STATS216v – Introduction to Statistical Learning
Stanford University, Summer 2016

Practice Final

Instructions: (This is a practice final and will not be graded.)

- Remember the university honor code.
- Write your name and SUNet ID (ThisIsYourSUNetID@stanford.edu) on each page.
- There are 25 questions in total on the real final exam. All questions are of equal value and are meant to elicit fairly short answers: **each question can be answered using 1 - 5 sentences.**
- You may not access the internet during the exam.
- You are allowed to use a calculator, though any calculations in the exam, if any, do not have to be carried through to obtain full credit.
- You may refer to your course textbook and notes, and you may use your laptop provided that internet access is disabled.
- Please write neatly.

---

1. There has been recent debate in biology on whether generosity is hereditary. To investigate the question, a researcher runs a linear regression using the amount of money donated by a given person as the response, and a certain collection of predictors. Later he reruns the regression, but now includes the amount of money donated by the person's parents as a predictor. He finds that with the additional predictor the RSS of the model goes down, and therefore claims there is evidence to conclude that generosity is hereditary. Is his reasoning sound? Explain.

   It is not. The training RSS can never increase when we include another regressor in our linear regression model, and almost always decreases. Furthermore, even if parent donations are predictive of child donations, this would not imply that generosity is hereditary because it does not take a researcher to establish that wealth, and therefore opportunity to donate, is hereditary. Note: Either explanation is valid. It is not necessary to provide both.

2. Suppose you run a simple linear regression of a response Y against a single predictor X. You find that the R^2 is 0.862. What do you expect would happen to the R^2 if we instead treated X as the response and Y as the predictor? Explain.

   The R^2 value would still be 0.862. This is because in simple linear regression the R^2 is simply the square of the sample correlation coefficient between X and Y.

3. Your colleague is studying a collection of 100 manuscripts, 40 of which are signed and authored by Alexander Hamilton, 30 of which are signed and authored by James Madison, and 20 of which are signed and authored by John Jay. The remaining manuscripts are of unknown authorship, but each was written by one of these three individuals. Your colleague has identified a collection of stylistic features that can be extracted from each document that she feels should be indicative of authorship. She would like to use these features to identify the author of each of the unknown documents. Suggest two ways of carrying out this analysis, and describe one advantage that each has over the other.

   One option is to use multiclass LDA; a benefit of this method over the one to follow is that this method produces probability estimates. A second is to use One-vs.-all SVMs; a benefit of this approach is that it should work well even when the Gaussianity assumption of LDA is a poor approximation of reality.

4. Explain how you could use the bootstrap to estimate the test MSE of an arbitrary regression procedure.

   I would produce an OOB estimate! That is, I would repeatedly sample bootstrap datasets, train my procedure on each dataset, and, for each point in my original dataset not included in a bootstrap dataset, compute the squared prediction error for that bootstrap dataset model on that out-of-bag datapoint. Averaging over all of those squared prediction errors and taking the square root yields an OOB estimate of test MSE.

5. Assume that you have p predictors available in your dataset.

   (a) What is a (non-computational) motivation for considering $m = \sqrt{p}$ predictors over $m = p$ predictors at each split in a random forest?

   By not using all predictors at each split, we produce a more diverse set of models with less correlated predictions; averaging less correlated predictions leads to greater variance reduction.

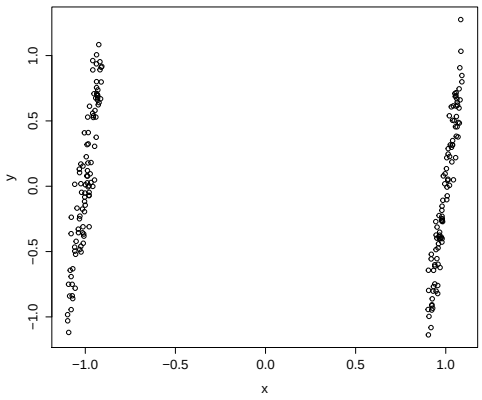
   (b) What is an advantage of considering $m = \sqrt{p}$ predictors over $m = 1$ predictor at each split in a random forest?

   If there are many irrelevant predictors in the dataset and few relevant ones, using $m = 1$ can lead to larger, lower quality decision trees that do not generalize well, since the tree is simply required to split on a randomly selected (and likely irrelevant) feature at each decision node.

6. Suppose we fit a linear spline, but we have the constraint that at the knots the fitted curve must be both continuous and have continuous first derivative. What simpler method does this become?

   The continuous first derivative constraint means we will have simply a straight line. This becomes ordinary linear regression.

7. For the data plotted below, find two functions of x (let's call them $f(x)$ and $g(x)$) such that y is well approximated as a linear function of $f(x)$ and $g(x)$. That is, find $f(x)$ and $g(x)$ such that y can be reasonably modeled as $y = \beta_0 + \beta_1 f(x) + \beta_2 g(x) + \epsilon$ for ϵ small Gaussian noise. Explain your answer.



It appears that y is proportional to $x - 1$ when $x > 0$ and y is proportional to $x + 1$ when $x < 0$. Hence, we can propose the linear model $y = \beta_0 + \beta_1 \mathbb{I}(x > 0)(x - 1) + \beta_2 \mathbb{I}(x < 0)(x + 1) + \epsilon$.

8. You are considering a binary classification problem in which the decision boundary separating your classes is a cubic polynomial in your $p = 10,000$ input predictors. However, it is computationally prohibitive for you to explicitly construct the 167 billion cubic interaction terms $x_{ij}x_{ik}x_{il}$ associated with each datapoint. Suggest a way to find a classification rule that separates your classes without explicitly forming cubic interaction terms.

I would fit an SVM with a cubic polynomial kernel $k(x, y) = (1 + \langle x, y \rangle)^3$.

The method might very well perform poorly with the actual test set.