Министерство образования Республики Беларусь

Учреждение образования

БелорусскиЙ государственный университет

информатики и радиоэлектроники

Факультет компьютерных систем и сетей

Кафедра информатики

ОТЧЕТ

по лабораторной работе №5

**Метод опорных векторов**

|  |  |  |
| --- | --- | --- |
| Студент |  | И.Ю. Фалько |
| Преподаватель |  | М. В. Стержанов |

Минск 2019

СОДЕРЖАНИЕ

[Условие лабораторной работы 3](#_Toc25744885)

[Ход выполнения 3](#_Toc25744886)

[Заключение 12](#_Toc25744887)

# Условие лабораторной работы

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.

# Ход выполнения

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.

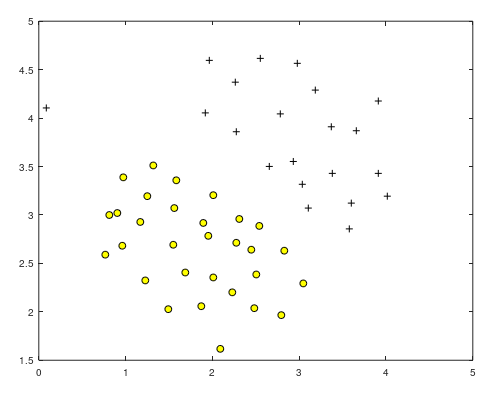
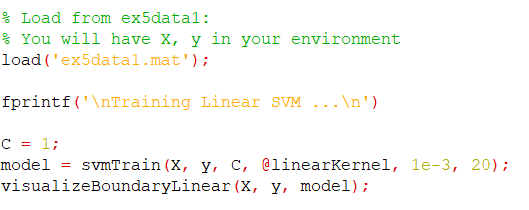


Рисунок 1. – точечный график тренировочного набора

1. Обучите классификатор с помощью библиотечной реализации SVM с линейным ядром на данном наборе.



1. Постройте разделяющую прямую для классификаторов с различными параметрами C = 1, C = 100 (совместно с графиком из пункта

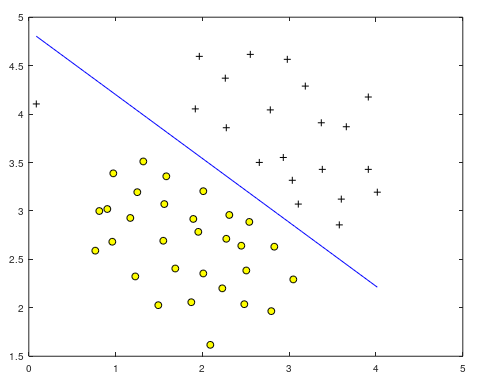


Рисунок 3. – Разделяющая прямая SVM при параметре C = 1

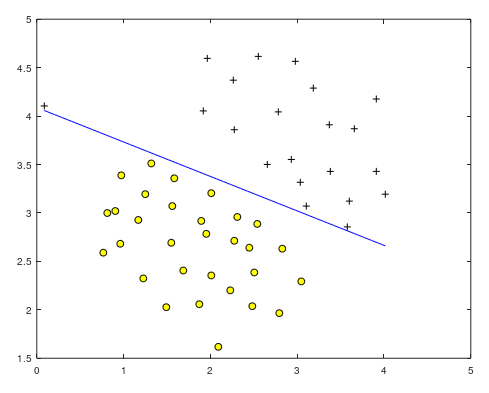
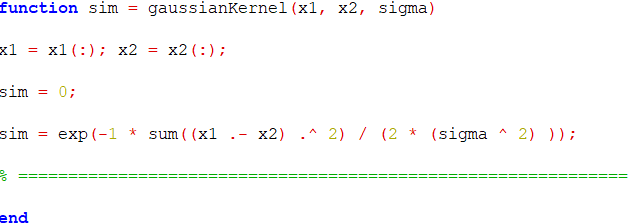


Рисунок 4. – Разделяющая прямая SVM при параметре C = 100

1. Реализуйте функцию вычисления Гауссового ядра для алгоритма SVM.



1. Загрузите данные **ex5data2.mat** из файла.

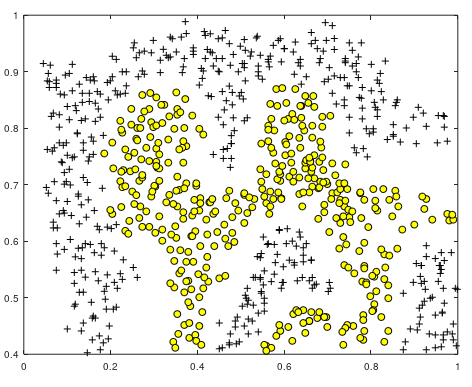
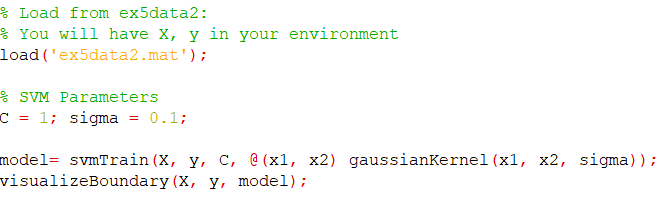


Рисунок 6. – точечный график тренировочного набора

* 1. Обработайте данные с помощью функции Гауссового ядра.

Обучите классификатор SVM.

Визуализируйте данные вместе с разделяющей кривой (аналогично пункту 4).



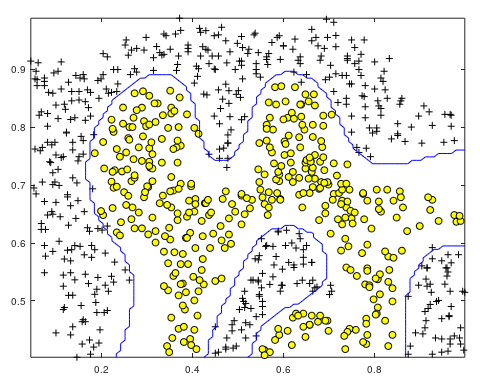


Рисунок 7. – Визуализация данных вместе с разделяющей кривой SVM для тренировочного набора

1. Загрузите данные **ex5data3.mat** из файла.

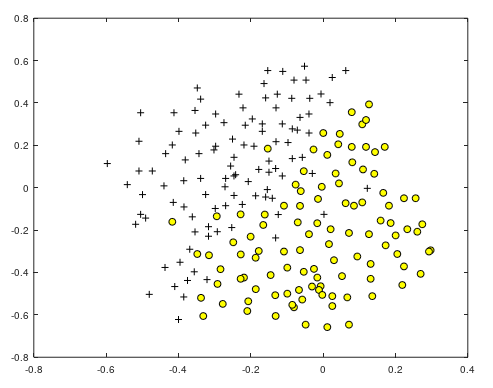
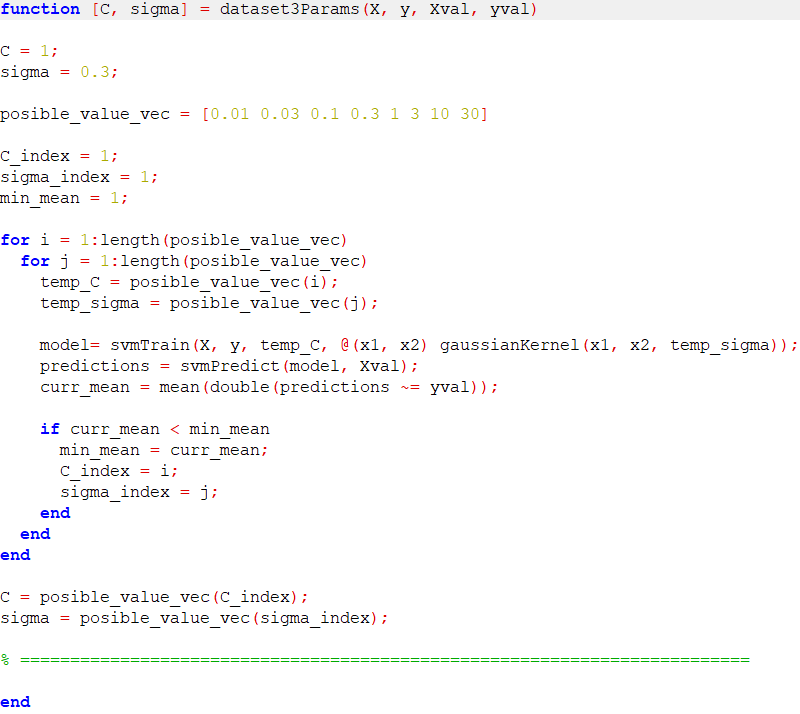


Рисунок 8. – точечный график тренировочного набора

1. Вычислите параметры классификатора SVM на обучающей выборке, а также подберите параметры C и σ2 на валидационной выборке.



1. Визуализируйте данные вместе с разделяющей кривой (аналогично пункту 4).

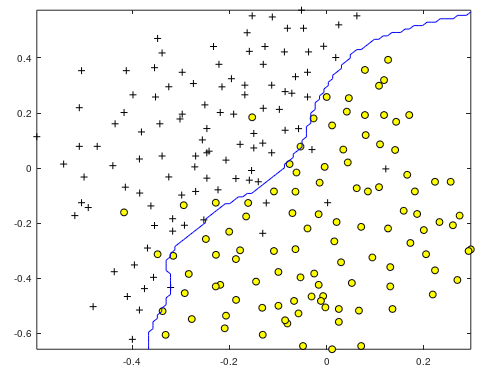


Рисунок 9. – Визуализация данных вместе с разделяющей кривой SVM для тренировочного набора

13-17 Загрузите данные **spamTrain.mat** из файла.

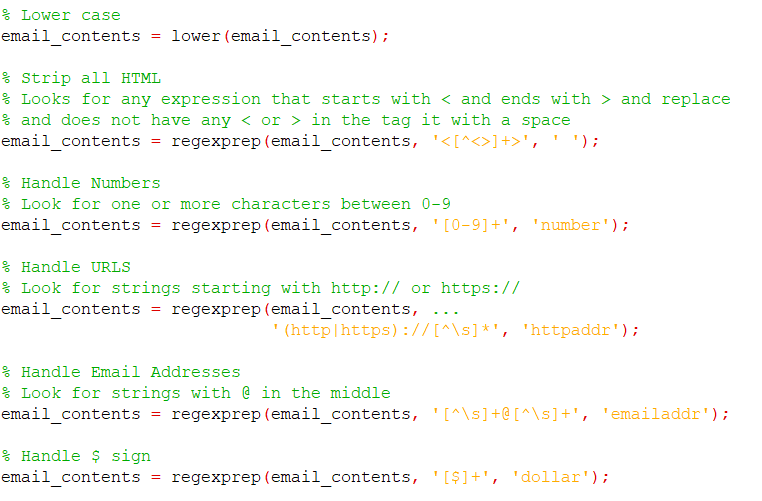
Обучите классификатор SVM.

Загрузите данные **spamTest.mat** из файла.

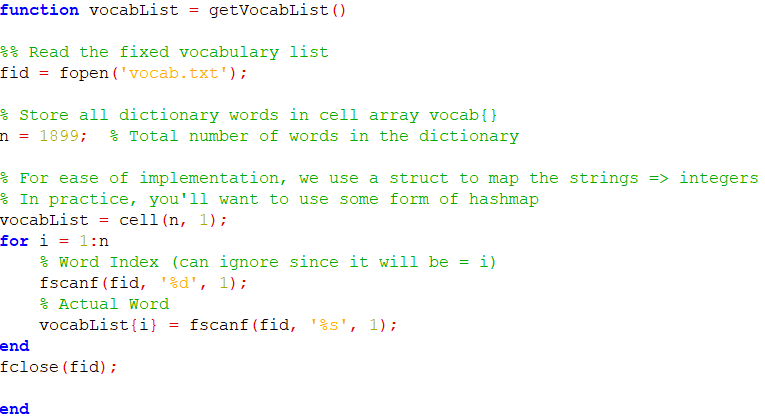
Подберите параметры C и σ2.

Реализуйте функцию предобработки текста письма, включающую в себя:

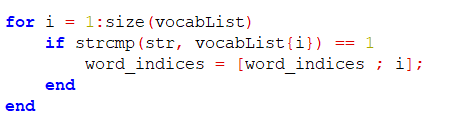
* 1. перевод в нижний регистр;
  2. удаление HTML тэгов;
  3. замена URL на одно слово (например, “httpaddr”);
  4. замена email-адресов на одно слово (например, “emailaddr”);
  5. замена чисел на одно слово (например, “number”);
  6. замена знаков доллара ($) на слово “dollar”;
  7. замена форм слов 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.



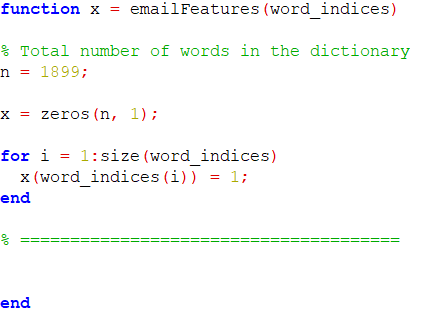
1. Загрузите коды слов из словаря **vocab.txt**.



1. Реализуйте функцию замены слов в тексте письма после предобработки на их соответствующие коды.



1. Реализуйте функцию преобразования текста письма в вектор признаков (в таком же формате как в файлах **spamTrain.mat** и **spamTest.mat**).



1. Проверьте работу классификатора на письмах из файлов **emailSample1.txt, emailSample2.txt, spamSample1.txt** и **spamSample2.txt**.

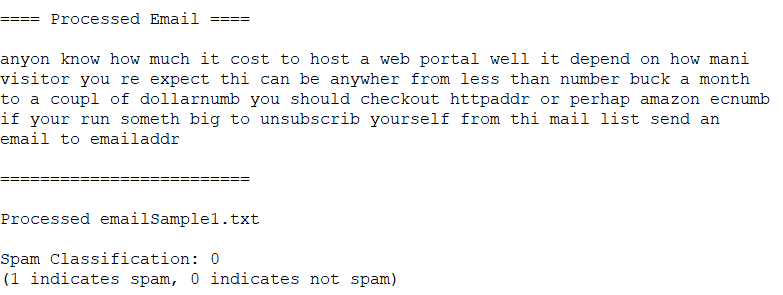


Рисунок 10. – Проверка работы классификатора на emailSample1.txt

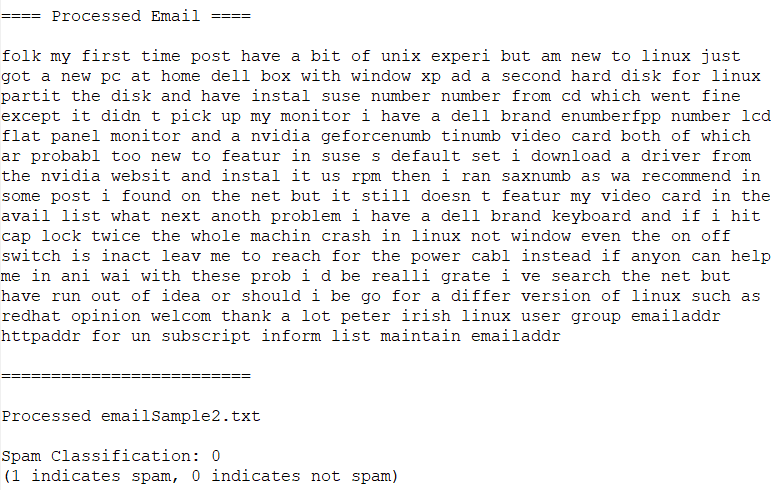


Рисунок 11. – Проверка работы классификатора на emailSample2.txt

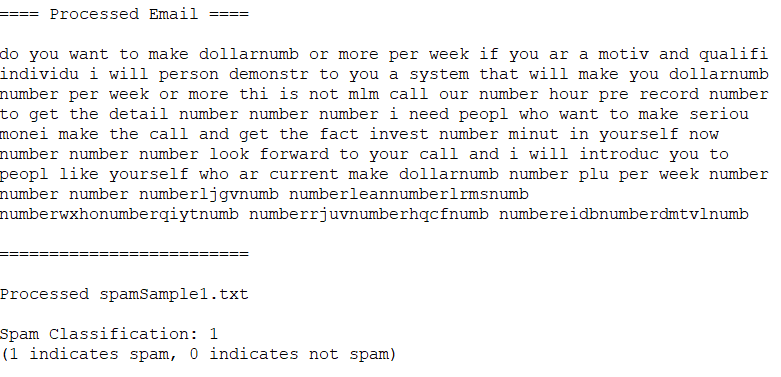


Рисунок 12. – Проверка работы классификатора на spamSample1.txt

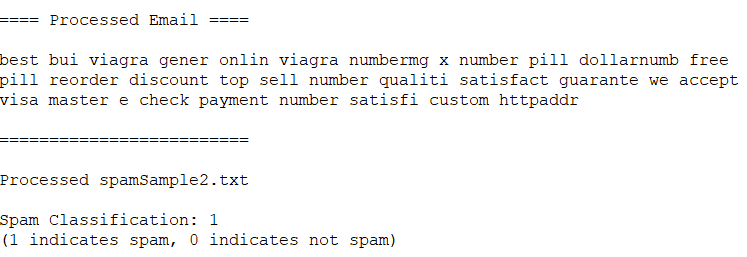


Рисунок 23. – Проверка работы классификатора на spamSample2.txt

# ЗАКЛЮЧЕНИЕ

В данной лабораторной работе были освоены навыки метод опорных векторов для решения задачи классификации, реализация функции вычисления Гауссового ядра для этого алгоритма.

Раннее для решения задач классификации использовалась логическая регрессия.