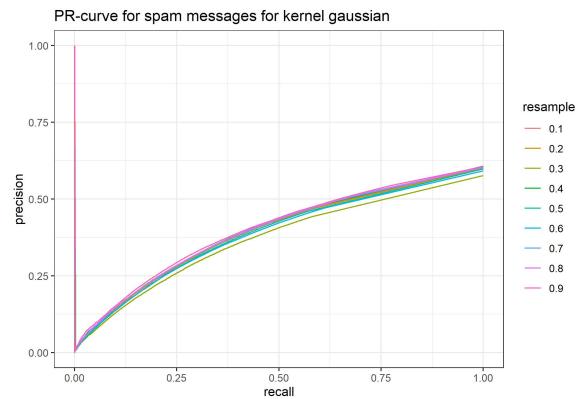


Рис. 6: PR-кривая классификации спама с помощью Knn с ядром Гаусса.

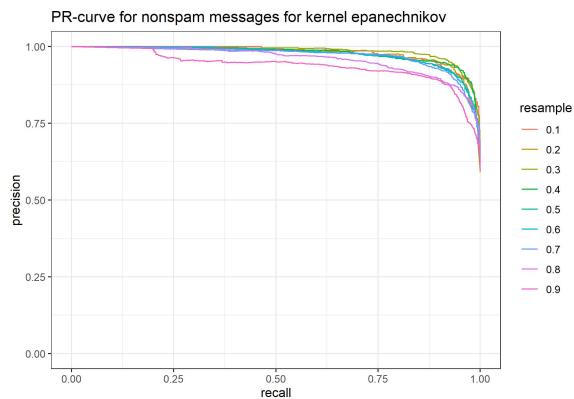


Рис. 7: PR-кривая классификации не спама с помощью Knn с ядром Епанечникова.

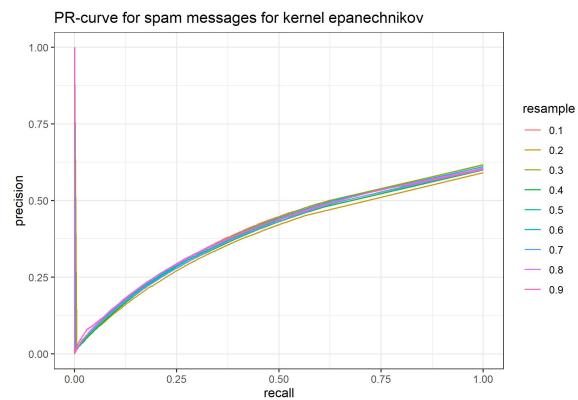


Рис. 8: PR-кривая классификации спама с помощью Knn с ядром Епанечникова.

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

AUC \ Доля данных	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Spam, Rect Kernel	0.378	0.362	0.368	0.369	0.373	0.367	0.393	0.431	0.404
Nonspam, Rect Kernel	0.951	0.95	0.938	0.948	0.93	0.937	0.927	0.923	0.906
Spam, Triag Kernel	0.383	0.403	0.386	0.379	0.393	0.398	0.391	0.401	0.399
Nonspam, Triag Kernel	0.985	0.97	0.971	0.972	0.972	0.968	0.969	0.958	0.951
Spam, Gauss Kernel	0.394	0.392	0.371	0.397	0.388	0.384	0.396	0.397	0.402
Nonspam, Gauss Kernel	0.981	0.978	0.976	0.974	0.981	0.972	0.969	0.969	0.954
Spam, Epan Kernel	0.406	0.385	0.404	0.391	0.398	0.402	0.394	0.401	0.402
Nonspam, Epan Kernel	0.979	0.975	0.984	0.978	0.974	0.974	0.97	0.959	0.937

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

Оценки получились достаточно противоречивыми - классификатор отлично отделяет письма класса «не спам», но дает обратный или случайный прогноз для писем из класса «спам». Данное утверждение не зависит от типа выбранного ядра.

### 3.2 ROC-кривые.

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

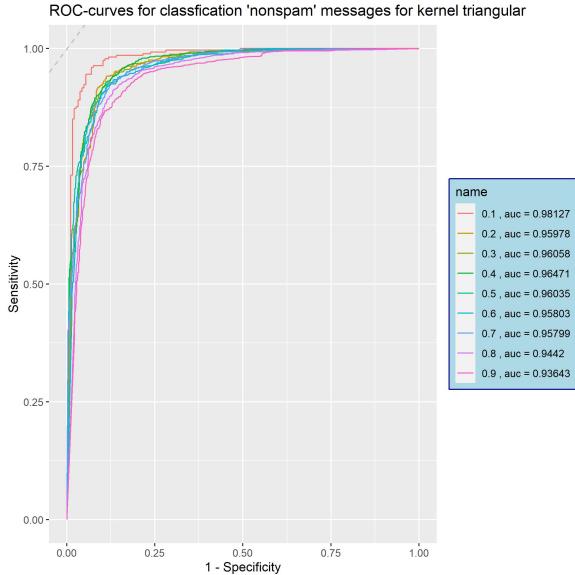


Рис. 9: ROC-кривая классификации не спама с помощью Knn с треугольным ядром.

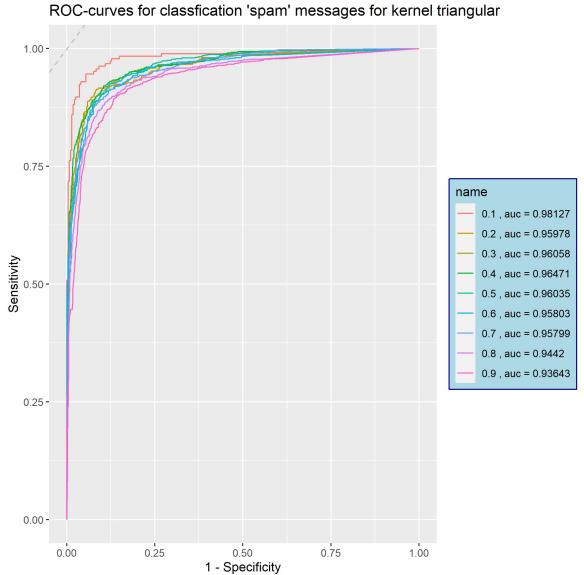


Рис. 10: ROC-кривая классификации спама с помощью Knn с треугольным ядром.

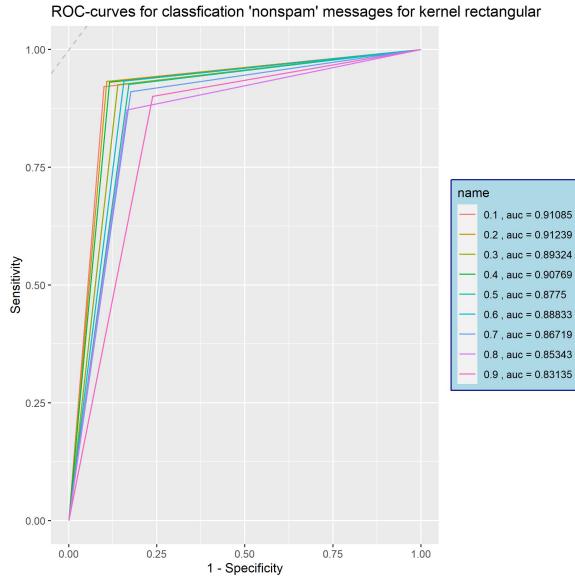


Рис. 11: ROC-кривая классификации не спама с помощью Knn с прямоугольным ядром.

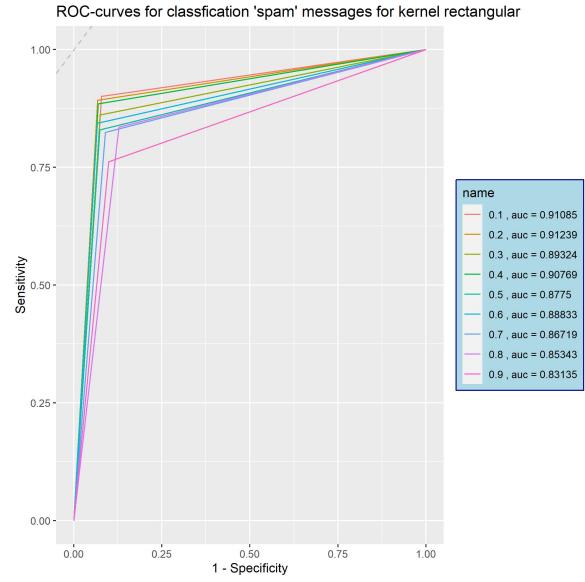


Рис. 12: ROC-кривая классификации спама с помощью Knn с прямоугольным ядром.

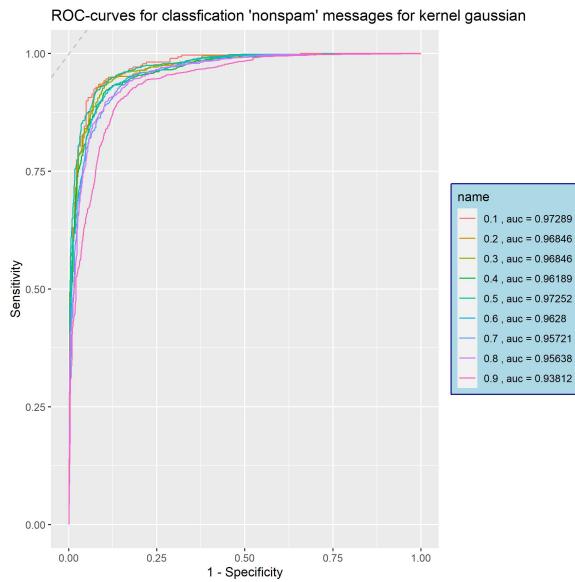


Рис. 13: ROC-кривая классификации не спама с помощью Knn с ядром Гаусса.

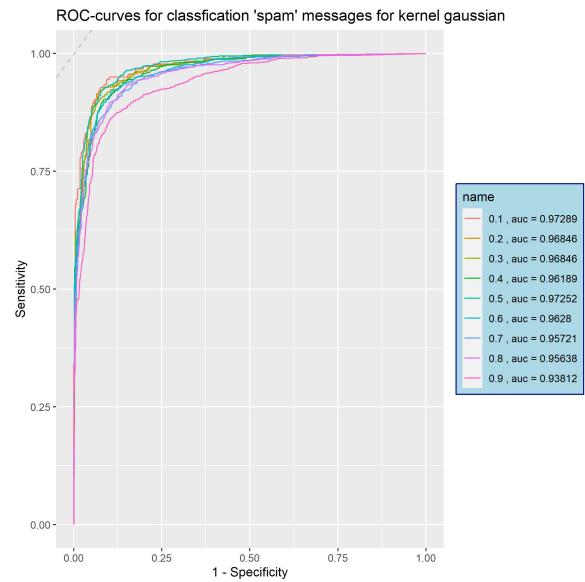


Рис. 14: ROC-кривая классификации спама с помощью Knn с ядром Гаусса.

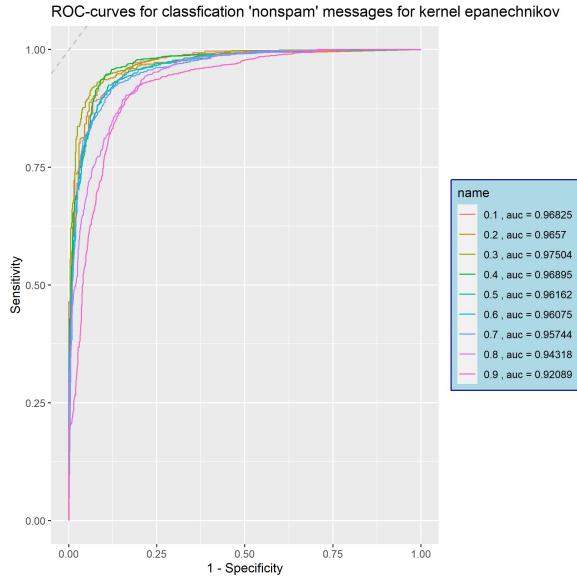


Рис. 15: ROC-кривая классификации не спама с помощью Knn с ядром Епанечникова.

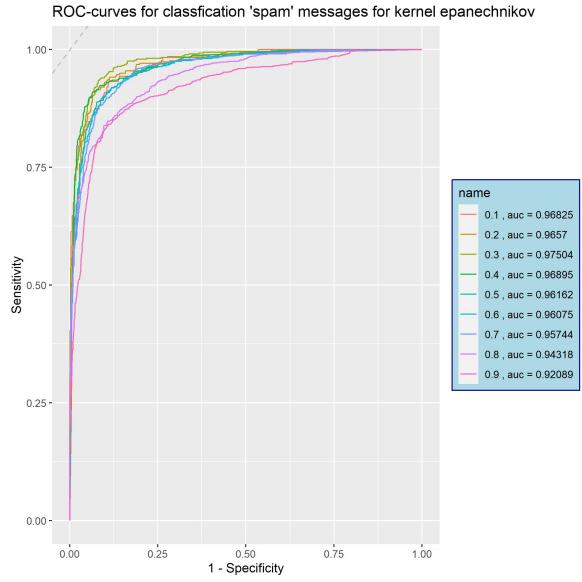


Рис. 16: ROC-кривая классификации спама с помощью Knn с ядром Епанечникова.

Из представленных зависимостей можно сделать вывод, что отношение размера тренировочной выборки к тестовой пусть и незначительно, но имеет значение: чем меньше доля тренировочной выборки в объеме данных, тем лучше качество классификации. Эта закономерность общая для любого типа сглаживающего ядра среди рассмотренного перечня. Но при этом худшие результаты работы показывает прямоугольное (равномерное ядро), лучшие - треугольное ядро.

## 4 Классификация исходов игры в «Крестики-Нолики».

### 4.1 Precision-Recall кривые.

Кривые «Precision-Recall» результатов классификации исходов игры в «Крестики-Нолики» выглядят следующим образом:

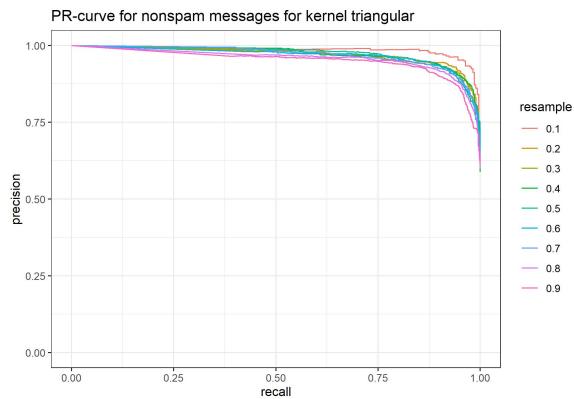


Рис. 17: PR-кривая классификации не спама с помощью Knn с треугольным ядром.

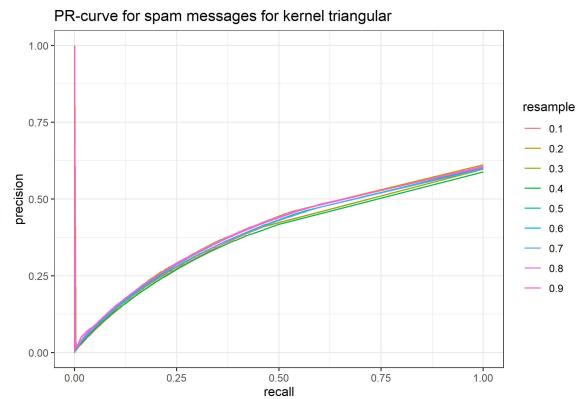


Рис. 18: PR-кривая классификации спама с помощью Knn с треугольным ядром.

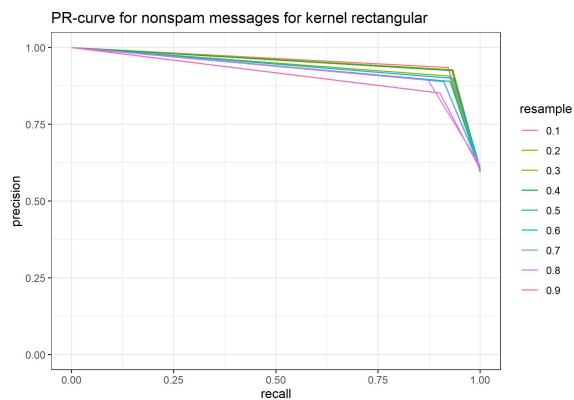


Рис. 19: PR-кривая классификации не спама с помощью Knn с прямоугольным ядром.

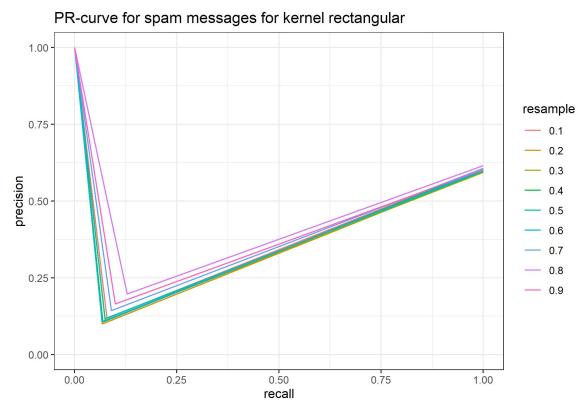


Рис. 20: PR-кривая классификации спама с помощью Knn с прямоугольным ядром.

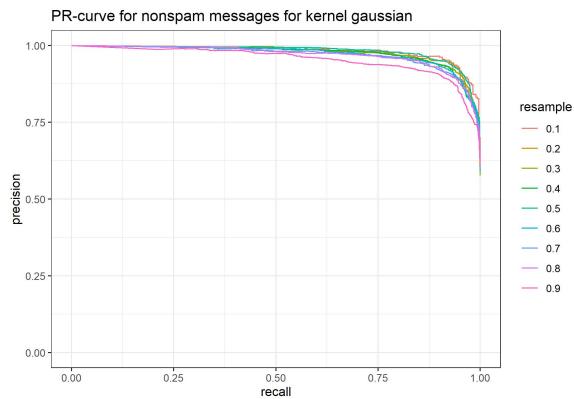


Рис. 21: PR-кривая классификации не спама с помощью Knn с ядром Гаусса.

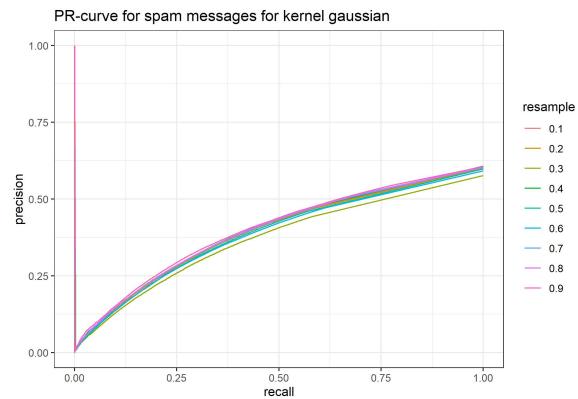


Рис. 22: PR-кривая классификации спама с помощью Knn с ядром Гаусса.

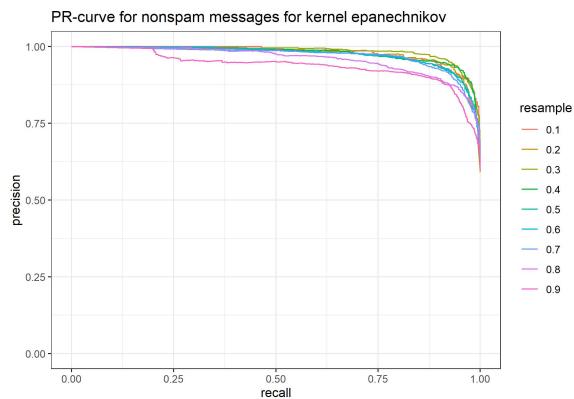


Рис. 23: PR-кривая классификации не спама с помощью Knn с ядром Епанечникова.

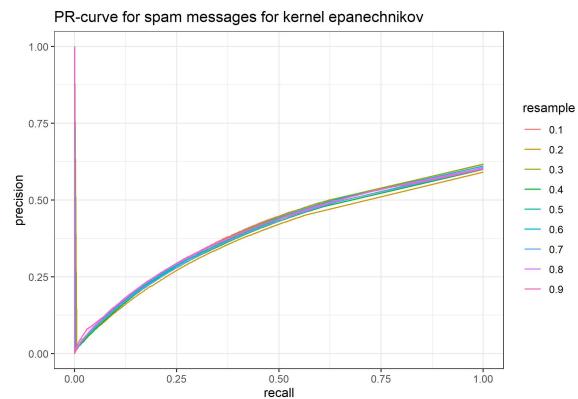


Рис. 24: PR-кривая классификации спама с помощью Knn с ядром Епанечникова.

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

AUC \ Доля данных	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Positive, Rect Kernel	0.204	0.189	0.180	0.222	0.196	0.199	0.196	0.225	0.236
Negative, Rect Kernel	1.0	1.0	0.998	0.998	0.956	0.931	0.856	0.736	0.585
Positive, Triag Kernel	0.185	0.183	0.201	0.2	0.186	0.195	0.2	0.202	0.205
Negative, Triag Kernel	1.0	0.996	0.999	0.991	0.986	0.99	0.972	0.876	0.741
Positive, Gauss Kernel	0.248	0.203	0.185	0.219	0.189	0.198	0.202	0.196	0.239
Negative, Gauss Kernel	0.994	1.0	0.998	0.998	0.991	0.963	0.887	0.845	0.664
Positive, Epan Kernel	0.178	0.190	0.213	0.195	0.199	0.196	0.202	0.200	0.227
Negative, Epan Kernel	0.998	1.0	0.999	0.998	0.998	0.993	0.0.933	0.858	0.716

Таблица 3: Таблица AUC значений для PR-кривых качества классификатора Байеса при индикации результатов игры в «Крестики-Нолики».

Результаты крайне противоречивы. С одной стороны, класс отрицательных исходов игры определяется даже слишком хорошо: во многих случаях значение AUC принимает значение 1, к которому стремится, но, по идее, достигать не должно. С другой стороны, абсолютно противоположную тенденцию поддерживает классификация положительных исходов игры: результаты хуже, чем на случайному классификаторе, складывается ощущение, что определяется противоположный класс исходов.

В целом же стоит отметить, что треугольное ядро сглаживания более устойчиво к изменению размера выборки: AUC не достигают абсурдно больших и малых значений, но при этом изменения площади под графиком при существенном увеличении тестовой выборки куда меньше по сравнению с остальными ядрами.

## 4.2 ROC-кривые.

ROC-кривые результатов классификации исходов игры в «Крестики-Нолики» выглядят следующим образом:

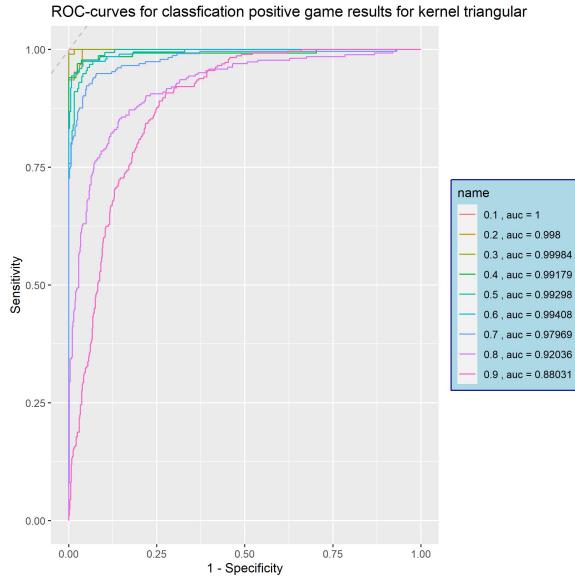


Рис. 25: ROC-кривая классификации не спама с помощью Knn с треугольным ядром.

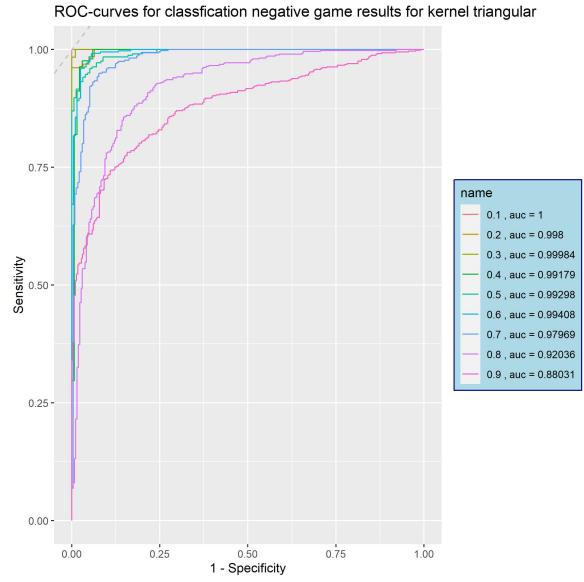


Рис. 26: ROC-кривая классификации спама с помощью Knn с треугольным ядром.

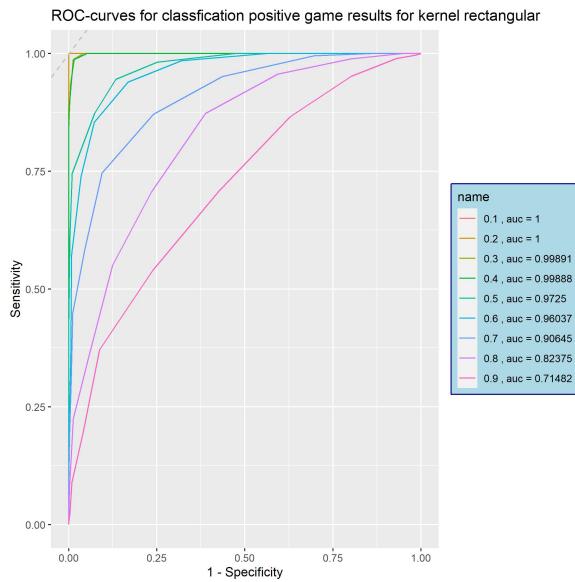


Рис. 27: ROC-кривая классификации положительных результатов с помощью Knn с прямоугольным ядром.

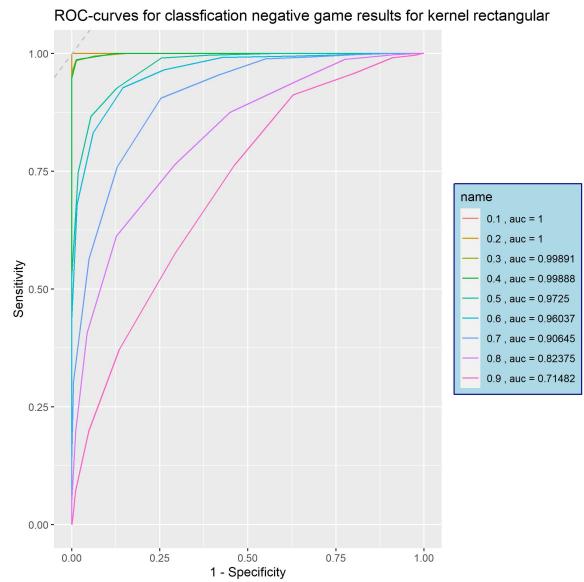


Рис. 28: ROC-кривая классификации негативных результатов с помощью Knn с прямоугольным ядром.

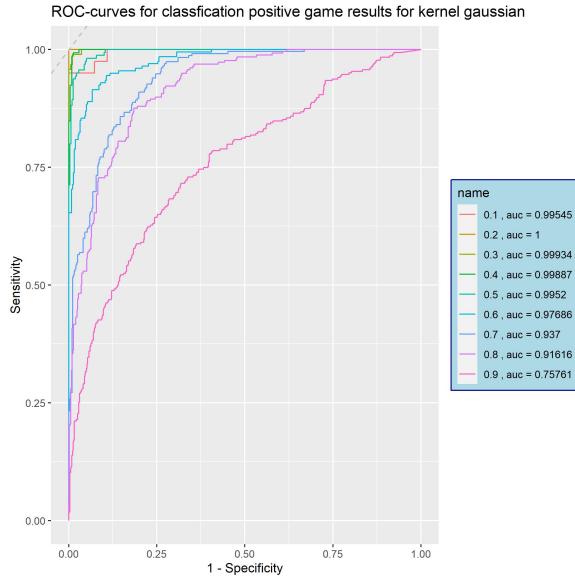


Рис. 29: ROC-кривая классификации положительных результатов с помощью Knn с ядром Гаусса.

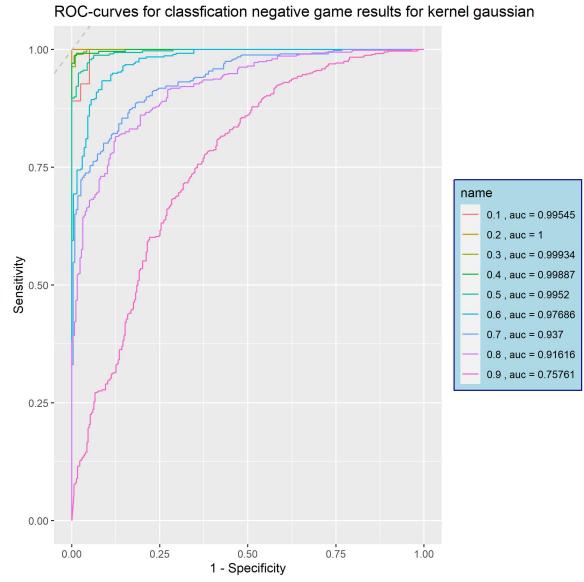


Рис. 30: ROC-кривая классификации негативных результатов с помощью Knn с ядром Гаусса.

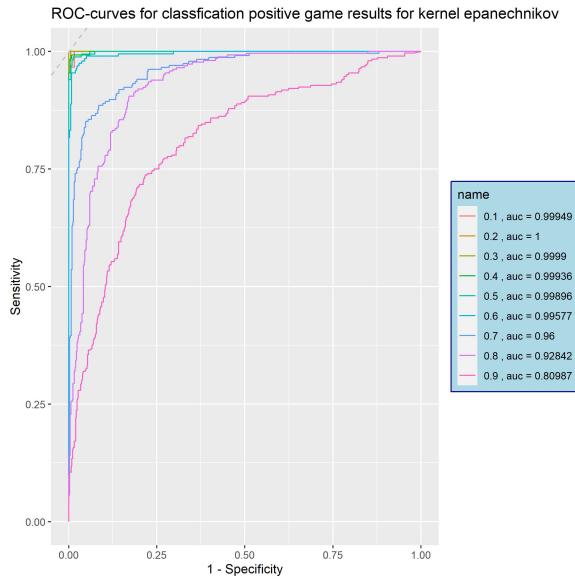


Рис. 31: ROC-кривая классификации положительных результатов с помощью Knn с ядром Епанечникова.

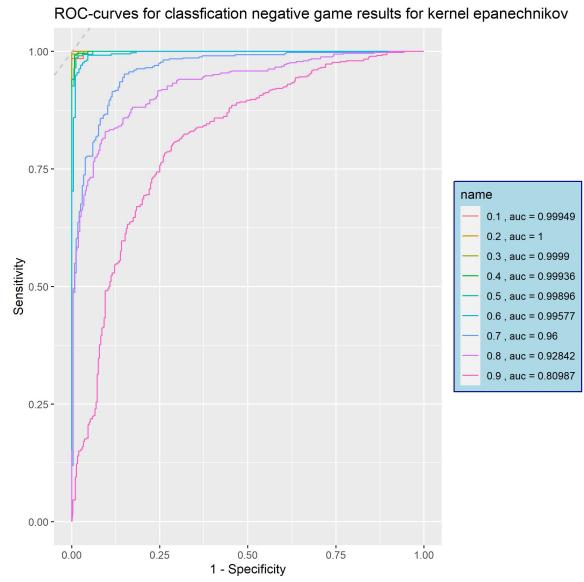


Рис. 32: ROC-кривая классификации негативных результатов с помощью Knn с ядром Епанечникова.

Данные иллюстрации продолжают тенденцию из прошлого пункта. Полученные результаты выглядят подозрительно, так как при малых размерах тестовой выборки по отношению к обучающей площади под кривыми как для положительных, так и для отрицательных результатов игры, приближаются и в некоторых случаях досягают 1, что является эталонным и обычно недостижимым значением в реальных

исследования. Как и ожидалось, в экспериментах, когда обучающая выборка превосходит тестовую, достигаются наилучшие результаты классификации.

В целом видно ухудшение качества классификации с увеличением тестовой выборки и уменьшением обучающего набора, что может свидетельствовать о корректности работы классификатора. С другой же стороны, качество классификации все равно ощутимо превосходит случайный выбор исхода игры, и причину этого хотелось бы установить в дальнейших исследованиях.

В качестве общего вывода по обоим пунктам можно сказать, что классификатор на основе метода  $k$  ближайших соседей работает тем лучше, чем большее отношение обучающей выборки к тестовой (как и в случае с наивным байесовским классификатором).

## 5 Классификация образцов стекла.

Исследуем зависимость качества классификации (Accuracy) от типа ядра, параметра расстояния Миньковского и величины  $k$ .

При определении влияния параметра числа ближайших соседей  $k$  исследуются значения от 2 до 25, параметр Миньковского принимается равным  $p = 2$ , исследуются 4 ядра: прямоугольное, треугольное, ядра Гаусса и Епанечникова. Получены следующие 4 зависимости:

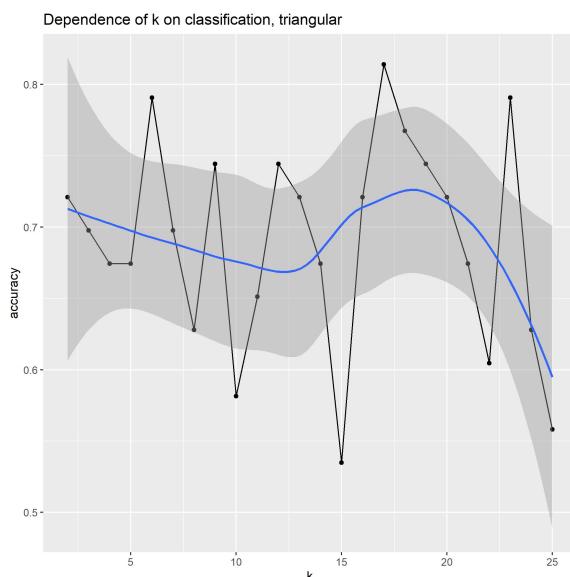


Рис. 33: Зависимость качества классификации от параметра  $k$  на треугольном ядре.

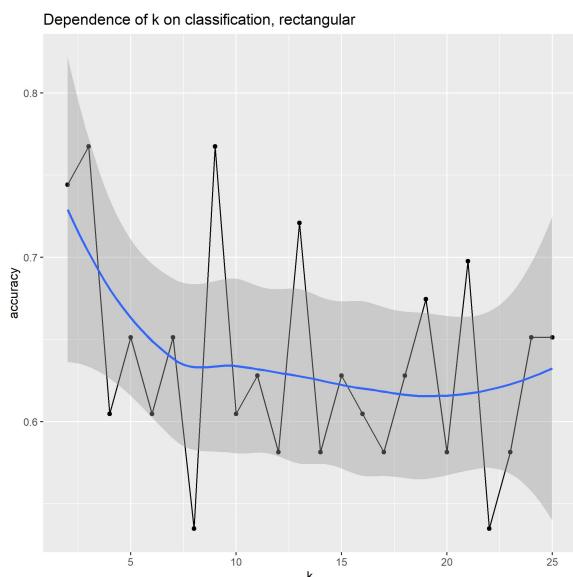


Рис. 34: Зависимость качества классификации от параметра  $k$  на прямоугольном ядре.

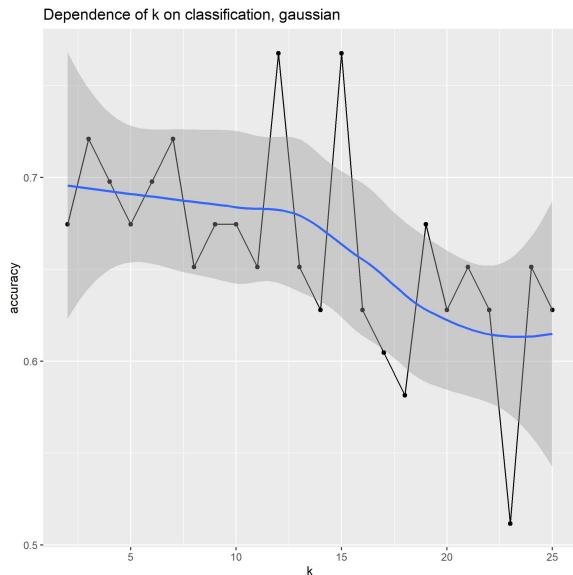


Рис. 35: Зависимость качества классификации от параметра  $k$  на ядре Гаусса.

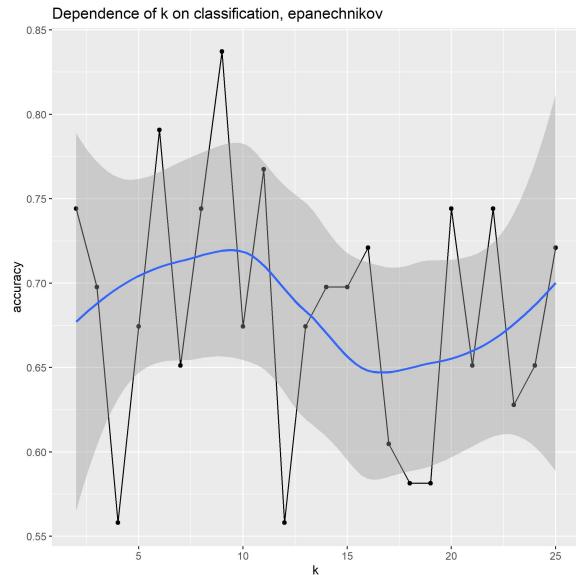


Рис. 36: Зависимость качества классификации от параметра  $k$  на ядре Епанечникова.

Наилучшую ошибку классификации метод  $k$  ближайших соседей показал на треугольном ядре и значение  $k = 17$ . В целом, зависимости для треугольного и Епанечникова ядер имеют явные экстремумы, остальные же 2 в основном состоят из участков монотонного убывания качества классификации с увеличением параметра  $k$ . Не для всех ядер справедлива эвристическая закономерность, согласно которой оптимальное  $k = \sqrt{n}$ , где  $n$  - объем тестовой выборки, как и не получается оптимальное значение при  $k = 7$ , каким оно установлено в программных пакетах R и Python.

При определении влияния параметра Миньковского значение  $p$  исследуются значения от 1 до 10, параметр числа ближайших соседей принимается равным  $k = 7$ , исследуются те же 4 ядра: прямоугольное, треугольное, ядра Гаусса и Епанечникова. Получены следующие 4 зависимости:

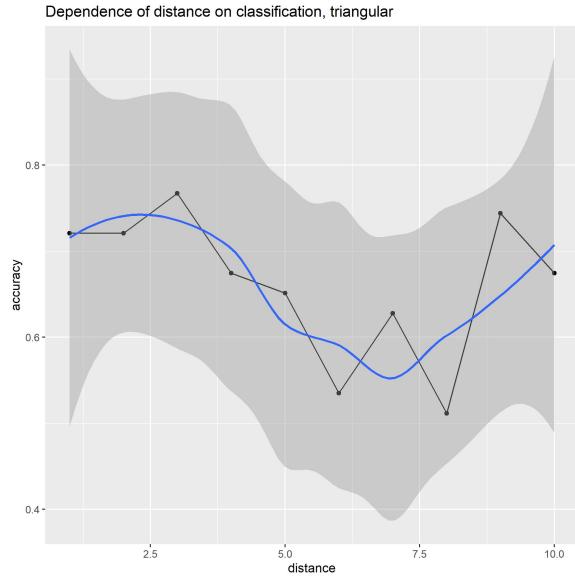


Рис. 37: Зависимость качества классификации от параметра Миньковского на треугольном ядре.

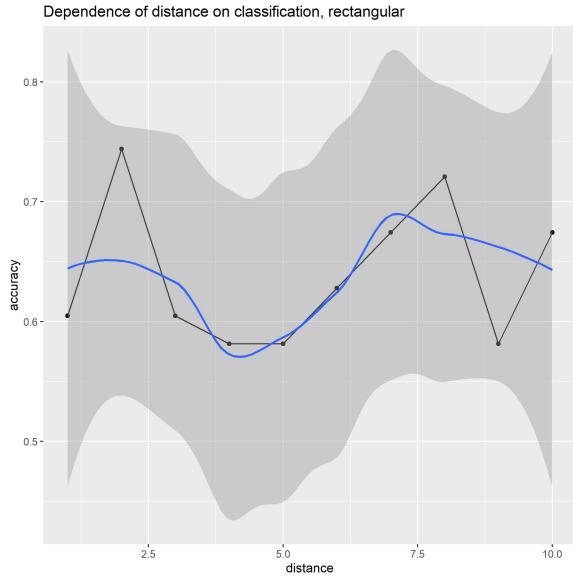


Рис. 38: Зависимость качества классификации от параметра Миньковского на прямоугольном ядре.

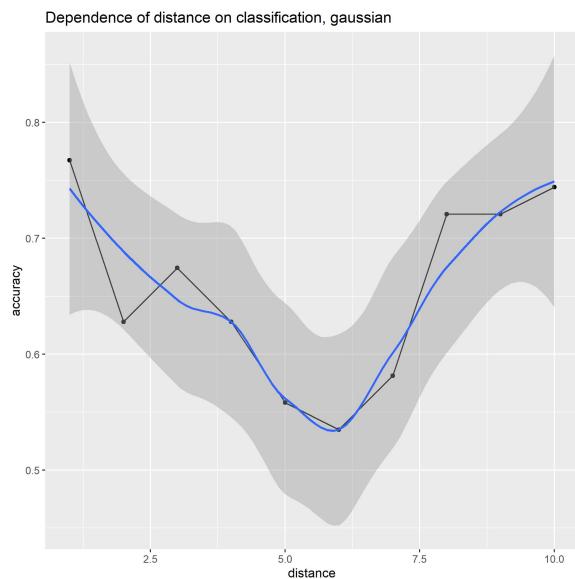


Рис. 39: Зависимость качества классификации от параметра Миньковского на ядре Гаусса.

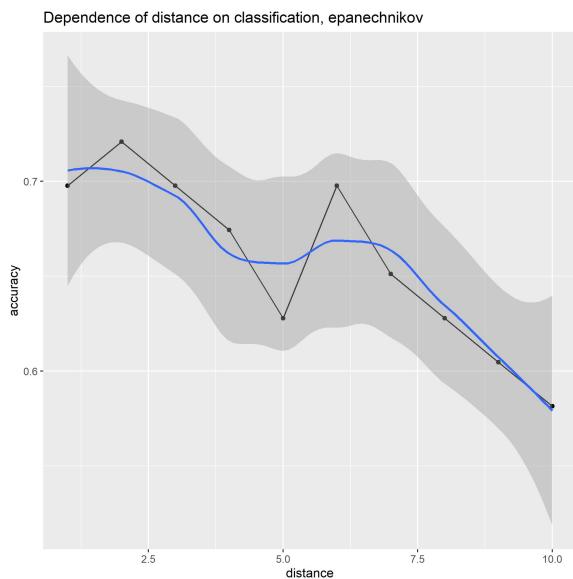


Рис. 40: Зависимость качества классификации от параметра Миньковского на ядре Епанечникова.

В полученных зависимостях параметры Миньковского, при которых достигаются лучшие ошибки классификации, отличаются для каждого ядра. Для треугольного ядра -  $p = 3$ , для прямоугольного -  $p = 2$ , для ядра Гаусса -  $p = 1$ , для ядра Епанечникова -  $p = 2$ . Получается, Евклидова метрика в целом является наиболее оптимальной. Использование метрик более высоких порядков, согласно полученным

графикам, нецелесообразно - достаточно евклидовой и манхэттенской метрик.

Определим теперь, какой из признаков оказывает наибольшее влияние на качество классификации. Для этого последовательно удалим каждый из набора данных и проведем классификацию на оставшемся перечне факторов.

Модель классификации строится при использовании параметров количества ближайших соседей  $k = 7$  и расстояния Миньковского  $p = 2$  на 4 ядрах: прямоугольном, треугольном, Епанчникова и Гаусса. Вычисление значения ошибки классификации происходит путем усреднения по 10 итерациям для каждого признака, таким образом достигается наиболее объективное значение и уменьшается элемент случайности.

Оценка в данном случае будет строиться по значению полученной точности классификации. Итоги проведенного эксперимента представлены на следующей гистограмме:

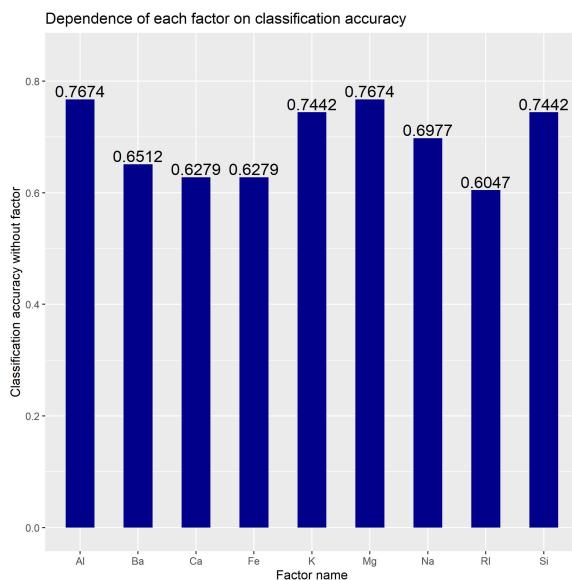


Рис. 41: Гистограмма точности (Accuracy) классификации с треугольным ядром.

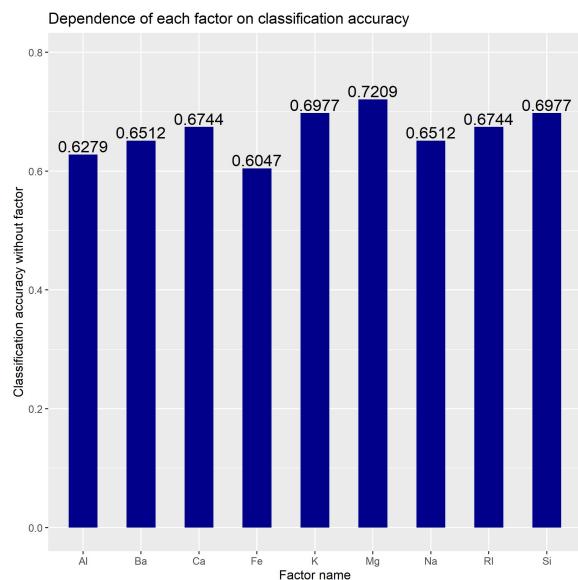


Рис. 42: Гистограмма точности (Accuracy) классификации с прямоугольным ядром.

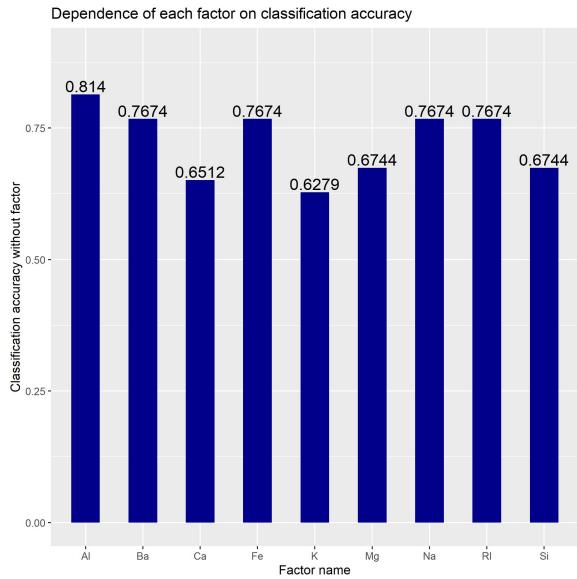


Рис. 43: Гистограмма точности (Accuracy) классификации с ядром Гаусса.

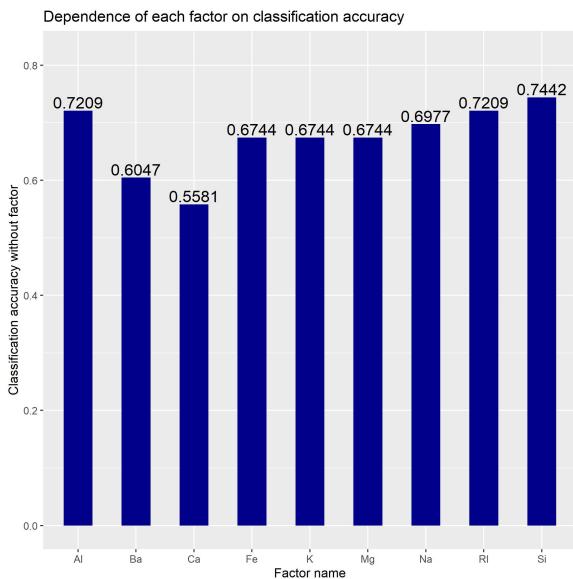


Рис. 44: Гистограмма точности (Accuracy) классификации с ядром Епанечникова.

Гистограммы по своему характеру достаточно сильно отличаются, но можно отметить, что качество классификации по всех случаях сильно уменьшается при удалении показателя содержания Кальция в образце, что наталкивает на вывод о его значимости в исходном наборе данных. Отсутствие же показателя содержания алюминия в 3 из 4 тестов увеличивает ошибку классификации слабее остальных, что позволяет по результатам эксперимента определить как наименее значимый.

Теперь в рамках небольшого теста построим модель на всем имеющемся наборе данных и классифицируем образец стекла с параметрами  $RI = 1.516$ ,  $Na = 11.7$ ,  $Mg = 1.01$ ,  $Al = 1.19$ ,  $Si = 72.59$ ,  $K = 0.43$ ,  $Ca = 11.44$ ,  $Ba = 0.02$ ,  $Fe = 0.1$ .

Классификацию будем проводить также с параметром расстояния Миньковского  $p = 2$ , параметром  $k = 7$ , сглаживание реализуется на 4 ядерных функциях: прямогульной, треугольной, ядрах Гаусса и Епанечникова.

Во всех 4 экспериментах образец был отнесен к классу **5 - «контейнерное стекло»**.

## 6 Классификация цветных точек.

Исходные данные имеют следующие распределения:

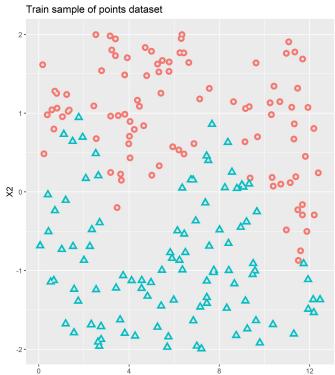


Рис. 45: Точки **обучаю-  
щей** выборки

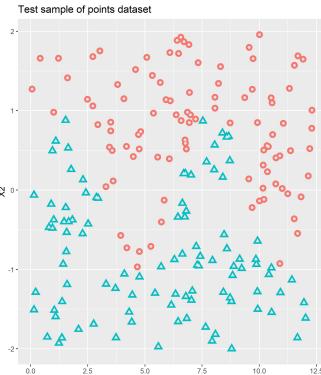


Рис. 46: Точки **тестовой** выборки

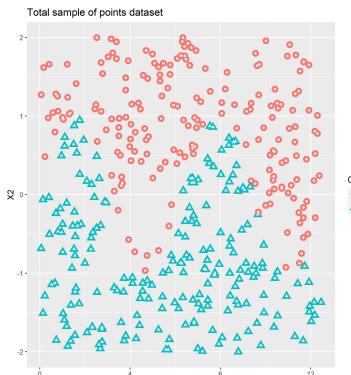


Рис. 47: **Общее** распреде-  
ление точек

Найдем оптимальное значение параметра  $k$  для 4 обозначенных ранее типов ядра: **Прямоугольного, Треугольного, Гауссова и ядра Епанечникова**. Во всех экспериментах параметр расстояния Миньковского имеет значение  $p = 2$ , то есть экземпляры данных сравниваются по Евклидовой метрике:

- Оптимальное число значимых соседей для треугольного ядра:  $k = 3, accuracy = 0.945$ .
- Оптимальное число значимых соседей для прямоугольного ядра:  $k = 2, accuracy = 0.94$ .
- Оптимальное число значимых соседей для ядра Гаусса:  $k = 2, accuracy = 0.94$ .
- Оптимальное число значимых соседей для ядра Епанечникова:  $k = 2, accuracy = 0.94$ .

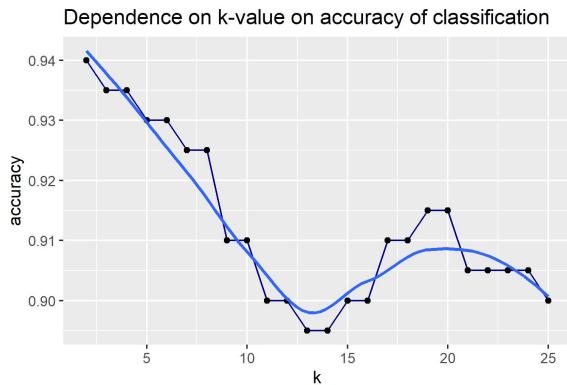


Рис. 48: Зависимость ошибки классифи-  
кации от  $k$  для треугольного ядра

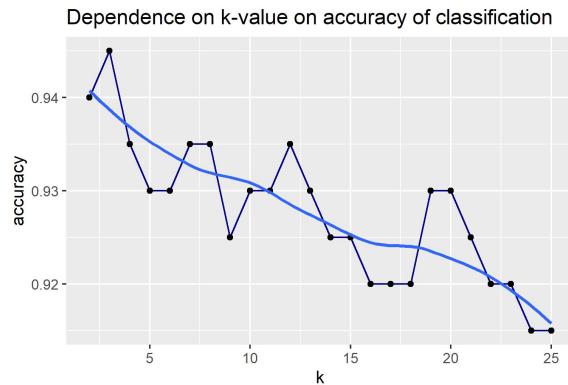


Рис. 49: Зависимость ошибки классифи-  
кации от  $k$  для прямоугольного ядра

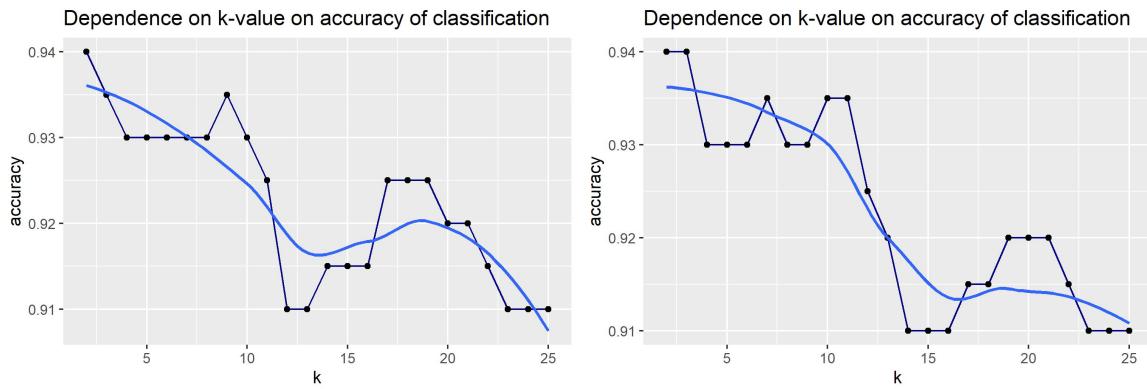


Рис. 50: Зависимость ошибки классификации от  $k$  для ядра Гаусса

Рис. 51: Зависимость ошибки классификации от  $k$  для ядра Епанечникова

Судя по полученным числовым значениям и графическим иллюстрациям результатов, наиболее оптимальное значение числа  $k$  достигается в самом начале экспирентов, на значении 2-3. При этом ошибка классификации находится в районе 6-7 %, а в целом по всем исследованиям для каждого ядра не превышает 10 %.

Стоит отметить отличающийся характер изменения параметра для треугольного ядра, строго монотонный, без явных экстремумов.

## 7 Классификация пассажиров Титаника.

Применим метод  $k$  ближайших соседей также для набора данных о пассажирах Титаника, уже использованный в прошлой работе по исследованию наивного классификатора Байеса.

Классификацию будем проводить при параллельном исследовании полученных значений на разных значениях  $k$ , при этом расстояние между элементами выборки будут оцениваться только по Евклидовой метрике.

Ядро	Треугольное	Премоугольное	Гауссово	Епанечниково
Классы пассажиров				
Выжившие	0.607	0.598	0.583	0.589
Погибшие	0.393	0.402	0.417	0.411

Таблица 4: Таблица результата классификации пассажиров Титаника по Полу, Возрасту, Пассажирским Тарифу и Классу, а также Числу Родственников.

Учитывая значения процента выживаемости, полученные на обучающем наборе данных по тем же признакам, можно сказать, что метод ближайших соседей достаточно хорош на всех использованных масштабирующих функциях. Хотя наиболее

близкий к обучающему набору данных результат получен на треугольном ядре, это не дает оснований оценивать его выше других значений, так как реальных меток классов на валидационном наборе данных не представлено.

## 8 Программная реализация вычислений.

Применение метода ближайших соседей осуществлялось при помощи программных средств языка программирования **R**, в частности таких пакетов, как **kknn**, **kernlab**, **ggplot2**, **hash**, **class** и других. Более подробно ознакомиться с программным кодом и обучающими/тестовыми выборками можно в репозитории по следующей ссылке.

## 9 Заключение.

В результате проведения исследований были получены доказательства высокой эффективности классификации метода k ближайших соседей на различных наборах данных.

Были получены различные зависимости качества классификации от таких гиперпараметров метода, как число ближайших соседей **k**, параметр Миньковского **p** и тип ядра **kernel**.

Среди использованных функций сглаживания наименее эффективным оказалось равномерное (прямоугольное) ядро, остальные же, в целом, показали сравнимую и более высокую эффективность. Оптимальным параметром Миньковского оказалось значение, характеризующее Евклидову метрику, для 3 из 4 рассмотренных ядерных функций, для последнего (прямоугольного) ядра лучшей оказалась Манхэттенская метрика с параметром  $p = 1$ . Значение рассматриваемых ближайших соседей для всех метрик оказалось в районе 2-3, что является аномально малой величиной.

В проведенных экспериментах отношение тестовой выборки к обучающей несильно влияет на качество обучения, что вызывает интерес, потому что эффект должен быть обратным.

## 10 Список использованных источников информации.

1. Сайт преподавателя СПБПУ Л. В. Уткина - основной лекционный и теоретический материал, разбирающий наивный байесовский классификатор.
2. Сайт Школы Анализа Данных - дополнительный лекционный материал с подробным разъяснением компонент метода k ближайших соседей.