Санкт-Петербургский государственный университет

Отчет по научно-исследовательской работе

Тема: **Классификация текстов**

Направление: 010400 (01.03.02) – Прикладная математика и информатика

Выполнил студент гр. 222 _____ Мамаев Владислав Викторович

Научный руководитель,

к. ф.-м. н, доцент _____ Голяндина Нина Эдуардовна

Санкт-Петербург – 2019

Оглавление

1. Введение

Цель работы — познакомиться с методами машинного обучения на примере задачи классификации текстов.

Задачи:

- Описать и реализовать алгоритм, который будет различать тексты двух авторов на основе двух признаков.
- Описать и реализовать алгоритм получения двух наиболее подходящих для применения предыдущего алгоритма признаков из большого числа признаков.

2. Основная часть

2.1. Постановка задачи обучения с учителем

Сформулируем задачу машинного обучения с учителем:

X - множество описаний некоторых объектов, "описания объектов" будем отождествлять с *объектами*, подразумевая преставление объектов в виде элементов множества X. В дальнейшем в качестве объектов будем рассматривать $\mathbf{x} \in \mathbb{R}^m$ — m-мерные вещественные векторы.

Предполагается, что существует *целевая зависимость* — отображение

$$y^* : X \to Y$$

в некторое множество ответов Y, и значения этого отображения на некотором конечном множестве известны.

$X^n = \{\mathbf{x}_i\}_{i=1}^n$ — объекты для которых известны значения целевой функции.

$Y^n = \{y_i = y^*(\mathbf{x}_i), \mathbf{x}_i \in X^n\}_{i=1}^n$ — известные значения функции y^* на этих объектах.

$T^n = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ — набор пар "объект - ответ" будем называть *обучающей выборкой.*

В задаче классификации в качестве множества ответов Y выступает некоторое конечное множество, определяющее принадлежность объекта классу. Далее будем рассматривать задачу бинарной классификации: $|Y| = 2$ и кодировать принадлежность классам *метками* $Y = \{-1, 1\}$. Обозначим:

$$C_1 = \{\mathbf{x}_i \in X^n : y_i = 1\}$$

$$C_2 = \{\mathbf{x}_i \in X^n : y_i = -1\},$$

объекты выборки принадлежащие первому и второму классу соответственно. Задача обучения заключается в построении *алгоритма* или *решающей функции* $a(\mathbf{x}) : X \to Y$, который приближает целевую зависимость y^*.[1]

2.2. Сведение задачи к задаче оптимизации

Одним из подходов к построению $a(x)$ является метод *минимизации эмпирического риска.* Для построенения алгоритма рассматривается:

$$A = \{a(\mathbf{x})\}$$

— *модель алгоритмов*, некоторое семейство функций, вводится

$$L(y, y') : Y^2 \to \mathbb{R}_+ \tag{1}$$

— *функция потерь*, некоторая функция характеризующая величину отклонения ответа $y = a(\mathbf{x})$ от правильного ответа $y' = y^*(\mathbf{x})$ на произвольном объекте $x \in X$, и

$$Q(a, T^n) = \frac{1}{n} \sum_{1}^{n} L(a(\mathbf{x}_i), y_i) \tag{2}$$

— *эмпирический риск*, средняя ошибка a на выборке T^n. Метод заключается в нахождении алгоритма a_{opt}, для которого средняя ошибка на выборке будет минимальна:[2]

$$a_{opt} = \underset{a \in A}{\operatorname{argmin}}\, Q(a, T^n) = \underset{a \in A}{\operatorname{argmin}} \frac{1}{n} \sum_{1}^{n} L(a(\mathbf{x}_i), y_i) \tag{3}$$

Перейдём к описанию используемой модели.

Модель алгортитмов

В качестве модели A будем рассматривать функции вида:

$$a_{\mathbf{w},b}(\mathbf{x}) := sign(\mathbf{w}^{\mathrm{T}}\mathbf{x} - b),$$

где $\mathbf{w} \in \mathbb{R}^m, b \in \mathbb{R}$. Уравнение

$$\mathbf{w}^{\mathrm{T}}\mathbf{x} - b = 0 \tag{4}$$

задает гиперплоскость в пространстве $\mathbb{R}^m$, а выражение $sign(\mathbf{w}^{\mathrm{T}}\mathbf{x} - b)$ показывает в каком из полупространств, полученных при разделении исходного пространства этой гиперплоскостью лежит $\mathbf{x}$.

Определенные таким образом алгоритмы, показывают по какую сторону гиперплоскости лежит объект, а задача обучения при таком семействе алгоритмов — построение оптимальной гиперплоскости.

На рисунке 1 построенная гиперплоскость(прямая) "разделяет"два класса объектов. Выборку для которой возможно построить гиперполскость, такую что объекты одного класса лежат в одном полупространстве, полученном при разделении гиперплоскостью, будем называть *линейно разделимой:*

$$\exists \mathbf{w} : \left[\begin{cases} \mathbf{w}^{\mathrm{T}}\mathbf{x_i} - b < 0, \forall \mathbf{x_i} \in C_1 \\ \mathbf{w}^{\mathrm{T}}\mathbf{x_i} - b > 0, \forall \mathbf{x_i} \in C_2 \end{cases} \right. \left. \begin{cases} \mathbf{w}^{\mathrm{T}}\mathbf{x_i} - b < 0, \forall \mathbf{x_i} \in C_1 \\ \mathbf{w}^{\mathrm{T}}\mathbf{x_i} - b > 0, \forall \mathbf{x_i} \in C_2 \end{cases} \right]$$

Для того чтобы окончательно сформулировать задачу оптимизации, необходимо выбрать функцию потерь L (1).

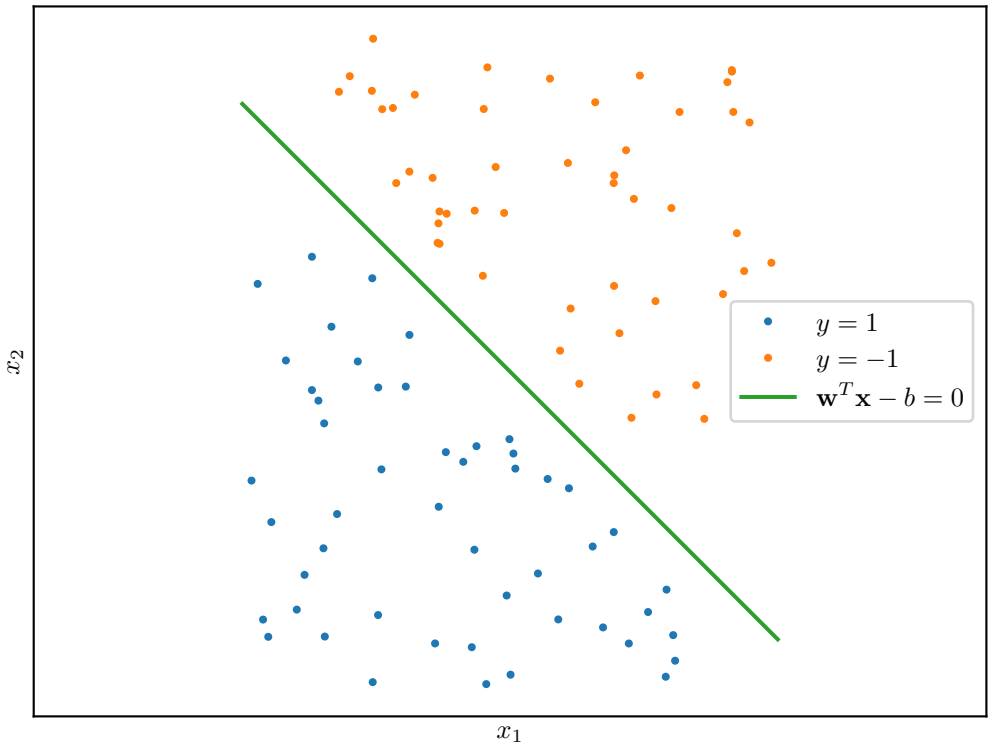

Рис. 1. Линейно разделимая выборка

Функция потерь

Естественным выбором функции потерь в задаче классификации будет:

$$L(y, y') := [y \neq y'] := \begin{cases} 1, y \neq y' \\ 0, y = y' \end{cases} \tag{5}$$

— *пороговая функция потерь*, индикатор ошибки. Функционал эмперического риска (2) при использовании пороговой функции потерь — это доля неверно классифицированных объектов обучающей выборки:

$$Q(a, T^n) = \frac{1}{n}\sum_1^n L(a(\mathbf{x}_i), y_i) = \frac{1}{n}\sum_1^n [a(\mathbf{x}_i) \neq y_i] = \frac{1}{n}\sum_1^n [a(\mathbf{x}_i) \neq y_i] \tag{6}$$

При использовании пороговой функции потерь, задача обучения — минимизация числа неверно классифицируемых объектов обучающей выборки.

2.3. Построение алгоритма классификации

Для алгоритмов из выбранного ранее семейства и множества $Y = \{-1, 1\}$ пороговая функция потерь (5), записывается следующим образом:

$$\begin{aligned} L(a(\mathbf{x}_i), y_i) &= [sign(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i - b) \neq y_i] \\ &= [sign(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i - b)y_i = -1] \\ &= [(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i - b)y_i < 0] \end{aligned} \tag{7}$$

Обозначим:

$$M_a(\mathbf{x}_i) = M_{a_{\mathbf{w},b}}(\mathbf{x}_i) = M_{\mathbf{w},b}(\mathbf{x}_i) = (\mathbf{w}^{\mathrm{T}}\mathbf{x}_i - b)y_i$$

—*отступ* объекта $\mathbf{x}_i$ относительно алгоритма $a = a_{\mathbf{w},b}$. При $|\mathbf{w}| = 1$, отступ — расстояние от объекта до построенной гиперплоскости, взятое с отрицательным знаком, в случае, если объект классифицирован неверно. Функция потерь переписывается в виде:

$$L(a(\mathbf{x}_i), y_i) = [M_a(\mathbf{x}_i) < 0], \tag{8}$$

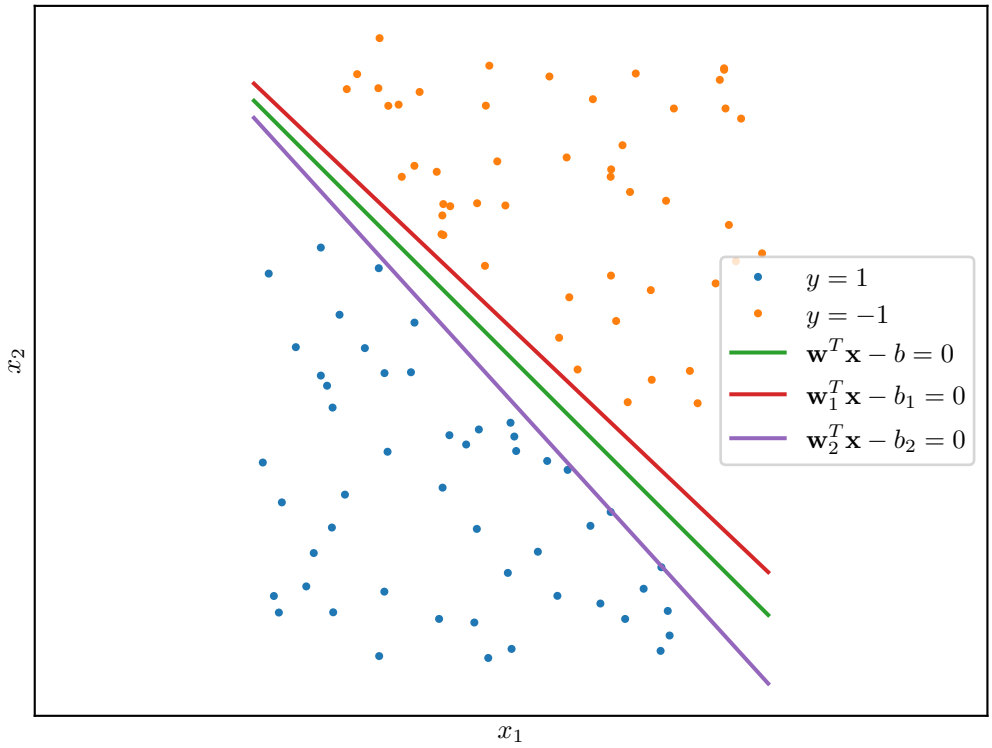

Рис. 2. Неколько разделяющих прямых

а функционал эмперического риска и задача оптимизации:

$$Q(a_{\mathbf{w},b}, T^n) = \frac{1}{n}\sum_{1}^{n}[M_a(\mathbf{x}_i) < 0] \tag{9}$$

$$a_{opt} = \underset{a_{\mathbf{w},b}\in A}{\operatorname{argmin}} Q(a_{\mathbf{w},b}, T^n) = \underset{\mathbf{w}\in\mathbb{R}^n, b\in\mathbb{R}}{\operatorname{argmin}} \frac{1}{n}\sum_{1}^{n}[(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i - b)y_i < 0] \tag{10}$$

Заметим, что уравнение плоскости 4 может быть домножено на некоторое число. В этом случае $w' = \alpha w$ и $b' = \alpha b$ задают ту же гиперплоскость, и решение задачи оптимизации не единственно. Этого можно избежать, задав ограничение на норму $\mathbf{w}$, однако это не решит проблему единственности.

На рисунке 2 можно увидеть, что несколько прямых разделют выборку с нулевой ошибкой: $\forall x : M(x) > 0$.

Попробуем обеспечить единственность в случае линейно разделимой выборки.

Для этого добавим следущее условие на оптимальные $\mathbf{w}', b'$:

$$(\mathbf{w}', b)' = \underset{\mathbf{w},b}{\operatorname{argmax}}(\min_{\mathbf{x}\in X^n} \frac{M_{\mathbf{w},b}(x)}{\|\mathbf{w}\|}) = \underset{\mathbf{w},b}{\operatorname{argmax}}(\frac{1}{\|\mathbf{w}\|} \min_{\mathbf{x}\in X^n} M_{\mathbf{w},b}(x))$$

— оптимальные $\mathbf{w}', b'$ должны максимизировать минимальное расстояние от объекта до разделяющей гиперплоскости.

Отметим, что минимальные для каждого из классов значения отступов, при выполнение этого условия, совпадают:

$$\begin{cases} (\mathbf{w}', b') = \operatorname{argmax}_{\mathbf{w},b} \min_{\mathbf{x}\in X^n} \frac{M_{\mathbf{w},b}(\mathbf{x})}{\|\mathbf{w}\|} \\ \forall i : M_{\mathbf{w}',b'}(\mathbf{x}_i) > 0 \end{cases} \implies \min_{\mathbf{x}\in C_1} M_{\mathbf{w}',b'}(\mathbf{x}) = \min_{\mathbf{x}\in C_2} M_{\mathbf{w}',b'}(\mathbf{x})$$

Доказательство. Предположим ,что это не так, для определенности пусть:

$$\min_{\mathbf{x}\in C_1} M_{\mathbf{w}',b'}(\mathbf{x}) > \min_{\mathbf{x}\in C_2} M_{\mathbf{w}',b'}(\mathbf{x}),$$

Существуют объекты разных классов на которых достигаются минимумы:

$$\exists \mathbf{x}_1 \in C_1, \mathbf{x}_2 \in C_2, \mathbf{x}_1 = \underset{\mathbf{x}\in C_1}{\operatorname{argmin}} M_{\mathbf{w}',b'}(\mathbf{x}), \mathbf{x}_2 = \underset{\mathbf{x}\in C_2}{\operatorname{argmin}} M_{\mathbf{w}',b'}(\mathbf{x})$$

При этом:

$$M_{\mathbf{w}',b'}(\mathbf{x}_1) > M_{\mathbf{w}',b'}(\mathbf{x}_2)$$

Тогда положив $b'' = b' - \frac{M_{\mathbf{w}',b'}(\mathbf{x}_1) - M_{\mathbf{w}',b'}(\mathbf{x}_2)}{2}$:

$$\begin{aligned} M_{\mathbf{w}',b''}(\mathbf{x}_i) &= (\mathbf{w}'^{\mathrm{T}}\mathbf{x}_2 - b'')y_i \\ &= (\mathbf{w}'^{\mathrm{T}}\mathbf{x}_i - b' - \frac{M_{\mathbf{w}',b'}(\mathbf{x}_1) - M_{\mathbf{w}',b'}(\mathbf{x}_2)}{2})y_i \\ &= M_{\mathbf{w}',b'}(\mathbf{x}_i) - \frac{M_{\mathbf{w}',b'}(\mathbf{x}_1) - M_{\mathbf{w}',b'}(\mathbf{x}_2)}{2} y_i \end{aligned}$$

$$\begin{aligned} M_{\mathbf{w}',b''}(\mathbf{x}_1) &= \frac{M_{\mathbf{w}',b'}(\mathbf{x}_1) + M_{\mathbf{w}',b'}(\mathbf{x}_2)}{2} \\ M_{\mathbf{w}',b''}(\mathbf{x}_2) &= \frac{M_{\mathbf{w}',b'}(\mathbf{x}_1) + M_{\mathbf{w}',b'}(\mathbf{x}_2)}{2} \end{aligned}$$

— получим равенство отступов. Покажем, что минимальность для каждого из классов сохраняется:

$$\begin{aligned}\min_{\mathbf{x}_i \in C_1} M_{\mathbf{w}',b''}(\mathbf{x}_i) &= \min_{\mathbf{x}_i \in C_1}(M_{\mathbf{w}',b'}(\mathbf{x}_i) - \frac{M_{\mathbf{w}',b'}(\mathbf{x}_1) - M_{\mathbf{w}',b'}(\mathbf{x}_2)}{2} y_i) \\ &= \min_{\mathbf{x}_i \in C_1}(M_{\mathbf{w}',b'}(\mathbf{x}_i) - \frac{M_{\mathbf{w}',b'}(\mathbf{x}_1) - M_{\mathbf{w}',b'}(\mathbf{x}_2)}{2}) \\ &= \min_{\mathbf{x}_i \in C_1}(M_{\mathbf{w}',b'}(\mathbf{x}_i)) - \frac{M_{\mathbf{w}',b'}(\mathbf{x}_1) - M_{\mathbf{w}',b'}(\mathbf{x}_2)}{2}\end{aligned}$$

$$\begin{aligned}\min_{\mathbf{x}_i \in C_2} c &= \min_{\mathbf{x}_i \in C_2}(M_{\mathbf{w}',b'}(\mathbf{x}_i) - \frac{M_{\mathbf{w}',b'}(\mathbf{x}_1) - M_{\mathbf{w}',b'}(\mathbf{x}_2)}{2} y_i) \\ &= \min_{\mathbf{x}_i \in C_2}(M_{\mathbf{w}',b'}(\mathbf{x}_i) + \frac{M_{\mathbf{w}',b'}(\mathbf{x}_1) - M_{\mathbf{w}',b'}(\mathbf{x}_2)}{2}) \\ &= \min_{\mathbf{x}_i \in C_2}(M_{\mathbf{w}',b'}(\mathbf{x}_i)) + \frac{M_{\mathbf{w}',b'}(\mathbf{x}_1) - M_{\mathbf{w}',b'}(\mathbf{x}_2)}{2}\end{aligned}$$

Из этого следует, что $(\mathbf{w}, b'')$ оптимальнее $(\mathbf{w}, b')$:

$$\min_{\mathbf{x} \in X} M_{\mathbf{w}',b''}(\mathbf{x}) = \min_{\mathbf{x} \in C_1} M_{\mathbf{w}',b''}(\mathbf{x}) = \min_{\mathbf{x} \in C_2} M_{\mathbf{w}',b''}(\mathbf{x}_i) > \min_{\mathbf{x} \in C_2} M_{\mathbf{w}',b'}(\mathbf{x}_i) = \min_{\mathbf{x} \in X} M_{\mathbf{w}',b'}(\mathbf{x})$$

□

Положим:

$$\min_{\mathbf{x} \in X^n} M_{\mathbf{w}',b'}(\mathbf{x}) = 1$$

тогда задача оптимизации преобразуется

$$\begin{cases} (\mathbf{w}', b') = \operatorname{argmax}_{\mathbf{w},b} \min_{\mathbf{x} \in X^n} \frac{M_{\mathbf{w},b}(\mathbf{x})}{\|\mathbf{w}\|} \\ \forall i : M_{\mathbf{w}',b'}(\mathbf{x}_i) > 0 \\ \min_{\mathbf{x} \in X^n} M_{\mathbf{w}',b'}(\mathbf{x}) = 1 \end{cases} \Longrightarrow$$

$$\Longrightarrow \begin{cases} (\mathbf{w}', b') = \operatorname{argmax}_{\mathbf{w},b} \frac{1}{\|\mathbf{w}\|} \\ \forall i : M_{\mathbf{w}',b'}(\mathbf{x}_i) \geq 1 \end{cases} \Longrightarrow$$

$$\Longrightarrow \begin{cases} (\mathbf{w}', b') = \operatorname{argmin}_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 \\ \forall i : M_{\mathbf{w}',b'}(\mathbf{x}_i) \geq 1 \end{cases}$$

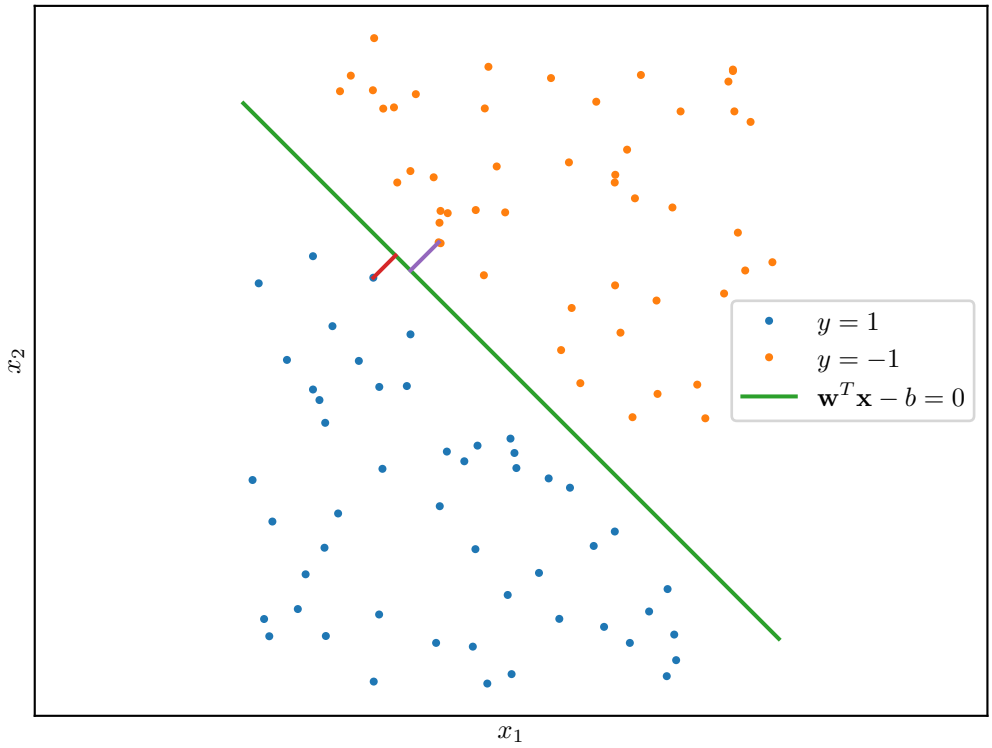

Рис. 3. Наименьшие отступы

Рассмотрев случай линейно разделимой выборки, попробуем адаптировать поставленную задачу для выборки не являющейся линейно разделимой. Для этого ослабим ограничения, позволив оступам быть меньше единицы и даже отрицательными. Разрешив ошибки построенного алгоритма на обучающей выборке, добавим к минимизируемой функции сумму этих ошибок. Получим следущую задачу:

$$\begin{cases} (\mathbf{w}', b') = \text{argmin}_{\mathbf{w},b,\xi}\, \frac{1}{2}(\|\mathbf{w}\|^2 + C\sum_1^n \xi_i) \\ \forall i : M_{\mathbf{w}',b'}(\mathbf{x}_i) \geq 1 - \xi_i \\ \forall i : \xi_i \geq 0 \end{cases}$$

Применим построенный алгоритм и оценим результаты.

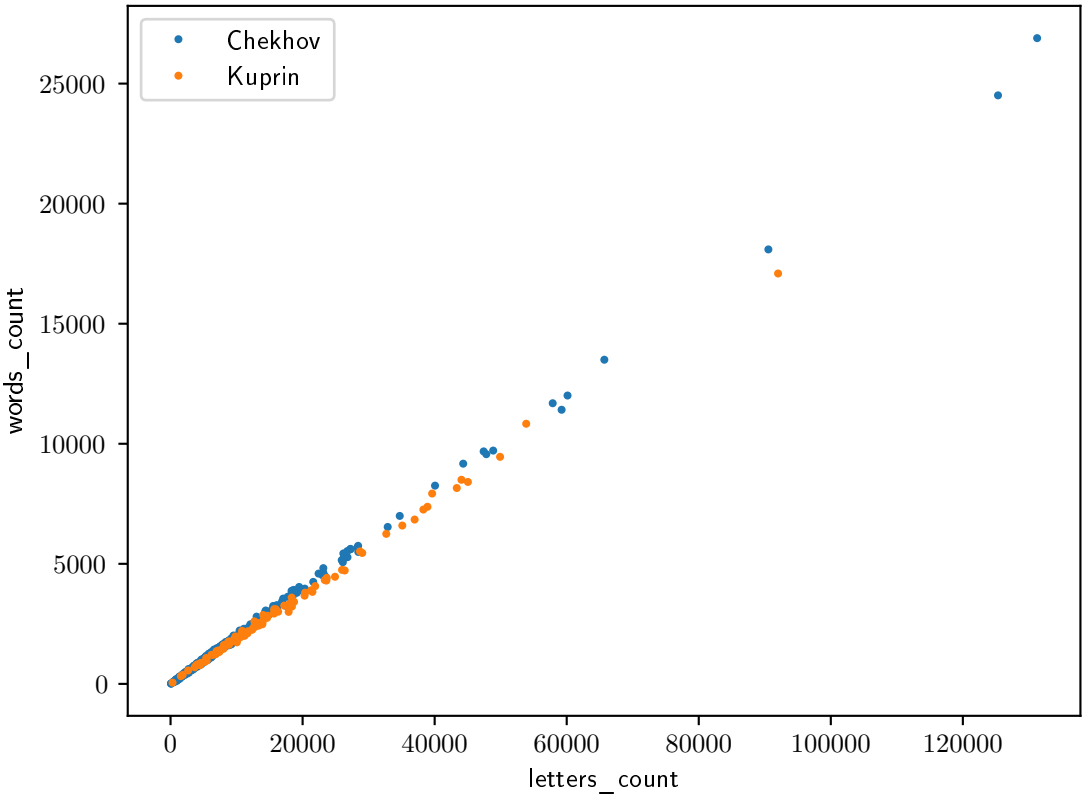

Рис. 4. Диаграмма рассеяния по количеству букв и количеству слов

2.4. Оценка результатов работы алгоритма

Для применения построеного алгоритма были собраны 669 рассказов А. П. Чехова и А. И Куприна. И посчитаны некоторые численные характеристики каждого из рассказов.

Для оценки результатов, все рассказазы разбиты на обучающую и тестовые выборки. Построив диаграммы рассеяния рисунки 4,5,6,7 по некоторым признакам, было решено провести обучение на некоторых парах признаков. Рисунки 8,9,10,11 демонстрируют постренные разделяющие прямые. Таблицы 1,3,5,7 демонстрируют точность классификации на тестовой выборке при $C = 1$. Таблицы 2,4,6,8 демонстрируют точность классификации на тестовой выборке при использовании скользящего контроля для выбора C.

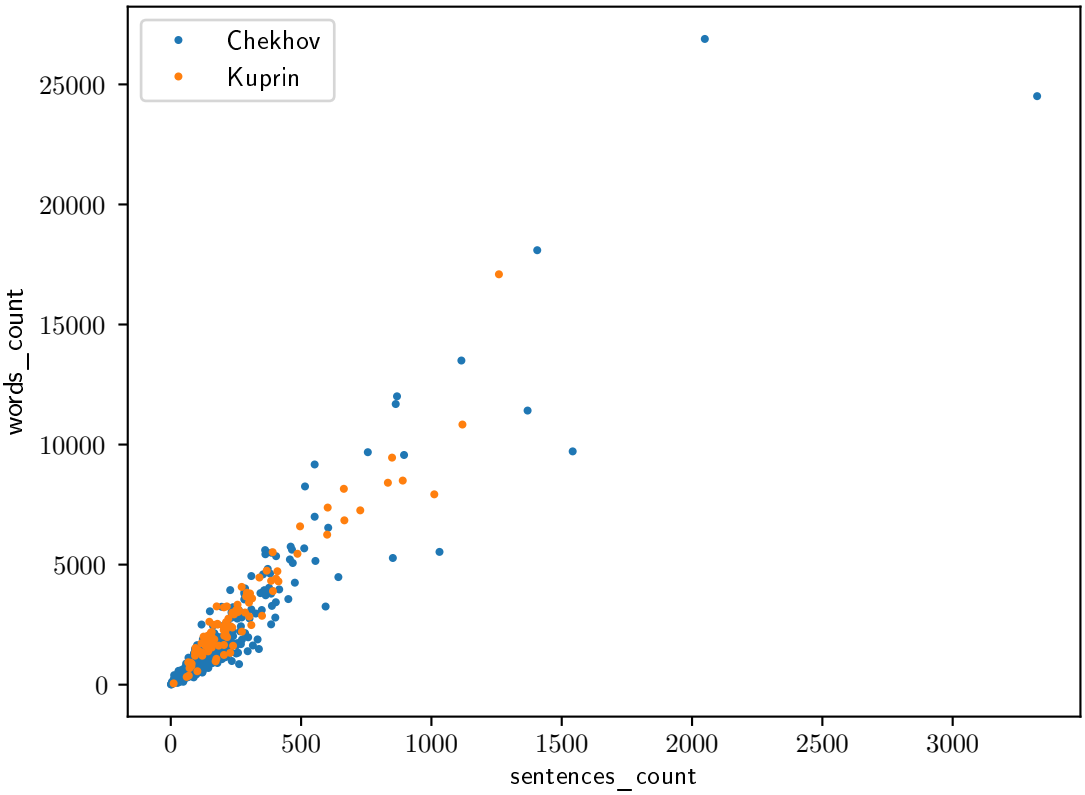

Рис. 5. Диаграмма рассеяния по количеству слов и количеству предложений

	presision	recall	count
class1	0.96	0.87	170
class2	0.53	0.81	31
accuracy	0.86		

Таблица 1. Результаты работы алгоритма для признаков: количество символов, количество слов, число слов.$C = 1$

	presision	recall	count
class1	0.96	0.87	170
class2	0.53	0.81	31
accuracy	0.86		

Таблица 2. Результаты работы алгоритма для признаков: количество символов, количество слов. C выбрано используя скользящий контроль

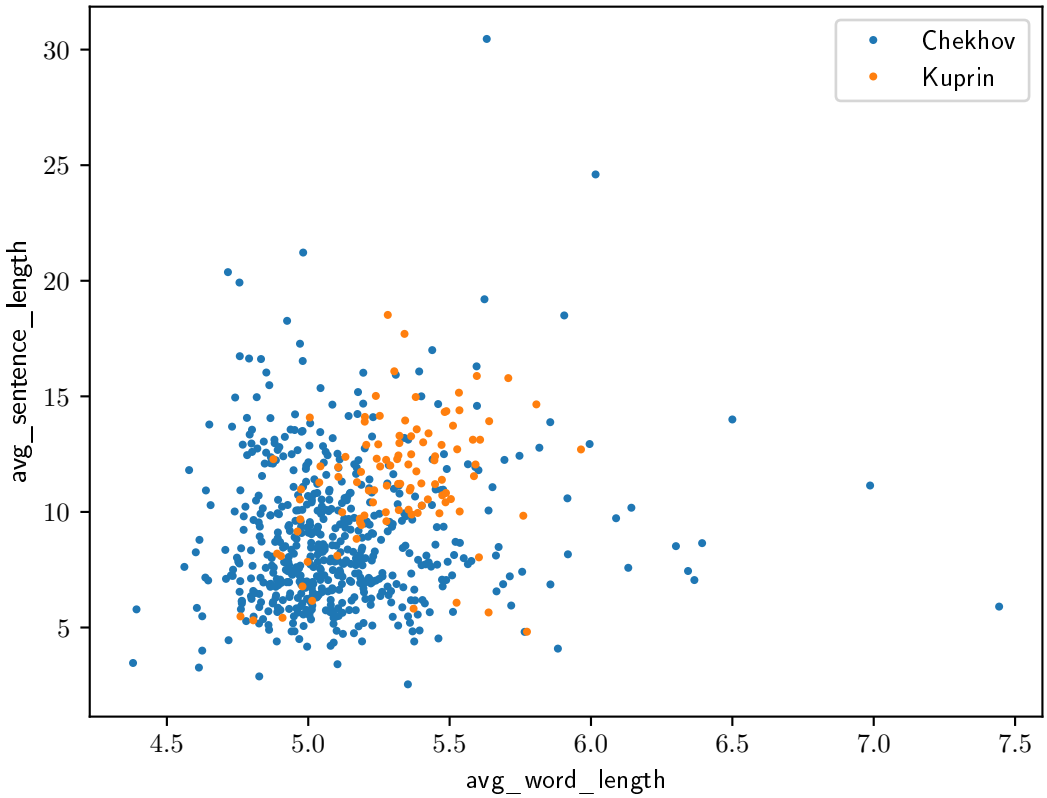

Рис. 6. Диаграмма рассеяния по средней длине слова и средней длине предложения

	presision	recall	count
class1	0.92	0.90	170
class2	0.50	0.55	31
accuracy	0.85		

Таблица 3. Результаты работы алгоритма для признаков: количество предложений, количество слов.$C = 1$

	presision	recall	count
class1	0.92	0.90	170
class2	0.50	0.55	31
accuracy	0.85		

Таблица 4. Результаты работы алгоритма для признаков: количество предложений, количество слов. C выбрано используя скользящий контроль

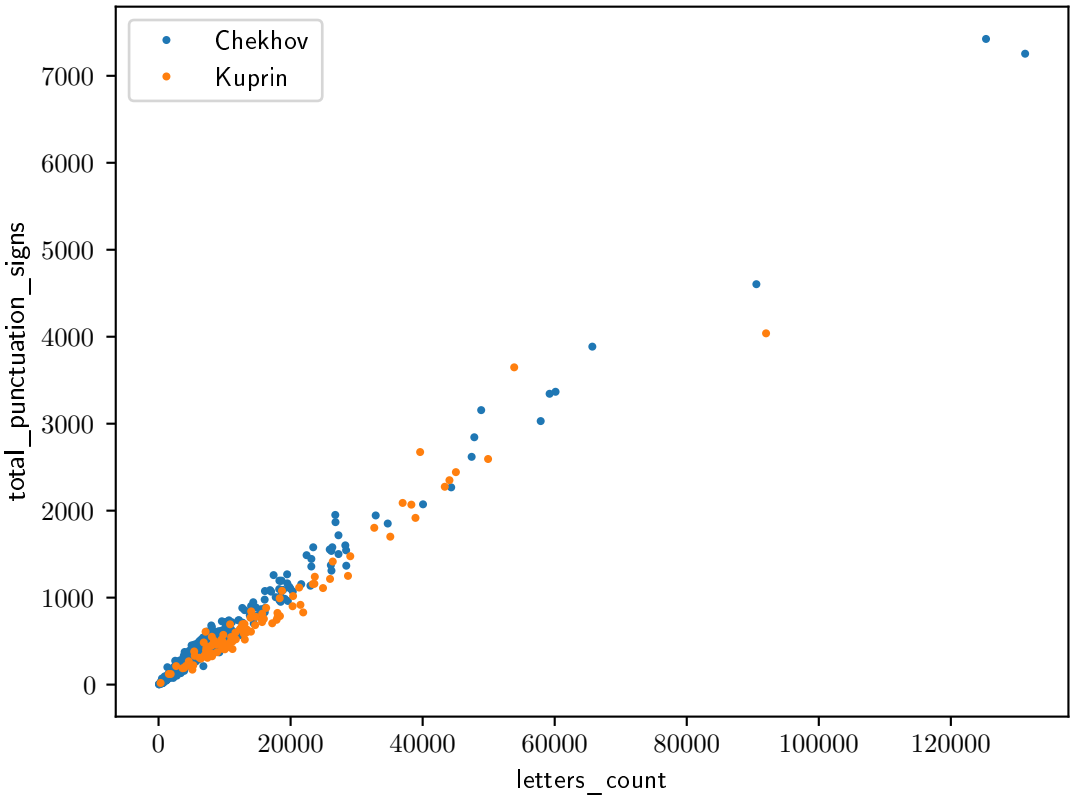

Рис. 7. Диаграмма рассеяния по количеству букв и количеству знаков препинания

	presision	recall	count
class1	0.88	0.96	170
class2	0.53	0.26	31
accuracy	0.85		

Таблица 5. Результаты работы алгоритма для признаков: средня длина слова, средня длина предложения. $C = 1$

	presision	recall	count
class1	0.89	0.87	170
class2	0.37	0.42	31
accuracy	0.80		

Таблица 6. Результаты работы алгоритма для признаков: средня длина слова, средня длина предложения. C выбрано используя скользящий контроль

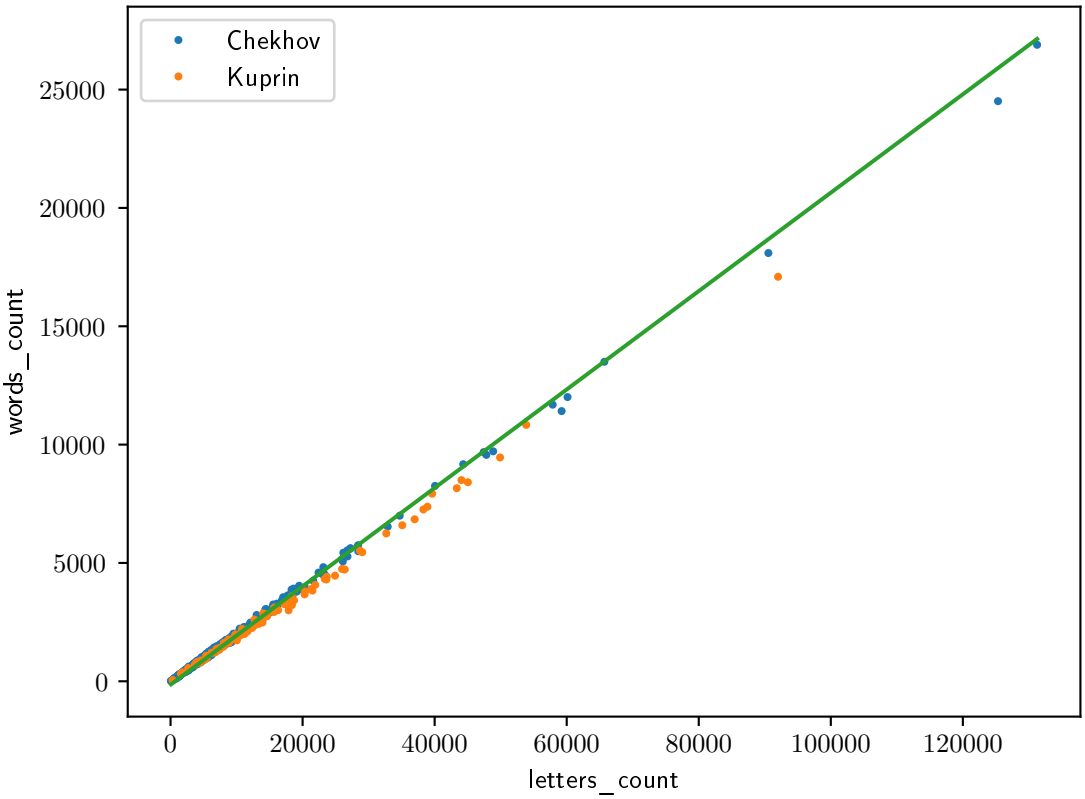

Рис. 8. Диаграмма рассеяния по количеству букв и количеству слов и разделяющая прямая

	presision	recall	count
class1	0.92	0.93	170
class2	0.59	0.55	31
accuracy	0.87		

Таблица 7. Результаты работы алгоритма для признаков: количество символов, количество знаков пунктуации. $C = 1$

	presision	recall	count
class1	0.92	0.93	170
class2	0.59	0.55	31
accuracy	0.87		

Таблица 8. Результаты работы алгоритма для признаков: количество знаков пунктуации. C выбрано используя скользящий контроль

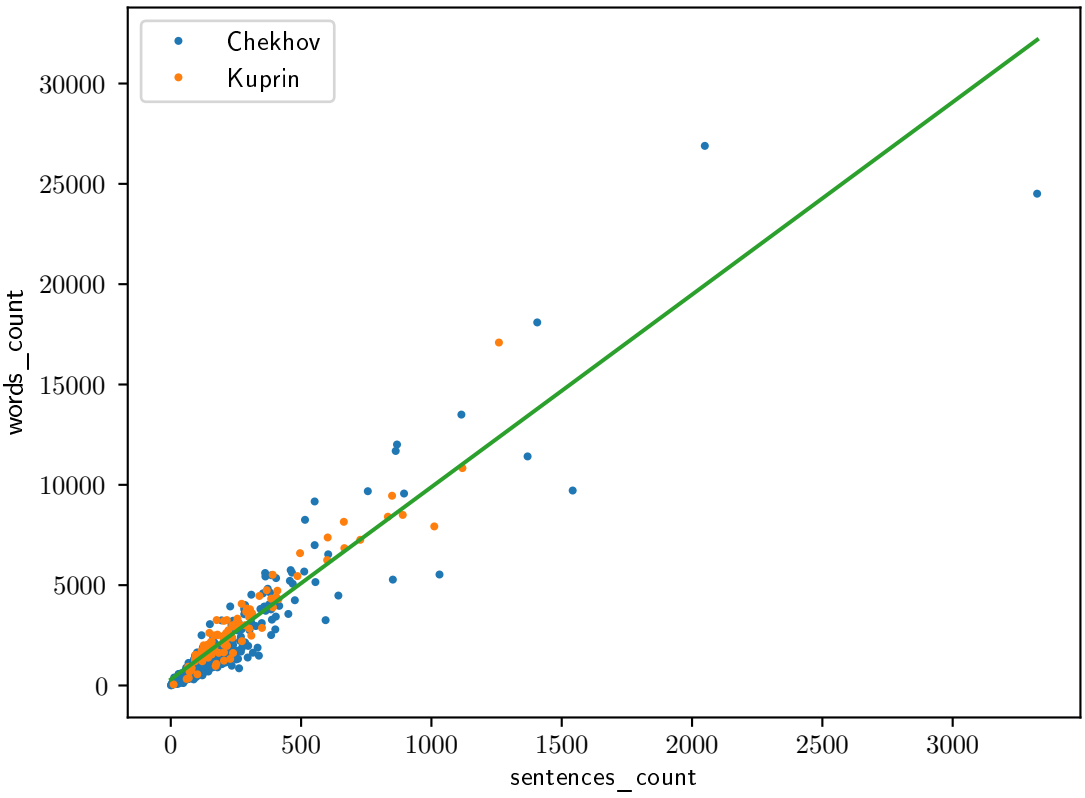

Рис. 9. Диаграмма рассеяния по количеству слов и количеству предложений и разделяющая прямая

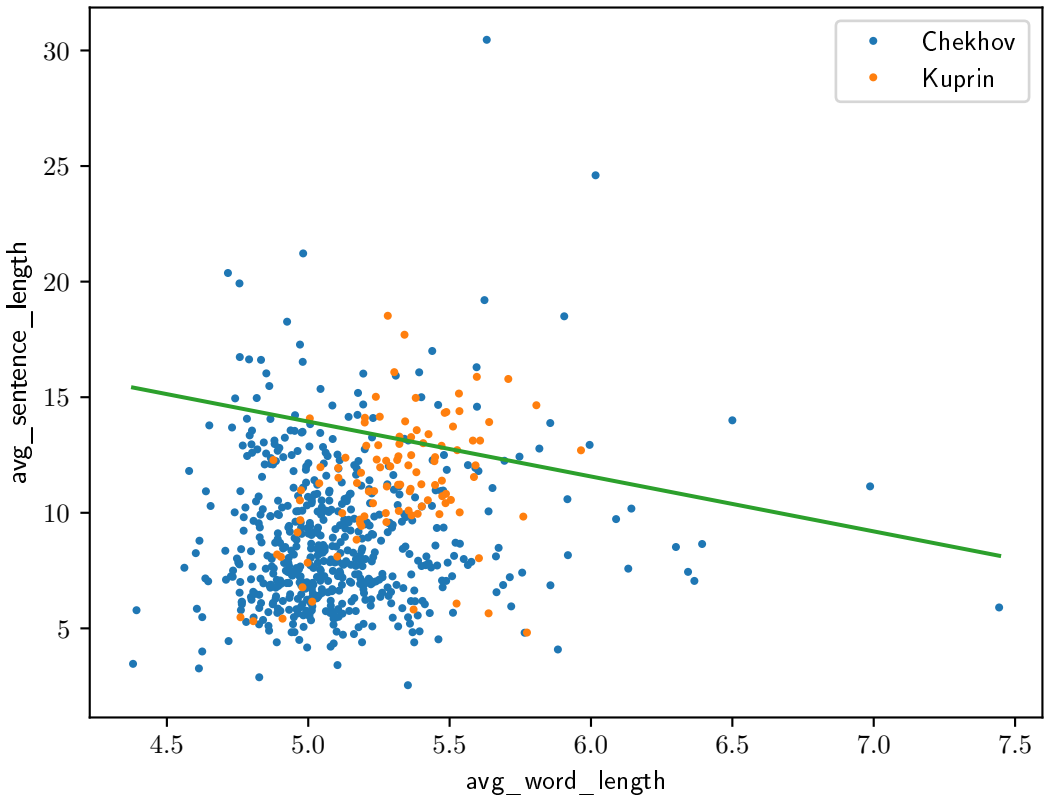

Рис. 10. Диаграмма рассеяния по средней длине слова и средней длине предложения и разделяющая прямая

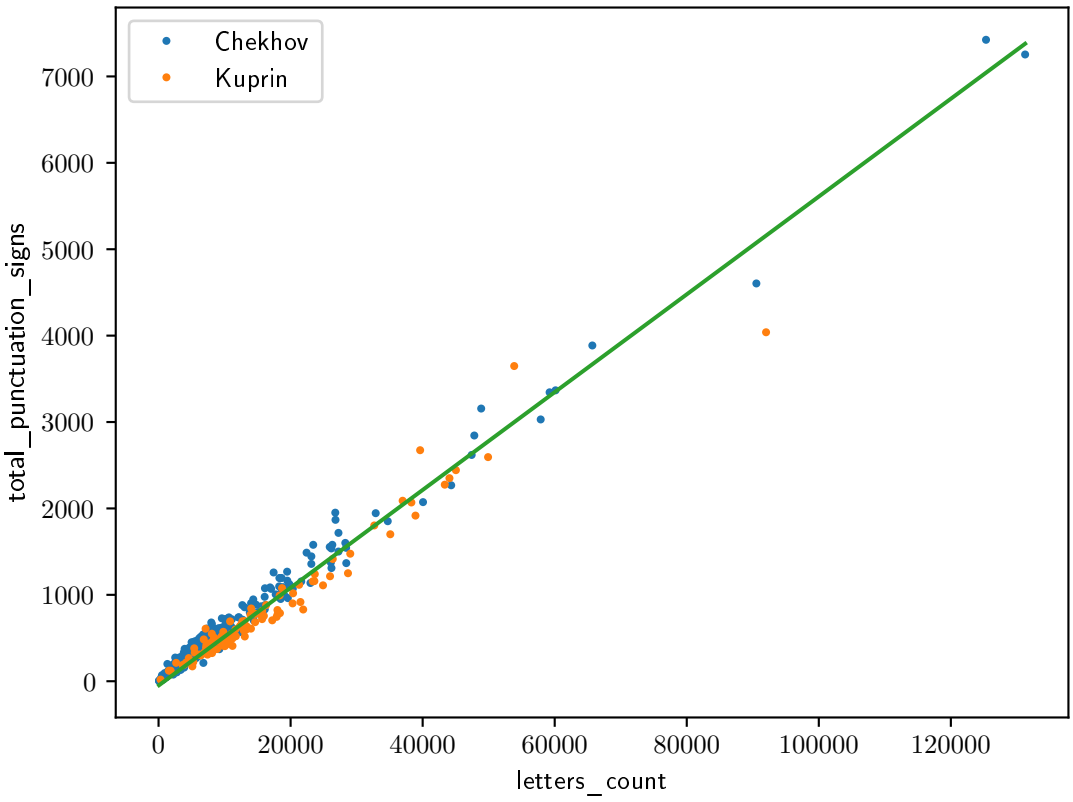

Рис. 11. Диаграмма рассеяния по количеству букв и количеству знаков препинания и разделяющая прямая

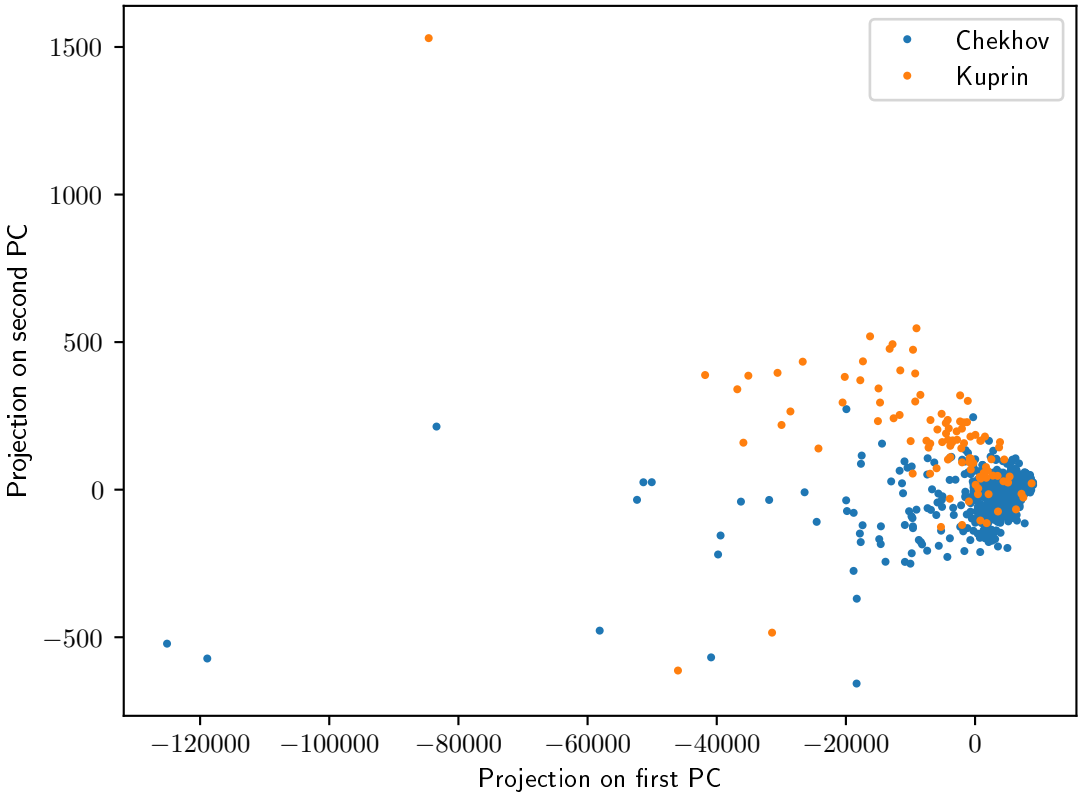

Рис. 12. Проекция на главные компоненты.

2.5. Метод главных компонент

Попробуем из m признаков получить 2, которые лучше других описывают объекты обучающей выборки: для этого вычислим сингулярное разложение центрированной и транспонированное матрицы объектов и потроим проекции объектов выборки на сингулярные вектора соответствующие двум наибольшим сингулярным числам.[3]

2.6. Оценка результатов работы модифицированного алгоритма

Рисунок 13 и таблица 9 демонстрируют результаты работы алгоритма с предварительным построением двух признаков.

3. Заключение

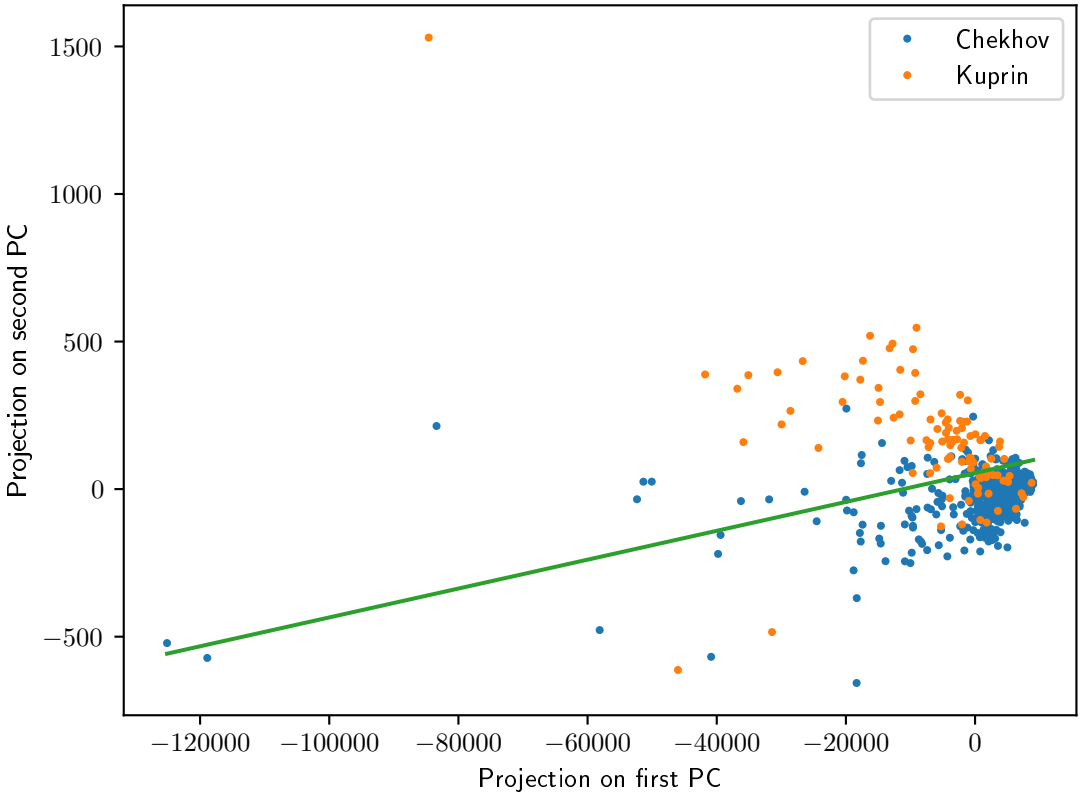

Рис. 13. Проекция на главные компоненты и разделяющая прямая

	presision	recall	count
class1	0.96	0.91	170
class2	0.61	0.81	31
accuracy	0.89		

Таблица 9. Результаты работы модифицированного алгоритма

Список литературы

1. Обучение по предцедентам. — URL: http://www.machinelearning.ru/wiki/index.php?title=%D0%9E%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%BB8%D0%B5_%D0%BF%D0%BE_%D0%BF%D1%80%D0%B5%D1%86%D0%B5%D0%B4%D0%B5%D0%BD%D1%82%D0%B0%D0%BC.

2. Минимизация эмпирического риска. — URL: http://www.machinelearning.ru/wiki/index.php?title=%D0%9C%D0%B8%D0%BD%D0%B8%D0%BC%D0%B8%D0%B7%D0%B0%D1%86%D0%B8%D1%8F_%D1%8D%D0%BC%D0%BF%D0%B8%D1%80%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%BE%D0%B3%D0%BE_%D1%80%D0%B8%D1%81%D0%BA%D0%B0.

3. *Н.Э. Г.* Метод "Гусеница"-SSA: Анализ временных рядов. Учебное пособие. — Санкт-Петергбург, 2004. — URL: http://www.gistatgroup.com/gus/ssa_an.pdf.