Homework 6: Support Vector Machines

Overview

Due Sunday by 11:59 pm.

Fork the `problem-set-6` repository

Conceptual exercises

Non-linear separation

1. (15 points) *Generate* a simulated two-class data set with 100 observations and two features in which there is a visible (clear) but still non-linear separation between the two classes. *Show* that in this setting, a support vector machine with a radial kernel will outperform a support vector classifier (a linear kernel) on the training data. *Which* technique performs best on the test data? Make plots and report training and test error rates in order to support your conclusions.

SVM vs. logistic regression

We have seen that we can fit an SVM with a non-linear kernel in order to perform classification using a non-linear decision boundary. We will now see that we can also obtain a non-linear decision boundary by performing logistic regression using non-linear transformations of the features. Your goal here is to compare different approaches to estimating non-linear decision boundaries, and thus assess the benefits and drawbacks of each.

2. (5 points) Generate a data set with $n = 500$ and $p = 2$, such that the observations belong to two classes with some **overlapping, non-linear** boundary between them.
3. (5 points) Plot the observations with colors according to their class labels (y). Your plot should display X_1 on the x-axis and X_2 on the y-axis.
4. (5 points) Fit a logistic regression model to the data, using X_1 and X_2 as predictors.
5. (5 points) Obtain a predicted class label for each observation based on the logistic model previously fit. Plot the observations, colored according to the *predicted* class labels (*the predicted decision boundary should look linear*).
6. (5 points) Now fit a logistic regression model to the data, but this time using some *non-linear function* of both X_1 and X_2 as predictors (e.g. $X_1^2, X_1 \times X_2, \log(X_2)$, and so on).
7. (5 points) Now, obtain a predicted class label for each observation based on the fitted model with non-linear transformations of the X features in the previous question. Plot the observations, colored according to the new *predicted* class labels from the non-linear model (*the decision boundary should now be obviously non-linear*). If it is not, then repeat earlier steps until you come up with an example in which the predicted class labels and the resultant decision boundary are clearly non-linear.
8. (5 points) Now, fit a support vector classifier (linear kernel) to the data with *original* X_1 and X_2 as predictors. Obtain a class prediction for each observation. Plot the observations, colored according to the *predicted* class labels.
9. (5 points) Fit a SVM using a non-linear kernel to the data. Obtain a class prediction for each observation. Plot the observations, colored according to the *predicted* class labels.
10. (5 points) Discuss your results and specifically the tradeoffs between estimating non-linear decision boundaries using these two different approaches.

Tuning cost

In class we learned that in the case of data that is just barely linearly separable, a support vector classifier with a small value of `cost` that misclassifies a couple of training observations may perform better on test data than one with a huge value of `cost` that does not misclassify any training observations. You will now investigate that claim.

11. (5 points) Generate two-class data with $p = 2$ in such a way that the classes are just barely linearly separable.
12. (5 points) Compute the cross-validation error rates for support vector classifiers with a range of `cost` values. How many training errors are made for each value of `cost` considered, and how does this relate to the cross-validation errors obtained?
13. (5 points) Generate an appropriate test data set, and compute the test errors corresponding to each of the values of `cost` considered. Which value of `cost` leads to the fewest test errors, and how does this compare to the values of `cost` that yield the fewest training errors and the fewest cross-validation errors?
14. (5 points) Discuss your results.

Application: Predicting attitudes towards racist college professors

In this problem set, you are going to return to the GSS question from last week and predict attitudes towards racist college professors. Recall, each respondent was asked "Should a person who believes that Blacks are genetically inferior be allowed to teach in a college or university?" Given the kerfuffle over Richard J. Herrnstein and Charles Murray's *The Bell Curve* and the ostracization of Nobel laureate James Watson over his controversial views on race and intelligence, this analysis will provide further insight into the public debate over this issue.

`gss_*.csv` contain a selection of features from the 2012 GSS. The outcome of interest `colrac` is a binary variable coded as either `ALLOWED` or `NOT ALLOWED`, where 1 = the racist professor should be allowed to teach, and 0 = the racist professor should **not** be allowed to teach. Documentation for the other predictors (if the variable is not clearly coded) can be viewed here. Some data pre-processing has been done in advance for you to ease your model fitting: (1) Missing values have been imputed; (2) Categorical variables with low-frequency classes had those classes collapsed into an "other" category; (3) Nominal variables with more than two classes have been converted to dummy variables; and (4) Remaining categorical variables have been converted to integer values, stripping their original labels.

This week, building on last week's problem set, you will approach this classification problem from an SVM-based framework.

15. (5 points) Fit a support vector classifier to predict `colrac` as a function of all available predictors, using 10-fold cross-validation to find an optimal value for `cost`. Report the CV errors associated with different values of cost, and discuss your results.
16. (15 points) Repeat the previous question, but this time using SVMs with **radial** *and* **polynomial** basis kernels, with different values for `gamma` and `degree` and `cost`. **Present and discuss your results** (e.g., fit, compare kernels, cost, substantive conclusions across fits, etc.).