Домашнее Задание №5. Петраков Сергей э-301.

В данной работе мне предстоит провести классификацию качества вин в зависимости от их физико-химических свойств. Для решения задачи я буду использовать такой алгоритм Машинного обучения, как метод опорных векторов.

Часть 1. Анализ данных.

Перед непосредственной реализацией задачи классификации рассчитаем ключевые статистические показатели: среднее и стандартное отклонение признаков, результаты расчёта приведены в таблице ниже.

Среди доступных данных имеются следующие показатели: уровень содержания алкоголя (alcohol), постоянная кислотность (fixed acidity), летучая кислотность (volatile acidity), содержание лимонной кислоты (citric acid), остаточный сахар (residual sugar), хлориды (chlorides), свободный диоксид серы (free sulfur dioxide), общий диоксид серы (total sulfur dioxide), плотность (density), единица pH (pH), сульфаты (sulphates), уровень алкоголя (alcohol), качество (quality).

Characteristic	Mean	Standard deviation
alcohol	10.423	1.0657
chlorides	0.087	0.0471
citric acid	0.271	0.1948
density	0.997	0.0019
fixed acidity	8.320	1.7411
free sulfur dioxide	15.875	10.4602
pH	3.311	0.1544
quality	5.636	0.8076
residual sugar	2.539	1.4099
sulphates	0.658	0.1695
total sulfur dioxide	46.468	32.8953
volatile acidity	0.528	0.1791

Рис. 1. Таблица выборочных средних и стандартных отклонений для каждого из анализируемых показателей.

Дополнительно проведу расчёт корреляций между имеющимися показателями, это может помочь при предварительном анализе важности разных признаков на качество вина. Ниже я привёл тепловую диаграмму, которая иллюстрирует направление взаимосвязи между данными.

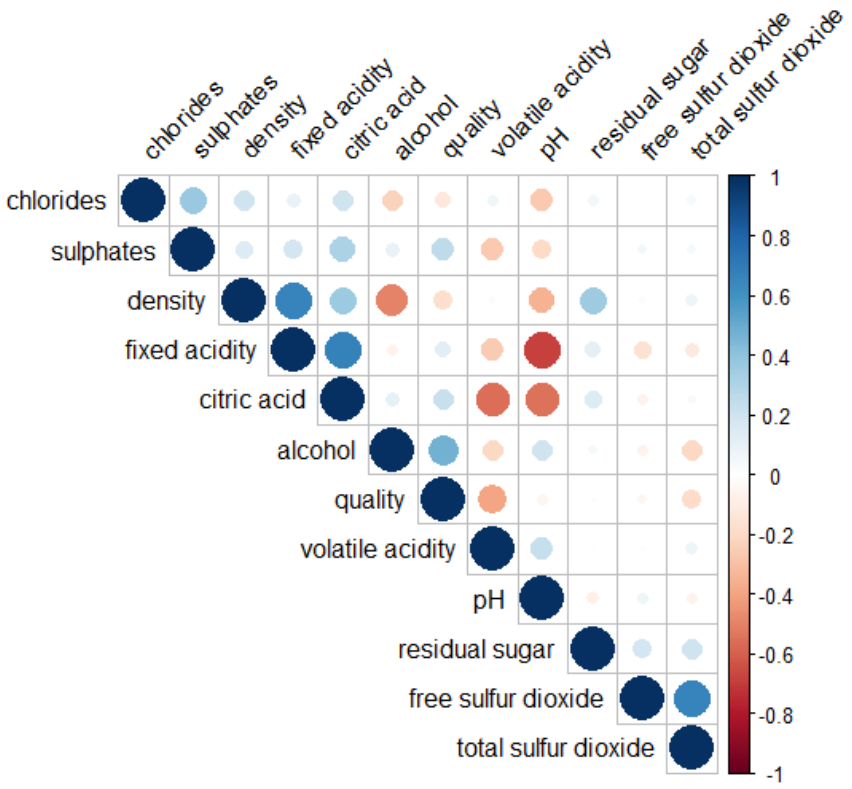

Рис. 2. Тепловая диаграмма корреляций между показателями.

Самые явные взаимосвязи:

1) Между постоянной кислотностью и уровнем pH заметная отрицательная связь.
2) Наблюдаются отрицательные взаимосвязи между уровнем содержания лимонной кислоты и показателями pH и остаточным сахаром.
3) также можно наблюдать отрицательную взаимосвязь между плотностью и уровнем алкоголя.
4) Что касается заметных положительных взаимосвязей, то можно сказать про связи плотности и постоянной кислотности, постоянной плотности и содержанием лимонной кислотной кислоты, свободным диоксидом серы и общим диоксидом серы.
5) Отметим, что относительно нашего целевого показателя – качества можно сказать, что сколько-нибудь явная линейная связь наблюдается с алкоголем (положительная) и с летучей кислотностью (отрицательная)

На певрвый взгляд, без применения методов машинного обучения можно предположить, что на качество в итоге сколько-нибудь значимо влияют сульфаты, плотность, лимонная кислота, алкоголь, летучая кислотность, общий диоксид серы исходя из анализа коррелограммы.

Для более подробного анализа характеристик распределения показателей, я привёл ниже таблицу с эмпирическими плотностями (по диагонали), совместного распределения показателей (нижний левый угол), показатели корреляции и их значимости (правый верхний угол). [1]

[1] После того, как я перешёл к новой бинарной целевой переменной quality не было значительных изменений в коррелограммах, соответственно выводы остаются прежними. Это продемонстрировано в коде.

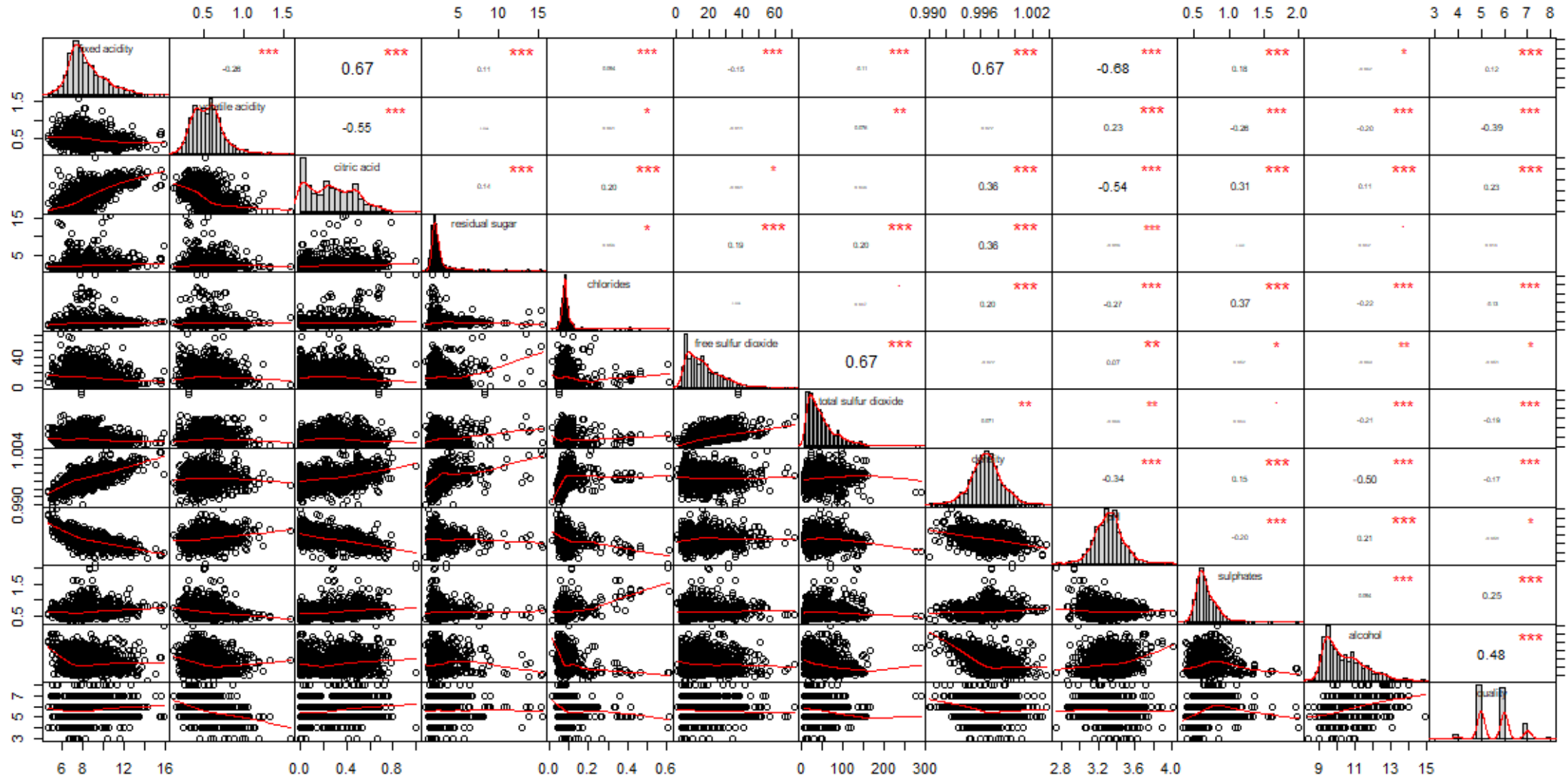

Рис. 3. Графический анализ распределений физико-химических характеристик вин.

Часть 2. Подготовка данных для работы SVM.

Для того, чтобы провести метод опорных векторов необходимо разделить выборку на обучающую и тестовую части, в моём случае я решил поделить исходную в пропорции 70% на 30% соответственно. Также отмечу, что для целевой переменной сохраняется баланс классов, то есть в обучающей выборке 70% нулей и 70% единиц нашей целевой переменной quality.

Часть 3. SVM.

Теперь, когда был произведён первичный анализ данных и они подготовлены для применения метода опорных векторов, можно приступить к непосредственной реализации алгоритма.

Зададим ряд экзогенных параметров для нашей модели: положим C = 5, gamma = 2 (параметр ядра - Gaussian radial basis function), k = 10 (параметр для расчёта ошибки предсказание на обучающей выборке с помощью кросс-валидации).

В зависимости от этих показателей построим предсказание качества вина в зависимости от всех доступных нам характеристик по обучающей выборке.

Ошибка предсказания на тестовой выборке оставила 0.2860125, то есть примерно 28,6%, согласно результатам, получаем, что в классе некачественных вин (0) содержится 460 опорных векторов, а в классе качественных вин (1) 528. Суммарно мы имеем 988 опорных векторов. В качестве самопроверки я провёл процедуры в двух разных пакетах с помощью функций svm (e1071) и ksvm (kernlab), результаты совпали.

Рассчитаем ошибку кросс валидации для случая применения метода опорных векторов для классификации качества вина по параметрам уровень содержания алкоголя (alcohol) и остаточного сахара (residual sugar).

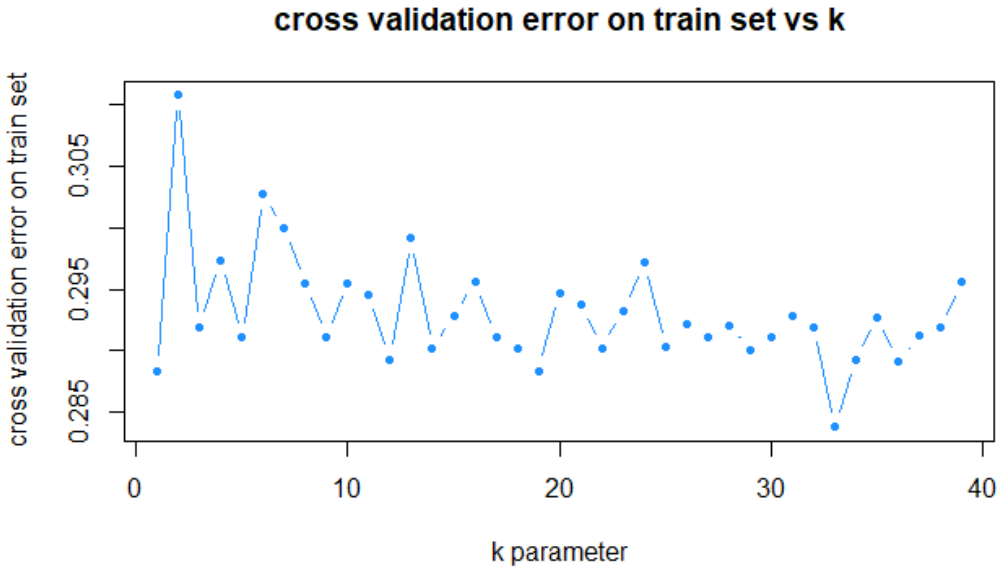

Рис. 4. Взаимосвязь параметра кросс валидации (количесвта фолдов на которые разделяется обучающая выборка с выделения искусственной тестовой подвыборки из состава обучающей, на которой происходит подсчёт ошибки кросс валидации) с ошибкой кросс валидации на обучающей выборки.

Интерпретация графика: в ситуации, когда обучающая подвыборка случайно перемешивается и из неё выделяется 33 новых фолда (подподвыборки), из которых на 32 происходит обучение, а на последнем проверка результата, то есть подсчёт ошибки – ошибки кросс-валидации, мы наблюдаем её самую наименьшую среди всех обозримых альтернативных вариантов деления обучающей выборки (ошибка кросс-валидации на обучающей выборке меньше 28,5%). То есть происходит подсчёт ошибок на каждом из псевдо тестовом фолде, который мы выделили благодаря кросс-валидации (в данном случае на 33), так делается 33 раза, ошибка усредняется и получаем ошибку кросс-валидации. Такая логика работает для любого k, в данном случае хуже всего разделять на 2 фолда обучающую подвыборку, там самая большая ошибка кросс-валидации на обучающей выборке (превышает 30,5%).

Для случая моего варианта (k = 10) ошибка кросс-валидации составила 0.2982143, то есть примерно 29,8%.

Иллюстрации разных результатов.

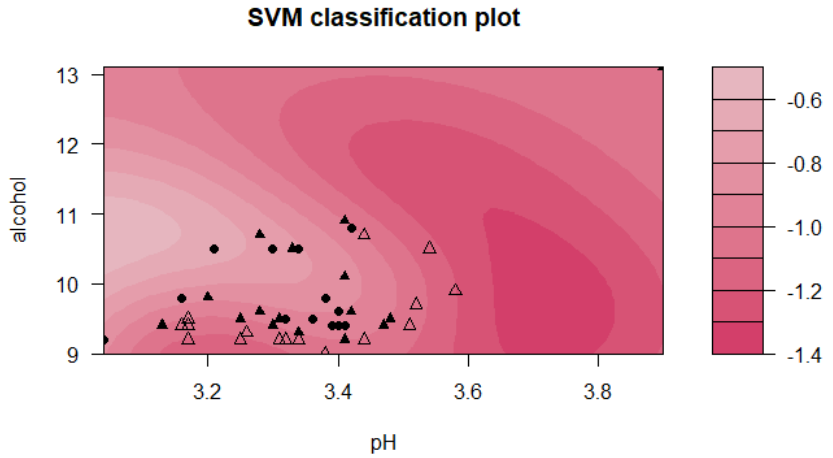

Рис. 5. SVM с параметрами C = 1.9, gamma = 0.2 для случая оптимизации качества по параметру алкоголя и pH.

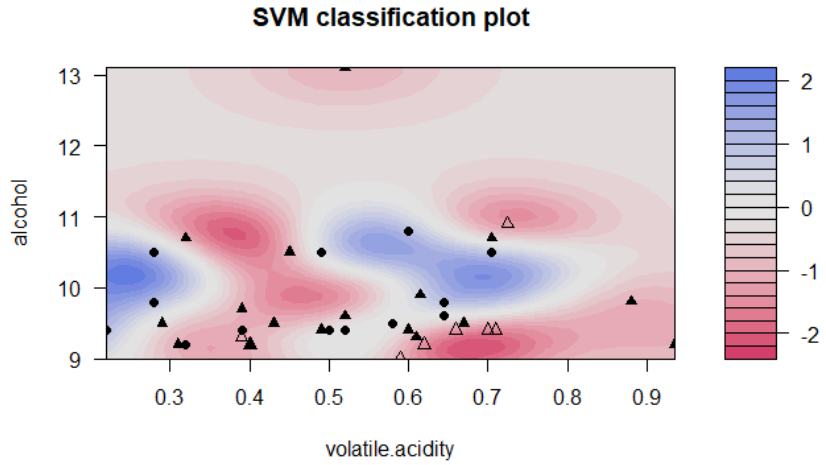

Рис. 6. SVM с параметрами C = 5, gamma = 2 для случая оптимизации качества по параметру алкоголя и летучей кислотности.

Оптимизация параметров.

Теперь подберём по обучающей выборке наилучшие значения параметров C и gamma с помощью кросс-валидации, где за количество фолдов отвечает параметр k, соответствующий 10.

Для параметра C будем рассматривать в качестве претендентов значения от 1 до 30 с шагом 1, а для gamma от 0,1 до 3 с шагом 0,1.

В результате оптимизации параметров модели сэмлинговым методом 10 фолдовой кросс-валидации получаем, что наилучшая модель имеет параметры C = 2, gamma = 0,2; При этом минимальная ошибка, которую удалось достичь при этих параметрах составила 0.2267857, то есть примерно 22,7%. Что касается опорных векторов, то модель показала, что в некачественном классе (0) их 353, тогда как в качественном классе (1) их 384, соответственно суммарное значение равняется 737.

Дополнительная оптимизация.

Мною было опробовано несколько способов улучшения показателя ошибки: я менял такие показатели модели, как cost, gamma, epsilon, tolarance, cachesize. В итоге мною было замечено, что в случае радиального ядра подбор параметров оптимизации может быть разным с зависимости от того, с какими иными параметрами идёт сопоставление, например, полагая cost = 2, gamma = 0.2, и оптимизируя показатели epsilon, tolarance, cachesize можно получить скор хуже чем до их оптимизации, поскольку важнейшие показатели cost и gamma меняют своё оптимальное значение при изменении epsilon, tolarance, cachesize. Важным замечанием является тот факт, что радиальное ядро лучше линейного.

Лучший результат у меня показала модель tunedModel_2 с радиальным ядром, 10 фолдами кросс-валидации, cost = 1.6 и gamma = 0.26. Её скор составил 0.2223214, конкуренты, уступившие ей расписаны в скрипте кода.