Министерство образования и науки Российской Федерации
Санкт-Петербургский государственный технический университет
Институт прикладной математики и механики
Кафедра «Телематика»

ЛАБОРАТОРНАЯ РАБОТА

ПО ТЕМЕ

«Метод поиска ближайшего соседа»

по направлению 02.04.01.02 «Организация и управление суперкомпьютерными системами»

Выполнил:
Студент гр. 13643.1 Титов А.И.
Проверил: Уткин Л.В.

Санкт-Петербург
2019

Оглавление

Постановка задачи

Требуется выполнить следующие задачи:

1. Исследовать, как объем обучающей выборки и количество тестовых данных, влияет на точность классификации или на вероятность ошибочной классификации в примере крестики-нолики и примере о спаме e-mail сообщений.

2. Построить классификатор для обучающего множества Glass, данные которого характеризуются 10-ю признаками:

    1. Id number: 1 to 214;
    2. RI: показатель преломления;
    3. Na: сода (процент содержания в соответствующем оксиде);
    4. Mg;
    5. Al;
    6. Si;
    7. K;
    8. Ca;
    9. Ba;
    10. Fe.

    Классы характеризуют тип стекла:

    1. окна зданий, плавильная обработка
    2. окна зданий, не плавильная обработка
    3. автомобильные окна, плавильная обработка
    4. автомобильные окна, не плавильная обработка (нет в базе)
    5. контейнеры
    6. посуда
    7. фары

    Перед построением классификатора необходимо удалить первый признак Id number. Это выполняется командой glass <- glass[,-1]. Построить графики зависимости ошибки классификации от значения k и от типа ядра. Исследовать, как тип метрики расстояния (параметр distance) влияет на точность классификации. Определить, к какому

типу стекла относится экземпляр с характеристиками: 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. Определить, какой из признаков оказывает наименьшее влияние на определение класса путем последовательного исключения каждого признака.

3. Для построения классификатора использовать заранее сгенерированные обучающие и тестовые выборки. Найти оптимальное значение k, обеспечивающее наименьшую ошибку классификации. Посмотреть, как выглядят данные на графике.

4. Разработать классификатор на основе метода ближайших соседей для данных Титаник (Titanic dataset).

1 Наборы данных «крестики-нолики» и «спам e-mail сообщений»

Для того чтобы исследовать, как объем обучающей выборки влияет на точность классификации были применен метод кросс-валидации. Были рассмотрены размеры обучающей выборки от 10% до 90% исходного набора данных. Для исследуемых наборов данных были построены графики зависимости (рис 1-2). Так как точность вычислений может отличаться от раза к разу - было подсчитано среднее значение точности из 5 переобучений алгоритма.

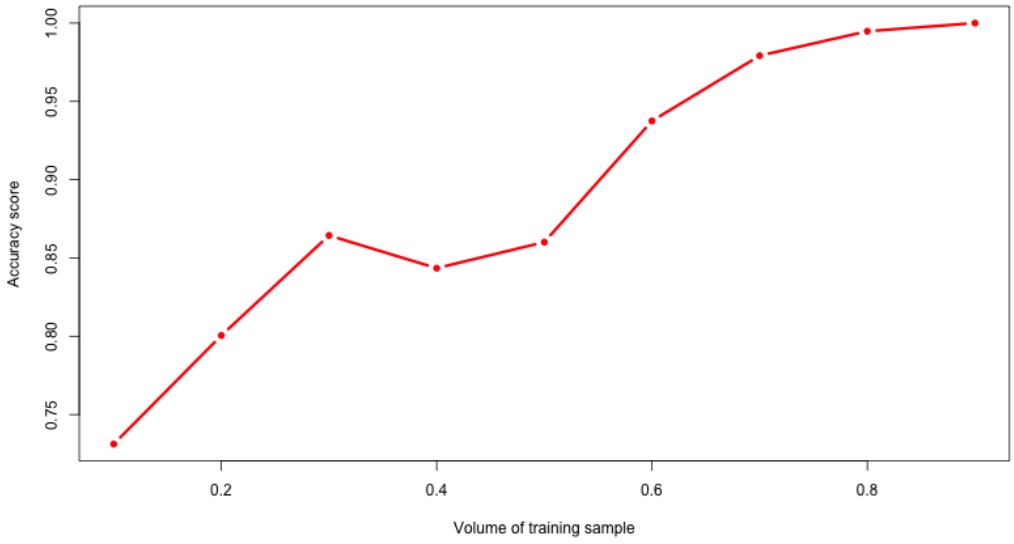

Рис. 1. График зависимости для набора «Крестики-нолики»

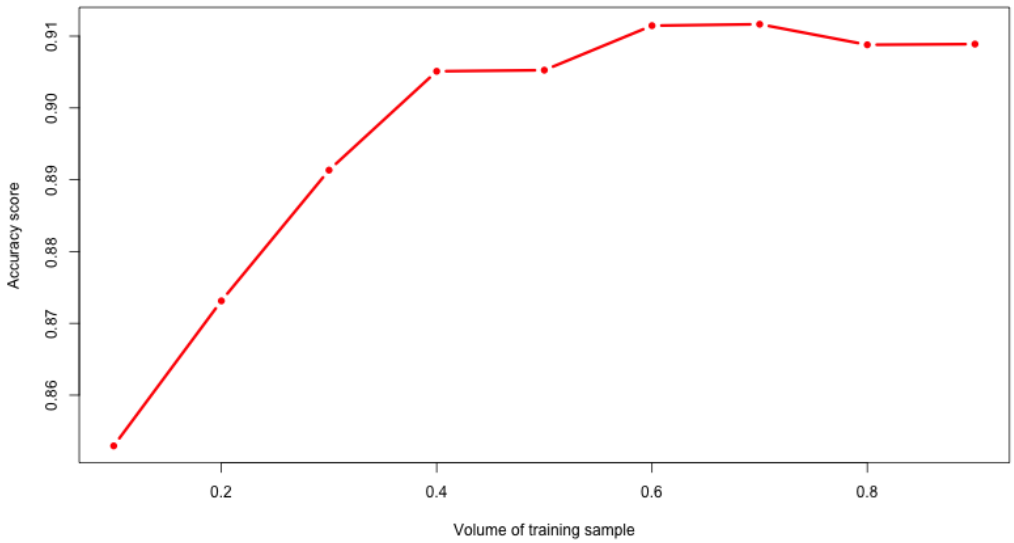

Рис. 2. График зависимости для набора «Спам email сообщений»

2 Набор данных «Glass»

Для того, чтобы исследовать работу алгоритма с предложенным набором данных из него был удален признак «Id number» (он не несет информационной нагрузки). Набор данных был разбит на обучающую и тестирующую выборки в соотношении 9:1.

Обученный алгоритм классифицировал данный в условии задания экземпляр в класс «5».

Также были проведены исследования того, как значение параметра расстояния Минковского и количество рассматриваемых соседей (k) влияет на точность классификации при использовании разных функций ядра (рис. 3).

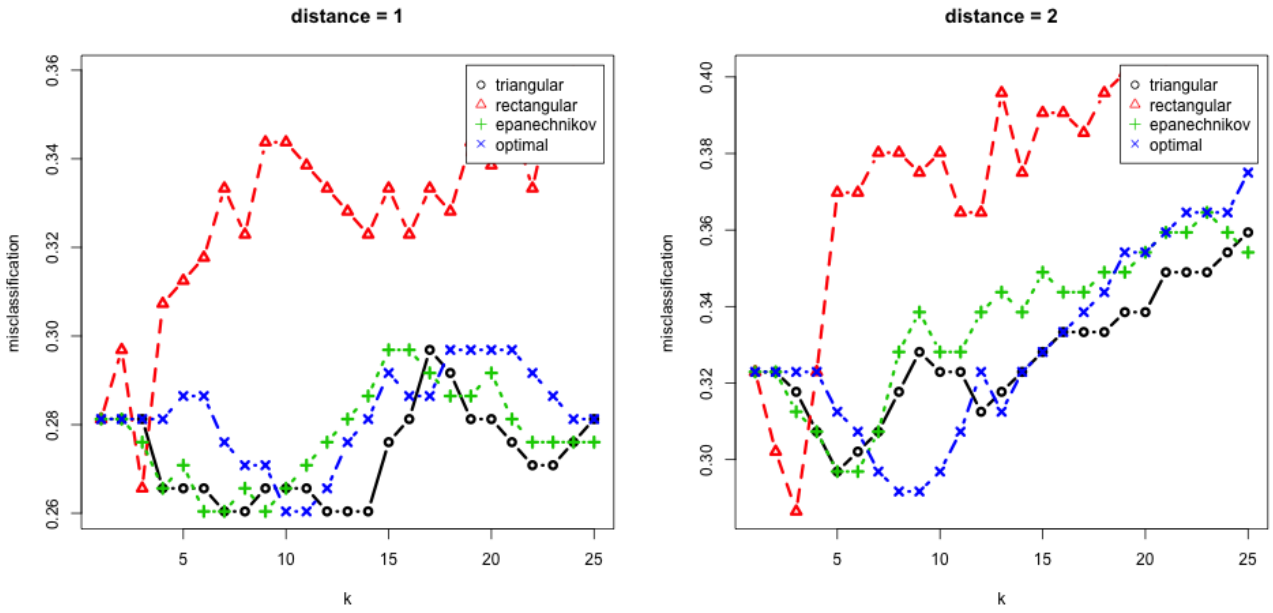

Рис. 3. График зависимости для набора «Glass»

Были выяснены оптимальные параметры классификатора для данного набора данных.

- Для параметра расстояния Минковского = 1:
  - Оптимальная функция ядра - треугольная;
  - Оптимальное количество соседей - 7.
- Для параметра расстояния Минковского = 1:
  - Оптимальная функция ядра - равномерная;
  - Оптимальное количество соседей - 3.

Был проведен анализ того, какой из признаков оказывает наименьшее влияние на определение класса. Для этого был построен график (рис. 4).

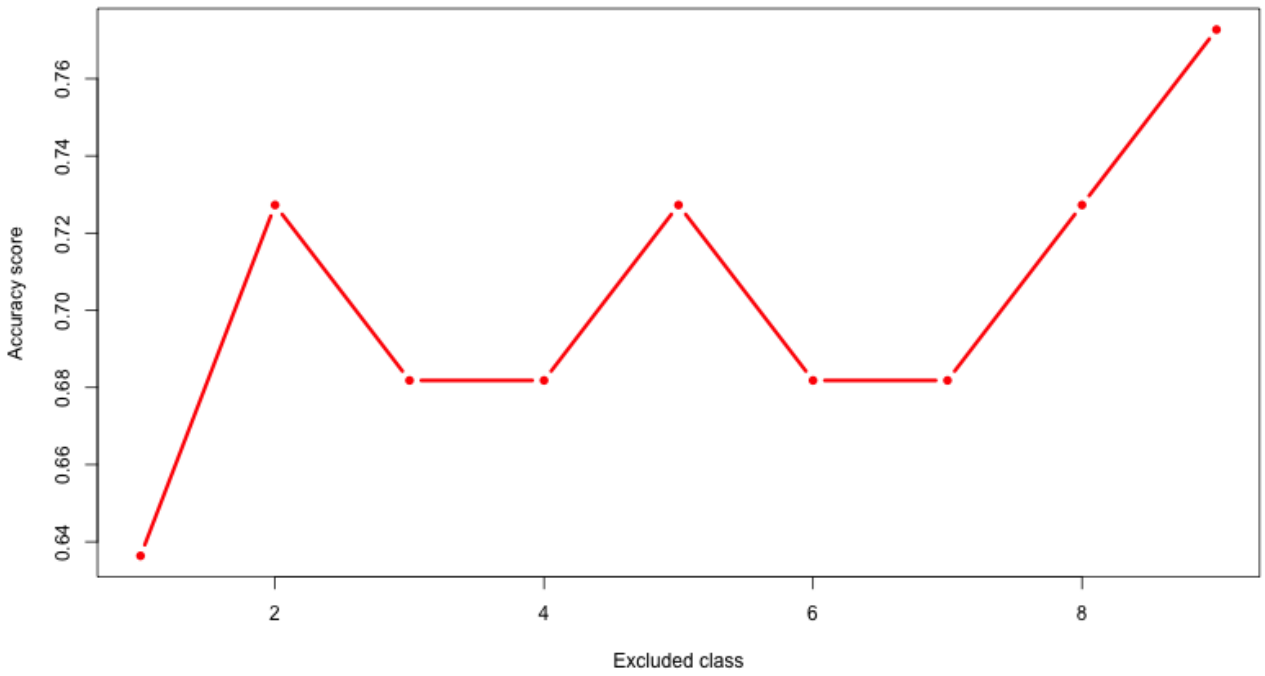

Рис. 4. График точности при исключении признаков для набора «Glass»

Говоря напрямую - сложно сказать, какой из признаков оказывает наименьшее влияние, потому что, не смотря на то, что замеры приведены по среднему из 5ти переобучений, возвращаемые значения имеют весьма случайный характер. Однако, чаще всего наименее влияющим оказывался 9ый признак.

3 Набор данных «svmdata4»

Для того, чтобы определить оптимальные значение k и тип функции ядра, был применен метод кросс-валидации. В результате были выявлены следующие значения параметров:

- Функция ядра - «optimal»;
- Количество соседей - 8.

Используя такие значения параметров была получена минимальная ошибка 0.03.

Также были построены графики, визуализирующие исходные данные и результат процесса кросс-валидации (рис. 5).

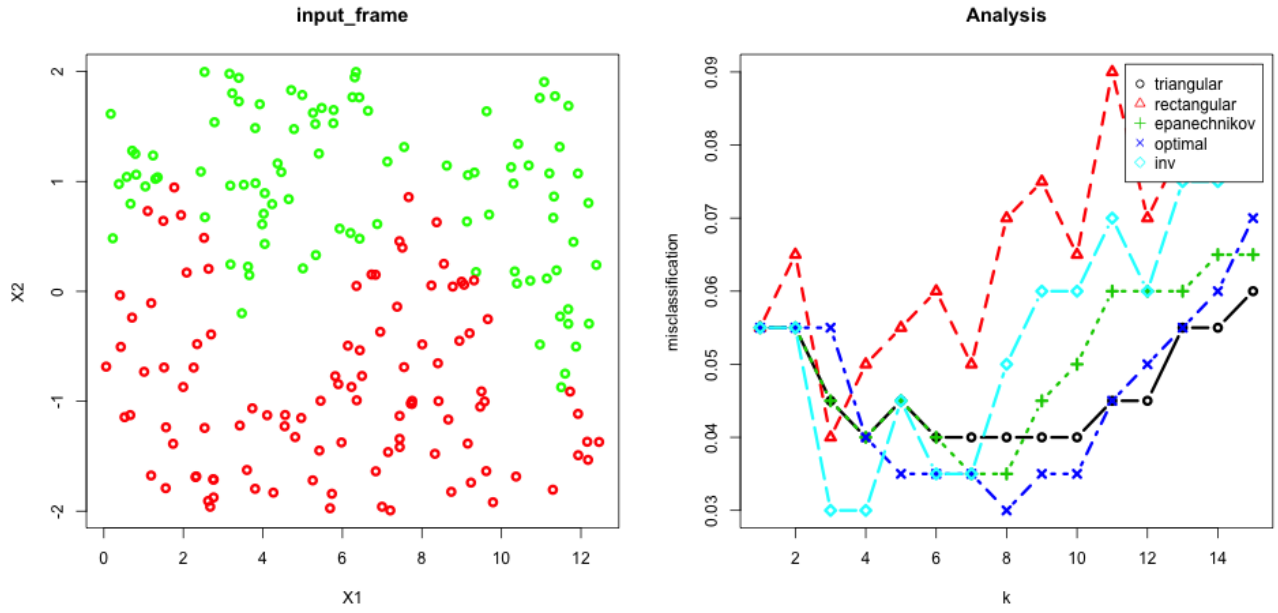

Рис. 5. Исходные данные и результаты кросс-валидации для набора «svmdata4»

4 Набор данных «Titanic»

Был использован метод кросс-валидации для оценки оценки работы данного алгоритма с набором данных «Titanic». На изображении ниже приведены результаты процесса кросс-валидации (рис. 6).

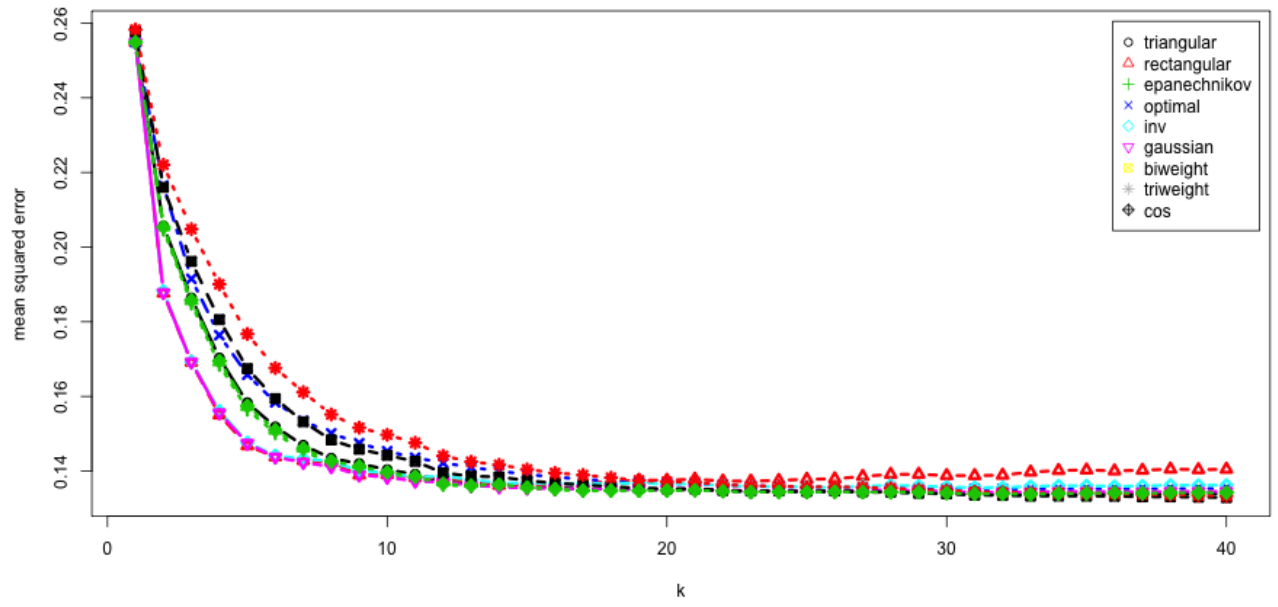

Рис. 6. Результаты кросс-валидации для набора «svmdata4»