NYC DATA SCIENCE ACADEMY

NYC Data Science Bootcamp

Support Vector Machines

Question #1: Wine Quality

Read in the `[10] Wine Quality.csv` dataset into your workspace. The data contains 1,599 observations of red Vinho Verde wines from the north of Portugal. The goal is to model wine quality based on various physicochemical measurements.

1. Perform some data munging:
    a. Recode the `quality` variable to be a factor variable with values of "Low" for quality ratings of 5 and below, and "High" for ratings of 6 and above.
    b. Scale and center the numeric vectors of your dataset.
2. Split the data into a training and test set with an 80% - 20% split, respectively. (**NB:** Use `set.seed(0)` so your results will be reproducible.)
3. Briefly explore some graphical EDA:
    a. Explain why a maximal margin classifier is impossible to fit to this data. (**NB:** Do not try to fit the maximal margin classifier.)
    b. Explain why a support vector classifier is generally more desirable.
4. Tune a support vector classifier with a cost ranging from 10^{-5} to $10^{-.5}$; try using the code snippet `cost = 10^(seq(-5, -.5, length = 50))`. **Caution:** This will take about a minute to run. (**NB:** Use `set.seed(0)` so your results will be reproducible.)
    a. What was the best cost parameter of the ones you tested?
    b. What was the best error rate corresponding to the best cost?
    c. Graphically view the cross-validated results. Is it plausible that you checked enough values of cost?
5. How many support vectors are there in your best support vector classifier?
6. What is the test error associated with the best support vector classifier you found in part 4?

7. Fit a support vector classifier to all of the data using the best cost parameter you found in part 4.
    a. How many support vectors does this support vector classifier have?
    b. Is the 555th observation a support vector?
8. What is the overall error rate for the support vector classifier you created in part 7?
9. Visualize the support vector classifier by examining the free sulfur dioxide and total sulfur dioxide cross-section; to do so, use the following line of code (modified with your object names): `plot(model, data, free.sulfur.dioxide ~ total.sulfur.dioxide)`.
10. Tune a support vector machine with a radial kernel. Check both cost and gamma values using the following code snippets: `cost = seq(.75, 1.25, length = 5), gamma = seq(.55, .95, length = 5)`. **Caution:** This will take about a minute to run. (**NB:** Use `set.seed(0)` so your results will be reproducible.)
    a. What was the best cost parameter of the ones you tested?
    b. What was the best gamma parameter of the ones you tested?
    c. What was the best error rate corresponding to the best cost & gamma?
    d. Graphically view the cross-validated results. Is it plausible that you checked enough values of cost and gamma?
11. How many support vectors are there in your best support vector machine?
12. What is the test error associated with the best support vector machine you found in part 10?
13. Fit a support vector machine to all of the data using the best cost and gamma parameters you found in part 10.
    a. How many support vectors does this support vector machine have?
    b. Is the 798th observation a support vector?
14. What is the overall error rate for the support vector machine you created in part 13?
15. Visualize the support vector machine by examining the free sulfur dioxide and total sulfur dioxide cross-section; to do so, use the following line of code (modified with your object names): `plot(model, data, free.sulfur.dioxide ~ total.sulfur.dioxide)`.
16. List a pro and con for both:

a. The best support vector classifier you found in part 7.
b. The best support vector machine you found in part 13.