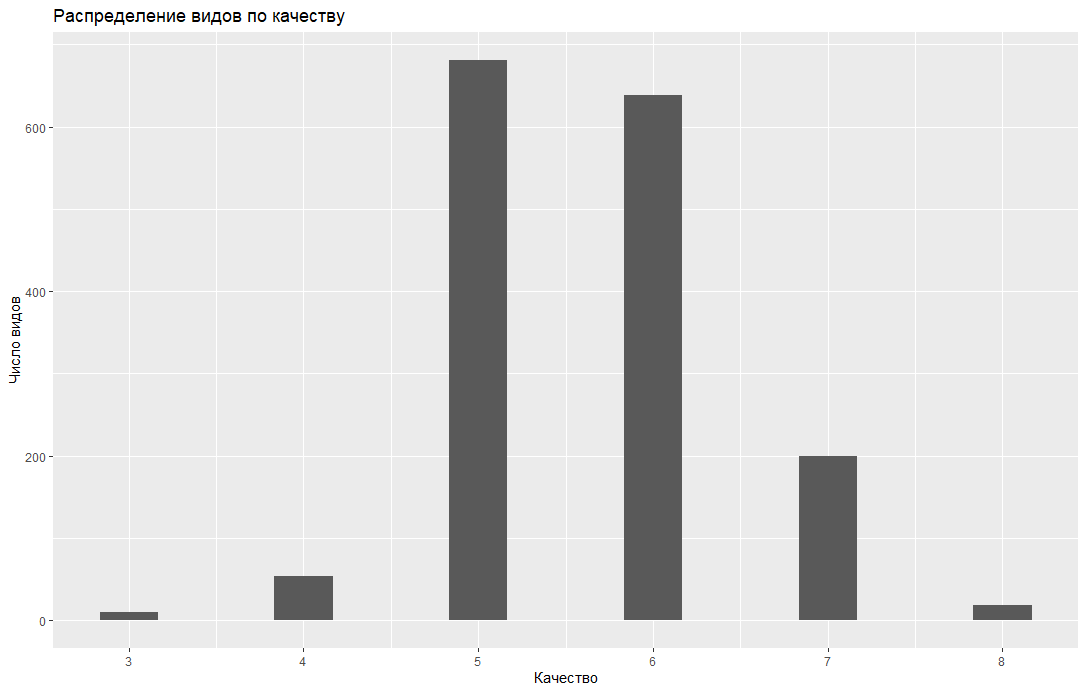
Supporting vectors machine

### C = 20, gamma = 0.5, k = 5

# Пункт 1

Ниже я представлю гистограмму распределения видов вин по качеству («quality»):



Таким образом, я выберу для пунктов 2-4 следующую классификацию: виды вин с качеством будут соответствовать классу «плохих» вин (class2), а виды вин с качеством будут соответствовать классу «хороших» вин (class1). Таким образом классификация будет наиболее сбалансированной, и, вероятно, даст лучшие результаты.

Вследствие такого деления виды вин разделяются по классам следующим образом:

class2 class1

744 855

# Пункт 2

Разделил выборку на обучающую и тестовую. В обучающую случайно были отобраны 75% наблюдений из исходной выборки. Для калибровки SVM-модели предсказания класса вина я задал параметры C = 20, . В используемых пакетах , однако, если и рассматриваются как канонические параметры ядерных функций, между ними, вроде как, должно быть такое соотношение: . В данном пункте и далее использовалось Гауссовское ядро. Предсказание класса в данном пункте задания проводилось по всем наличествующим характеристикам, кроме качества.

|  |  |  |
| --- | --- | --- |
| Pred/Real | Class2 | Class1 |
| Class2 | 137 | 50 |
| Class1 | 49 | 163 |

Ошибка классификации на тестовой выборке составила: 0.2481203.

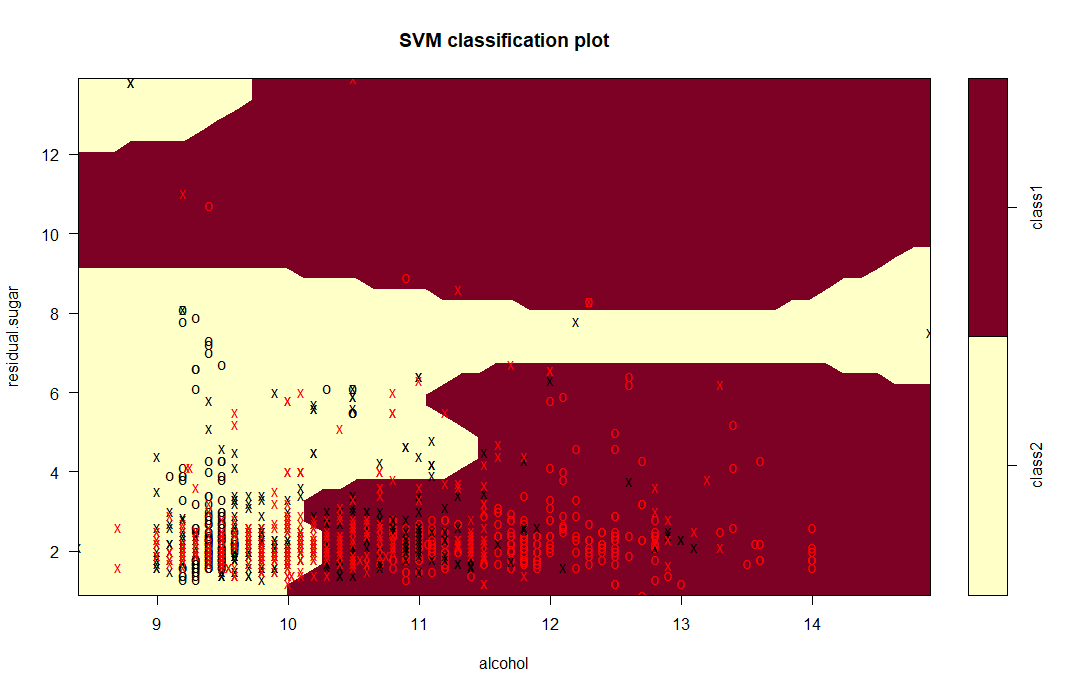
Опорных векторов в классе «плохих» вин (class2): 378, в классе «хороших» (class1): 438.

# Пункт 3

Предсказание класса в данном пункте задания проводилось только по характеристикам: alcohol и residual.sugar.

Ошибка классификации на обучающей выборке составила: 0.285.

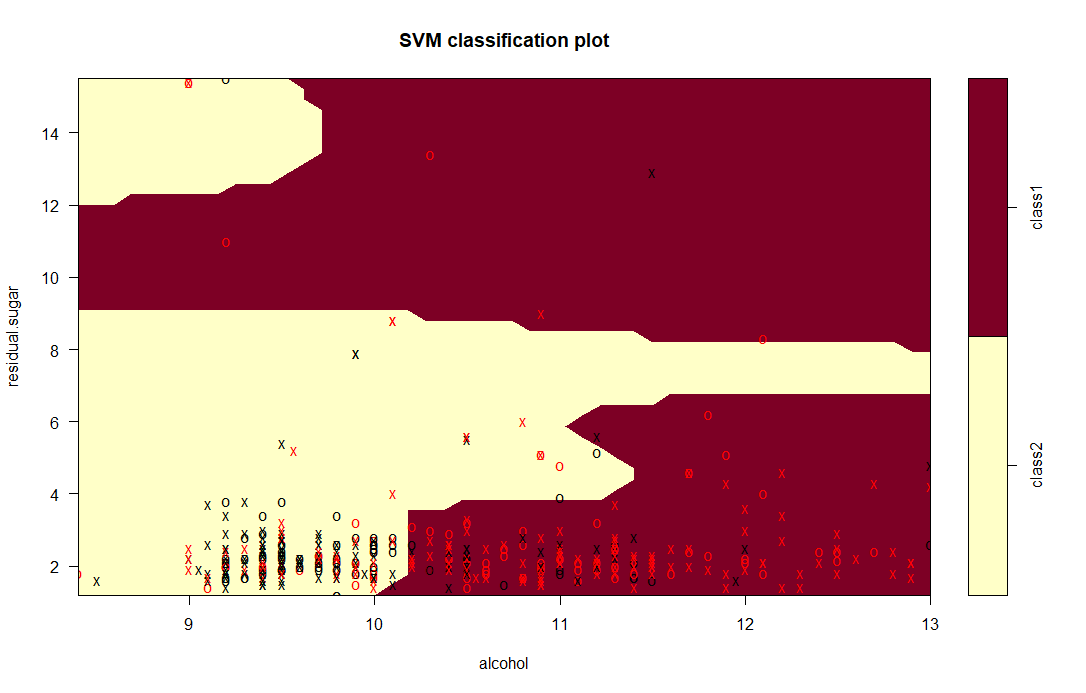
Кросс-валидационная ошибка предсказания класса на обучающей выборке составила 0.295 (по модели пакета kernlab).



На данном графике представлено расположение опорных векторов “X” и остальных наблюдений “0” в пространстве . Красные точки соответствуют классу «хороших» вин: class1, а чёрные – классу плохих: class2. Плоскость поделена на области двух цветов в пространстве . Этим областям в пространстве соответствуют указанные выше классы. Таким образом каждой точке пространства ставится в соответствие единственным образом класс из двух возможных (class2, class1), а соответствие это задаётся откалиброванной моделью.

**Лирическое отступление.** Хочется заметить, что и в данном пункте, и в пункте, где использовались все доступные характеристики, по итогу оказывалось достаточно много (примерно половина) опорных векторов, что указывает на недостаточность наличествующих факторов для «достаточно хорошего» предсказания класса.

Ниже представлю иллюстрацию того же типа, но для тестовой выборки.



# Пункт 4

Лучшими параметрами на основе кросс-валидационной ошибки по итогам калибровок модели классификации стали

Ошибка классификации для наилучшей модели на обучающей выборке составила: 0.05583333

Ошибка классификации для наилучшей модели на тестовой выборке составила: 0.2180451

# Пункт 5

Для подбора наилучшего среза по качеству для определения классов вина я итерирую по таким барьерам со значениями от 3 до 8. (т.к. качество дискретно и принимает значения от 3 до 8, вообще логично рассматривать от 4 до 7, но это не мешает никак получить результат). Далее, я оптимальными методами подбираю значения sigma и C и считаю ошибку предсказания класса на тренировочной выборке. Минимальной ошибке из ошибок подсчитанным вышеуказанным способом соответствует , как я и предполагал вначале.