Министерство образования Республики Беларусь

Учреждение образования

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ

Факультет компьютерных систем и сетей

Кафедра информатики

Дисциплина: Машинное обучение (МО)

Отчет о лабораторной работе №5

«Метод опорных векторов»

|  |
| --- |
| **Выполнил:** |
| **Студент гр. 858341** |
| **Немкович А. В.** |

|  |
| --- |
| **Проверил:** |
| **Стержанов М. В.** |

**Минск 2019**

Содержание

[Постановка задачи 3](#_Toc25275643)

[Ход выполнения 5](#_Toc25275644)

[Выводы 14](#_Toc25275645)

# Постановка задачи

Набор данных **ex5data1.mat** представляет собой файл формата \*.mat (т.е. сохраненного из Matlab). Набор содержит три переменные X1 и X2 (независимые переменные) и y (метка класса). Данные являются линейно разделимыми.

Набор данных **ex5data2.mat** представляет собой файл формата \*.mat (т.е. сохраненного из Matlab). Набор содержит три переменные X1 и X2 (независимые переменные) и y (метка класса). Данные являются нелинейно разделимыми.

Набор данных **ex5data3.mat** представляет собой файл формата \*.mat (т.е. сохраненного из Matlab). Набор содержит три переменные X1 и X2 (независимые переменные) и y (метка класса). Данные разделены на две выборки: обучающая выборка (X, y), по которой определяются параметры модели; валидационная выборка (Xval, yval), на которой настраивается коэффициент регуляризации и параметры Гауссового ядра.

Набор данных **spamTrain.mat** представляет собой файл формата \*.mat (т.е. сохраненного из Matlab). Набор содержит две переменные X - вектор, кодирующий отсутствие (0) или присутствие (1) слова из словаря vocab.txt в письме, и y - метка класса: 0 - не спам, 1 - спам. Набор используется для обучения классификатора.

Набор данных **spamTest.mat** представляет собой файл формата \*.mat (т.е. сохраненного из Matlab). Набор содержит две переменные Xtest - вектор, кодирующий отсутствие (0) или присутствие (1) слова из словаря vocab.txt в письме, и ytest - метка класса: 0 - не спам, 1 - спам. Набор используется для проверки качества классификатора.

**Задание.**

1. Загрузите данные **ex5data1.mat** из файла.
2. Постройте график для загруженного набора данных: по осям - переменные X1, X2, а точки, принадлежащие различным классам должны быть обозначены различными маркерами.
3. Обучите классификатор с помощью библиотечной реализации SVM с линейным ядром на данном наборе.
4. Постройте разделяющую прямую для классификаторов с различными параметрами C = 1, C = 100 (совместно с графиком из пункта 2). Объясните различия в полученных прямых?
5. Реализуйте функцию вычисления Гауссового ядра для алгоритма SVM.
6. Загрузите данные **ex5data2.mat** из файла.
7. Обработайте данные с помощью функции Гауссового ядра.
8. Обучите классификатор SVM.
9. Визуализируйте данные вместе с разделяющей кривой (аналогично пункту 4).
10. Загрузите данные **ex5data3.mat** из файла.
11. Вычислите параметры классификатора SVM на обучающей выборке, а также подберите параметры C и σ2 на валидационной выборке.
12. Визуализируйте данные вместе с разделяющей кривой (аналогично пункту 4).
13. Загрузите данные **spamTrain.mat** из файла.
14. Обучите классификатор SVM.
15. Загрузите данные **spamTest.mat** из файла.
16. Подберите параметры C и σ2.
17. Реализуйте функцию предобработки текста письма, включающую в себя:
    1. перевод в нижний регистр;
    2. удаление HTML тэгов;
    3. замена URL на одно слово (например, “httpaddr”);
    4. замена email-адресов на одно слово (например, “emailaddr”);
    5. замена чисел на одно слово (например, “number”);
    6. замена знаков доллара ($) на слово “dollar”;
    7. замена форм слов на исходное слово (например, слова “discount”, “discounts”, “discounted”, “discounting” должны быть заменены на слово “discount”). Такой подход называется stemming;
    8. остальные символы должны быть удалены и заменены на пробелы, т.е. в результате получится текст, состоящий из слов, разделенных пробелами.
18. Загрузите коды слов из словаря **vocab.txt**.
19. Реализуйте функцию замены слов в тексте письма после предобработки на их соответствующие коды.
20. Реализуйте функцию преобразования текста письма в вектор признаков (в таком же формате как в файлах **spamTrain.mat** и **spamTest.mat**).
21. Проверьте работу классификатора на письмах из файлов **emailSample1.txt, emailSample2.txt, spamSample1.txt** и **spamSample2.txt**.
22. Также можете проверить его работу на собственных примерах.
23. Создайте свой набор данных из оригинального корпуса текстов - <http://spamassassin.apache.org/old/publiccorpus/>.
24. Постройте собственный словарь.
25. Как изменилось качество классификации? Почему?
26. Ответы на вопросы представьте в виде отчета.

# Ход выполнения

**Задания 1-2**: Загрузка данных из файла ex5data1.mat и построение графика для загруженного набора данных: по осям - переменные X1, X2, а точки, принадлежащие различным классам должны быть обозначены различными маркерами.

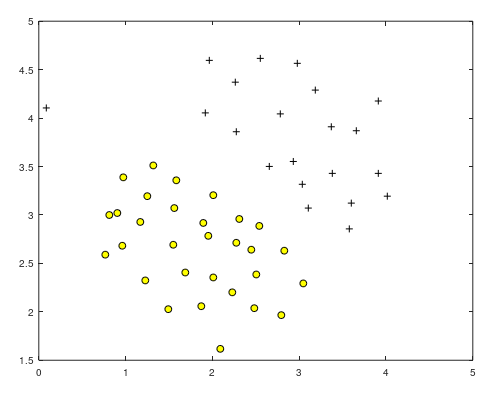


Рис 1. – точечный график тренировочного набора (ex5data1.mat)

**Задание 3**: Обучение классификатора с помощью библиотечной реализации SVM с линейным ядром на данном наборе.

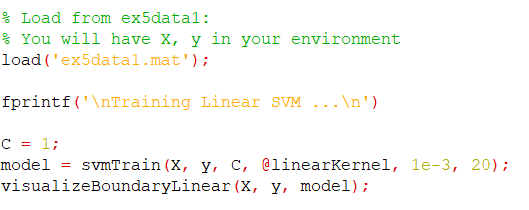


Рис 2. – Код использования библиотечной реализации SVM

**Задание 4**: Построение разделяющей прямой для классификаторов с различными параметрами C = 1, C = 100 (совместно с графиком из пункта 2).

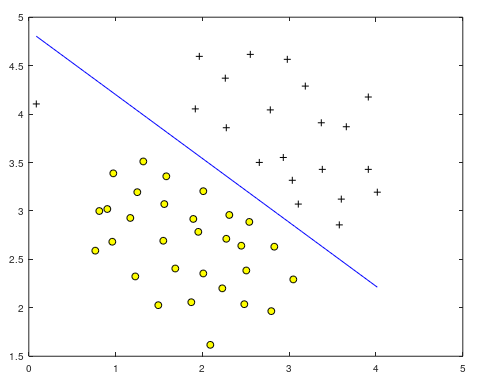


Рис 3. – Разделяющая прямая SVM при параметре C = 1

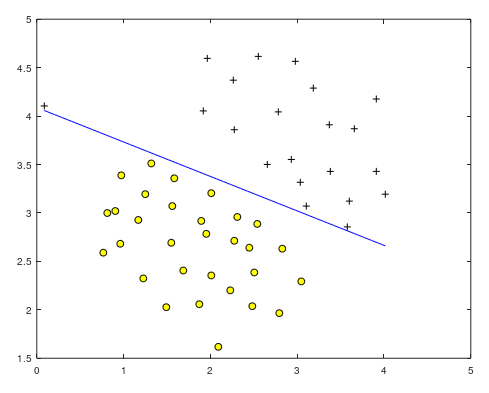
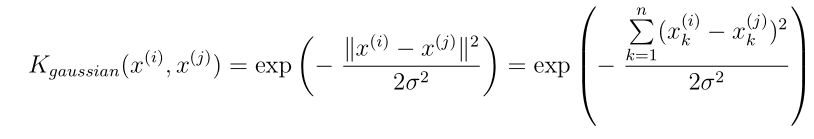


Рис 4. – Разделяющая прямая SVM при параметре C = 100

**Задание 5**: Реализация функции вычисления Гауссового ядра для алгоритма SVM.

Формула для вычисления Гауссового ядра:



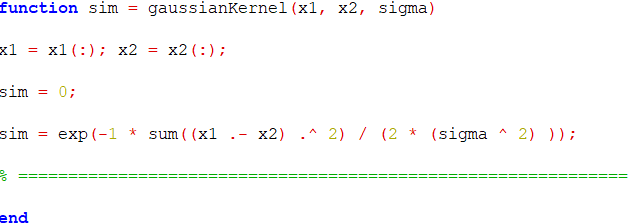


Рис 5. – Код функции вычисления Гауссового ядра

**Задание 6**: Загрузка данных ex5data2.mat из файла.

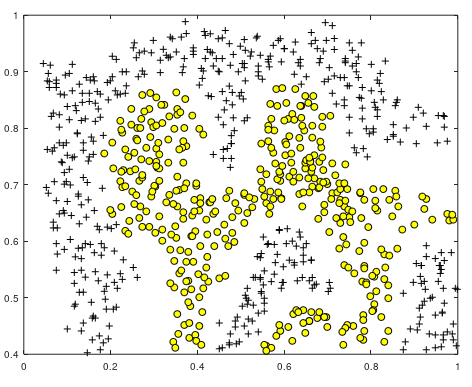


Рис 6. – точечный график тренировочного набора (ex5data2.mat)

**Задания 7-9**: Обработка данных с помощью функции Гауссового ядра. Обучение классификатора SVM. Визуализация данных вместе с разделяющей кривой (аналогично пункту 4)

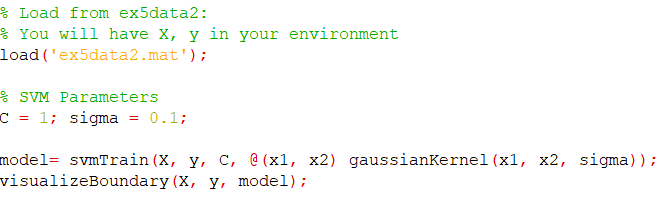


Рис 7. – Код обучения классификатора SVM с помощью функции Гауссового ядра

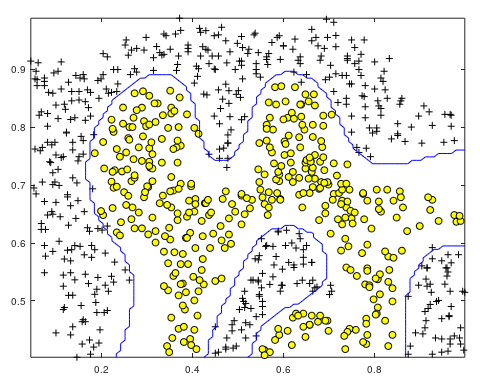


Рис 8. – Визуализация данных вместе с разделяющей кривой SVM для тренировочного набора (ex5data2.txt)

**Задание 10**: Загрузка данные ex5data3.mat из файла.

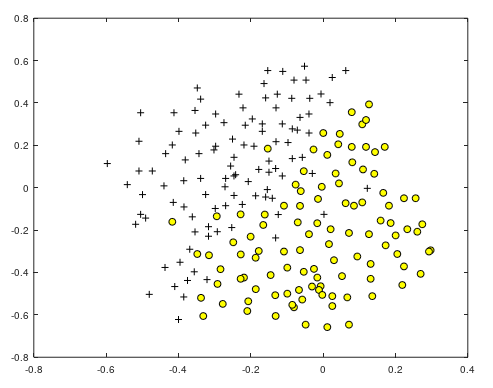


Рис 9. – точечный график тренировочного набора (ex5data3.mat)

**Задание 11**: Вычисление параметров классификатора SVM на обучающей выборке, а также подбор параметров C и σ2 на валидационной выборке.

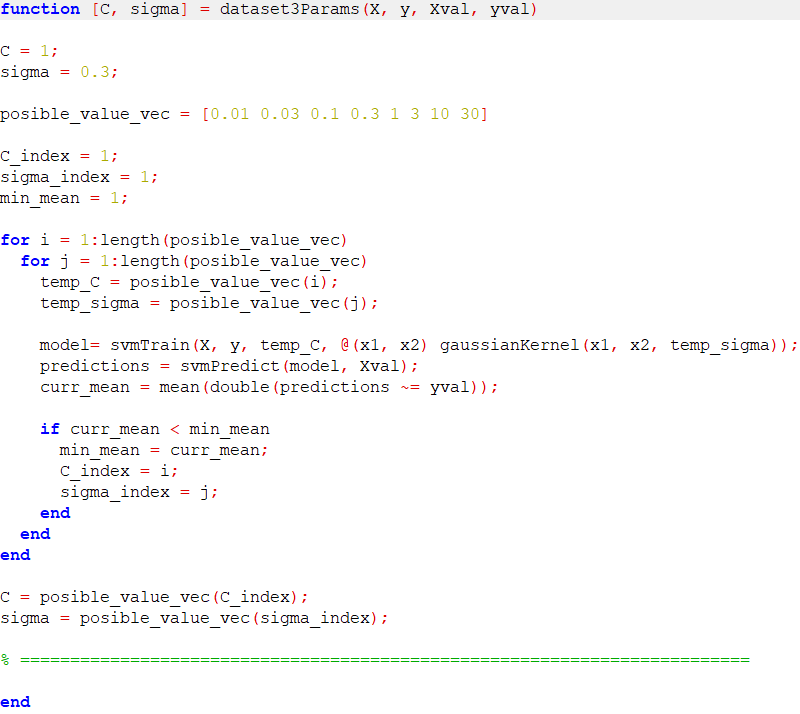


Рис 10. – Код функции подбора параметров C и σ2 на валидационной выборке

**Задание 12**: Визуализация данных вместе с разделяющей кривой (аналогично пункту 4).

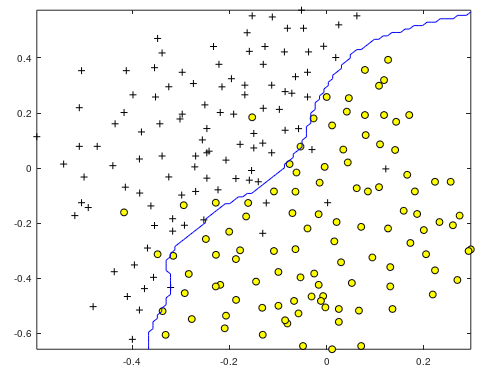


Рис 11. – Визуализация данных вместе с разделяющей кривой SVM для тренировочного набора (ex5data3.txt)

**Задание 17**: Реализация функции предобработки текста письма.

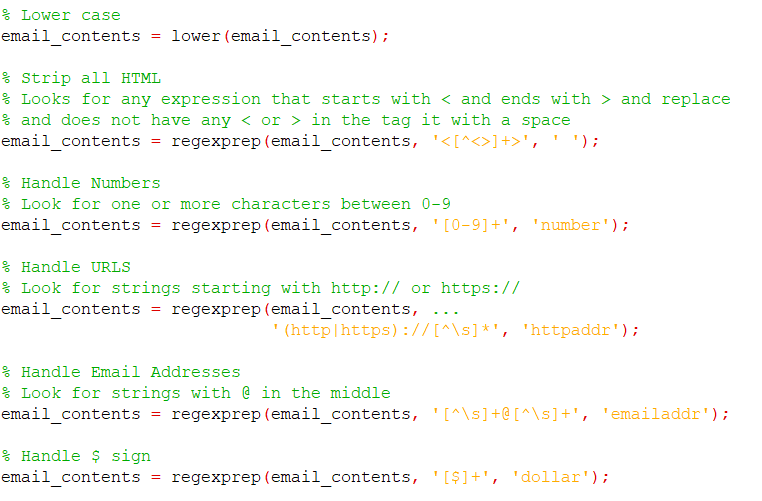


Рис 12. – Код функции предобработки текста письма

**Задание 18**: Загрузка кодов слов из словаря vocab.txt.

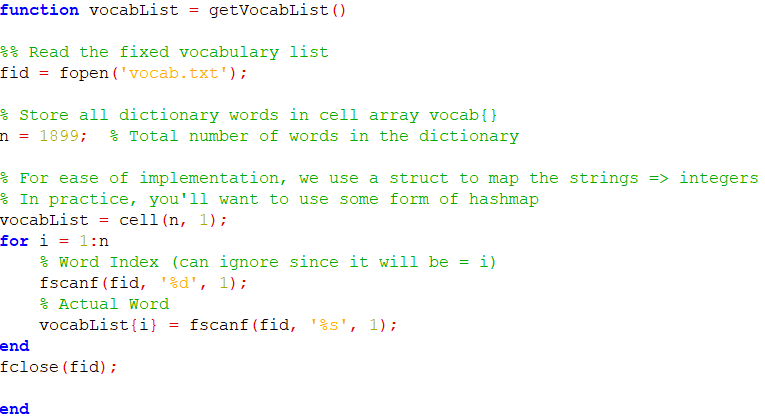


Рис 13. – Код загрузки кодов слов из словаря vocab.txt

**Задание 19**: Реализация функции замены слов в тексте письма после предобработки на их соответствующие коды.

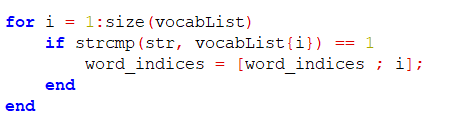


Рис 14. – Код замены слов в тексте письма на их соответствующие коды

**Задания 20**: Реализация функции преобразования текста письма в вектор признаков.

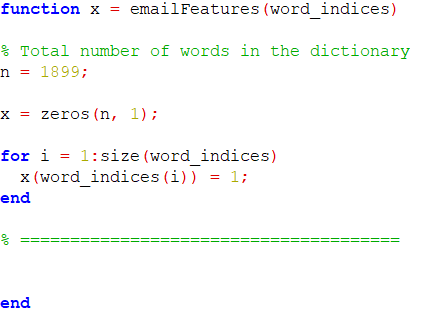


Рис 15. – Код функции преобразования текста письма в вектор признаков

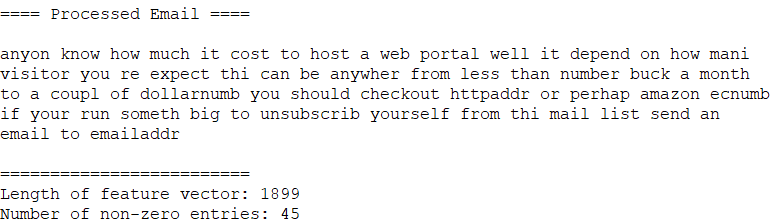


Рис 16. – Пример преобразования текста письма в вектор признаков

**Задание 21**: Проверка работы классификатора на письмах из файлов emailSample1.txt, emailSample2.txt, spamSample1.txt и spamSample2.txt.

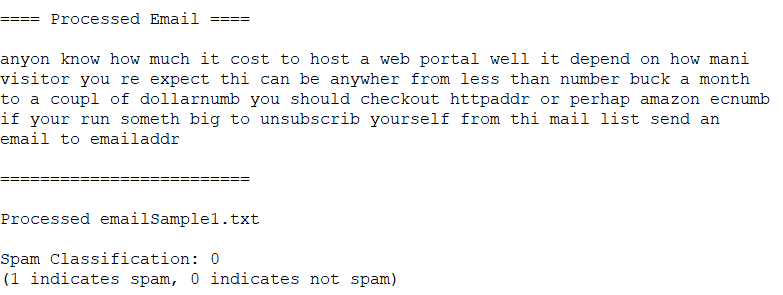


Рис 17. – Проверка работы классификатора на emailSample1.txt

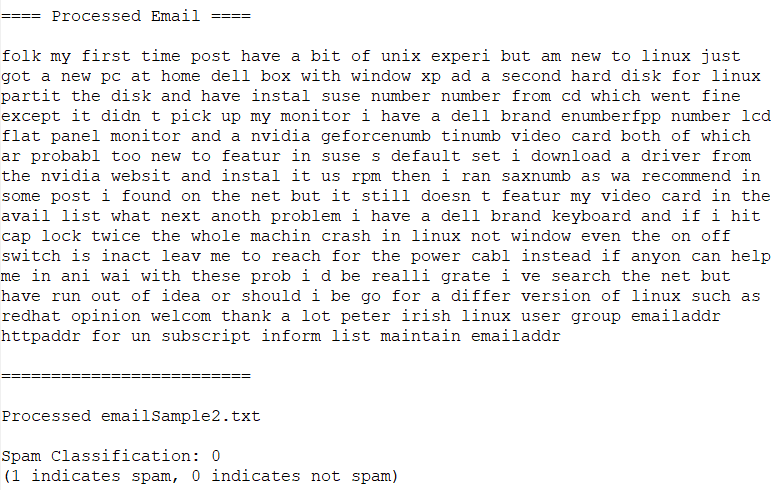


Рис 18. – Проверка работы классификатора на emailSample2.txt

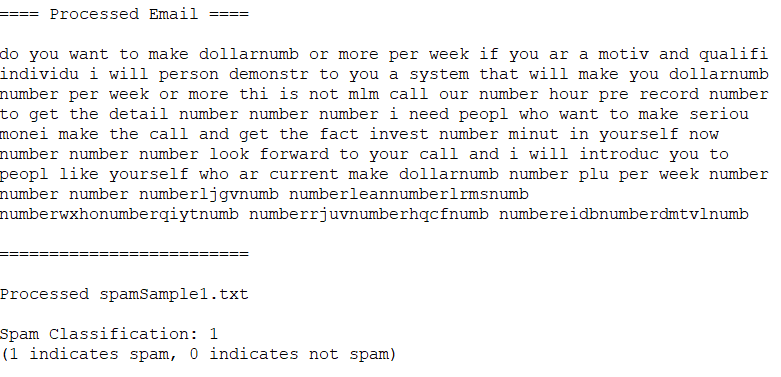


Рис 19. – Проверка работы классификатора на spamSample1.txt

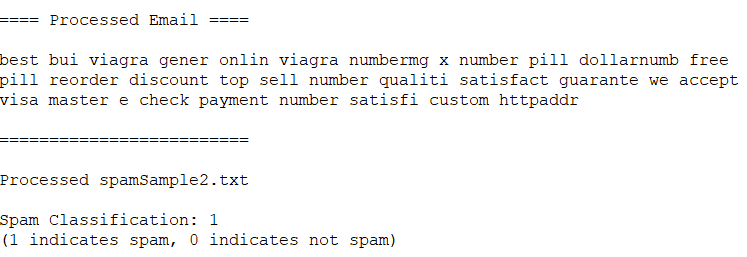


Рис 20. – Проверка работы классификатора на spamSample2.txt

# Выводы

Мы научились использовать метод опорных векторов для решения задачи классификации. В процессе мы научились реализовывать функцию вычисления Гауссового ядра для этого алгоритма.

Раннее для решения задач классификации мы раньше использовали логическую регрессию. Условия использования одного из двух методов представлены ниже:

Пусть *n –* количество признаков, *m* – размер тренировочного набора

Тогда, если *n* большое в сравнении с *m* – необходимо использовать логическую регрессию, либо метод опорных векторов (совместно с линейным ядром).

Если *n* мало (1-1000), а *m* среднее (10-10,000) – необходимо использовать метод опорных векторов совместно с Гауссовым ядром.

Если *n* мало (1-1000), а *m* большое (50,000+) – то необходимо создать/добавить признаков с дальнейшим использованием логической регрессии, либо метода опорных векторов (совместно с линейным ядром).