Санкт-Петербургский Политехнический Университет Петра Великого

Институт прикладной математики и механики

Кафедра Телематики при ЦНИИ РТК

Лабораторная работа

по теме

«Метод поиска ближайшего соседа»

Преподаватель Уткин Л.В.

Студент гр.43607/2 Лисенкова А.А

Санкт-Петербург

2018 г.

# Постановка задачи

В ходе выполнения лабораторной работы необходимо 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.

1. Для построения классификатора использовать заранее сгенерированные обучающие и тестовые выборки. Найти оптимальное значение k, обеспечивающее наименьшую ошибку классификации. Посмотреть, как выглядят данные на графике.
2. Разработать классификатор на основе метода ближайших соседей для данных **Титаник (Titanic dataset).**

# Ход работы

**1 задание**

Для того чтобы исследовать, как объём обучающей выборки и количество тестовых данных влияет на точность классификации выполнены следующие шаги:

1. Загрузка исходных данных из файла в структуру A\_raw.
2. Определение последовательности ‘x’: от 0.5 до 0.9 с шагом 0.05, которая отвечает за процентное соотношение обучающей и тестирующей выборок.
3. Для каждого разбиения обучающей и тестирующей выборки вызов функции calculate, в которой происходит основная часть вычислений:

* рандомизация исходных данных;
* использование весового метода ближайших соседей;
* оценка полученной модели;
* построение таблицы для сравнения прогнозируемых значений с исходными;
* вычисление значения вероятности ошибочной классификации.

1. Построение графика зависимости значения вероятности ошибочной классификации от объема обучающей выборки.

Вычисления производились на двух выборках: крестики-нолики и спам e-mail сообщений. Соответственно, в результате получилось 2 графика (см. Рис.1-2). Следует также отметить, что для достижения более правдоподобной оценки, действия из п.3 повторялись 5 раз, после чего находилось среднее арифметическое значения вероятности ошибочной классификации.

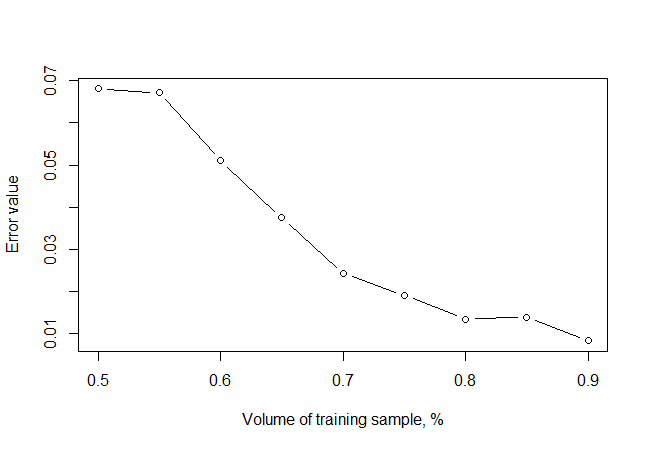


Рис. 1. График зависимости для примера «Крестики-нолики»

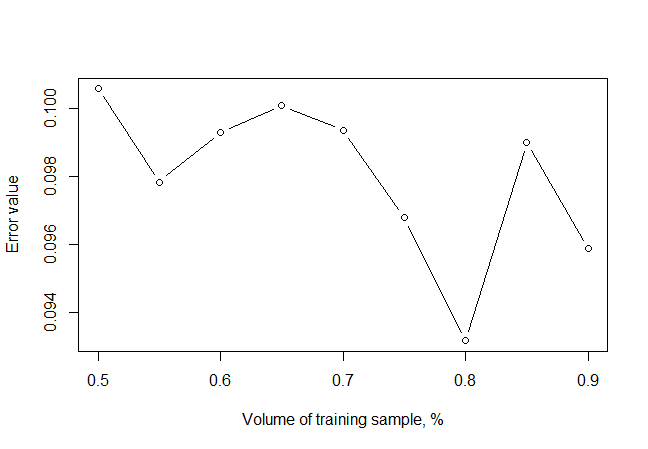


Рис. 2. График зависимости для примера «Спам e-mail сообщений»

**2 задание**

Для построения классификатора были выполнены следующие действия:

1. Загрузка данных в соответствующую структуру.
2. Удаление первого признака из исходного набора данных, т.к. он не несет никакой информационной нагрузки.
3. Разбиение выборки на обучающую и тестирующую в соотношении 9:1.
4. Определение к какому типу стекла относится экземпляр с данными характеристиками.
5. Использование метода кросс-валидации с различными значениями distance.
6. Построение графиков для данных, полученных из п.4.
7. Определение, какой из признаков оказывает наименьшее влияние на определение класса.

Согласно п.4, было выяснено, что заданный пример принадлежит к 5 классу.

После проведения исследования, как тип метрики расстояния влияет на точность классификации, построены графики, приведённые на Рис.3. Также, были получены следующие рекомендации для выбора оптимального параметра ядра и k:

|  |  |  |  |
| --- | --- | --- | --- |
|  | | distance | |
| 1 | 2 |
| параметры | ядро | triangular | rectangular |
| K | 22 | 3 |

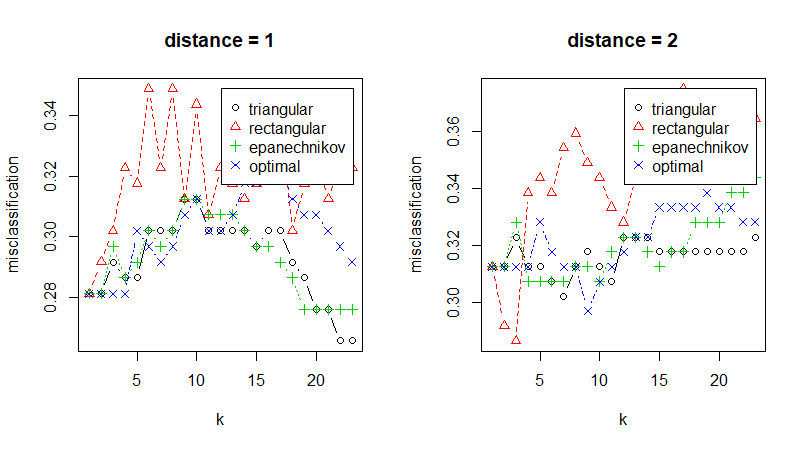


Рис. 3. Графики зависимости ошибки от значений k для двух значений distance

Для выявления, какой из признаков оказывает наименьшее влияние на определение класса, построен график, изображённый на Рис.4. Анализируя его, можно сделать вывод о том, что исключение 2 признака приведет к наибольшему значению ошибочной классификации.

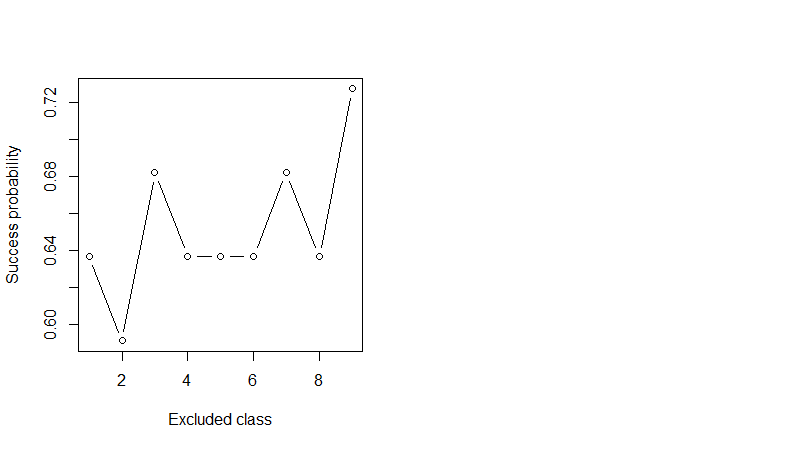


Рис.4 График зависимости вероятности успешной классификации от исключённого признака

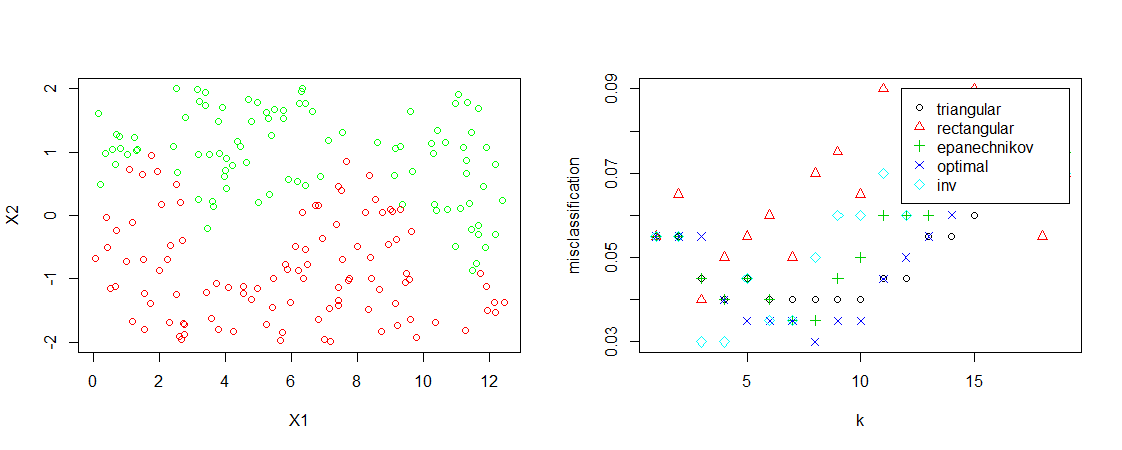
**3 задание**

Для нахождения оптимального значения k и вида ядра, которые обеспечивают наименьшую ошибку классификации, используется метод кросс-валидации. После применения его к обучающей выборке, получаются следующие значения:

Best kernel: optimal

Best k: 8

Кроме этого, построены графики, визуализирующие исходные данные (2 класса объектов) и найденный список значений, содержащий оптимальные параметры ядра и k (см. Рис.3).

Рис. 5 Исходные и результирующие данные

**4 задание**

Для выполнения задания был выбран метод кросс-валидации k ближайших соседей. Это техника валидации модели для проверки того, насколько успешно применяемый в модели статистический анализ способен работать на независимом наборе данных. Один цикл кросс-валидации включает разбиение набора данных на части, затем построение модели на одной части (называемой тренировочным набором), и валидация модели на другой части (тестовом наборе).

По результатам проведённой кросс-валидации были получены наилучшие значения для выбора ядра и величины k:

Best kernel: biweight.

Best k: 29.