Final Exam

Statistical Computing & Machine Learning

Spring 2016

1. LDA and QDA

Answer these questions about LDA and QDA:

1. Are they methods for regression, classification, both, or something else entirely?
2. Are they supervised or unsupervised?
3. What features in data would suggest use of LDA or QDA?
4. Which of LDA or QDA has lower bias? Which has lower variance? Or, does it depend on the data?

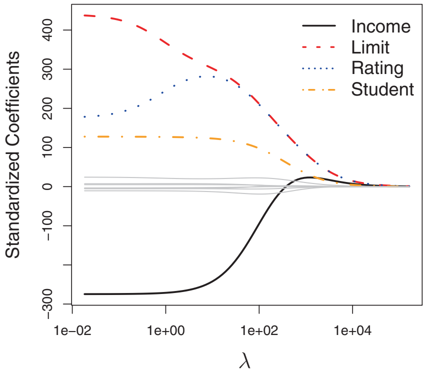
2. K-nearest neighbors

1. Why would one want to standardize predictor variables before using them in K-nearest neighbors?
2. Would using the rank transform of a predictor be a form of standardization? What's the usual form of standardization?
3. Is large K or small K associated with high model variance?

3. *Regularization in linear models*

Here's a figure showing and example of how linear model coefficients change with a "cost" parameter λ.

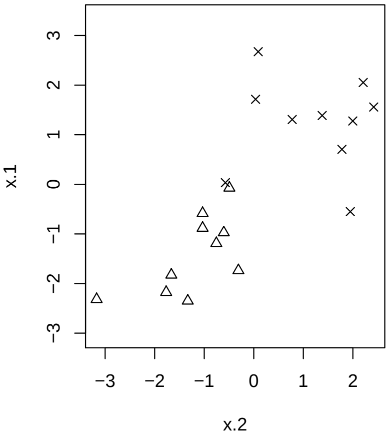
1. Is this pattern typically associated with ridge or lasso regression?
2. Sketch a reasonable graph (using your imagination) of what the pattern would look like for the other form of regression than you gave in (1).
3. Sketch a reasonable graph (from your imagination) showing both how the bias and variance of a fitted model change with λ.



4. *Support Vector Machines*

The graph in the margin shows several cases being used in the training set for building a classifier. There are two predictor variables, X_1 and X_2. The response variable is indicated by the shape of the dots.

Assume that you have set the cost parameter in a support vector machine (with a linear kernel) to a value that produces 4 support vectors. With just the calculations you can do in your head, circle the 4 support vectors and sketch in a plausible decision boundary and margins for your support vector machine.



5. *Variables working together*

Xavier and Zelda are discussing their Math 253 project to predict the yield of corn as a function of hours of sunlight, inches of rain, and amount of fertilizer used. They expect that the effect of rain on corn yield will depend on both the hours of sunlight and the amount of fertilizer. If they were building a linear regression modifier, they would incorporate one or more *interaction terms* into their model to account for this co-dependence. But they are planning to use a machine-learning form of predictor function.

For each of these model architectures indicate whether a new variable reflecting interaction **must** be added explicitly to capture the co-dependence.

a. k nearest neighbors
b. splines
c. ridge regression
d. lasso regression
e. principal components regression
f. a locally linear smoother
g. regression tree
h. boosting
i. random forests

6. Programming idioms

Describe a common use in machine learning for each of these functions or forms of expression. Assume that `Data_set` is a data frame with predictor variable `x` and response variable `y`. If the expression seems to reflect a mistaken conception, say so.

1. `sample()`
2. `Data_set[ - inds, ]`
3. `sum(y^2 - y_preds^2)`
4. `sum((y - y_preds)^2)`
5. `ifelse(p > 0.5, p, 1 - p)`
6. `sum(log(probs))`
7. `log(sum(probs))`
8. `prod(probs)`
9. `with(Data_set, cbind(1, x, x^2, x^3))`
10. `(x - sd(x)) / mean(x)`
11. `(x - mean(x)) / sd(x)`

7. Likelihood

Let's call the data you have collected on a certain topic D. You want to use these data to estimate a quantity of importance to you, which we'll call θ.

1. Here are several different probabilities

   i. $p(\theta|D)$
   ii. $p(D|\theta)$
   iii. $p(\theta)$
   iv. $p(D, \theta)$
   v. $p(D)$

   The following words may or may not be the conventional names for one or more of the above probabilities: **likelihood**, **posterior**, **leverage**, **prior**, **evidence**, **normalizing factor**, **influence**, **regularizing factor**. For each word that matches one of the probabilities given above, write the word next to the probability.

2. Now you are going to do some calculations. You can do these with paper and pen. If you need an arithmetic calculator to simplify fractions to a decimal form, etc. do so, but write down the calculation you are doing, not just the result.

   a. Abby and Bill are arguing about the value of θ. Both agree that θ is to be interpreted as the probability of outcome "H". Abby says that the data indicate that $\theta = 1/2$. Bill claims that $\theta = 1/3$ is a better value. The data set is $D = \{H, T, T, T, H, T, H\}$. Do a simple, appropriate calculation and explain whose claim it supports.

   b. Clara joins in the debate and argues that, even without the data, Abby's claim is more plausible than Bill's. Clara reckons that, even without the data, the odds are 2:1 in favor of Abby's claim. Incorporating Clara's view into appropriate calculations, explain whether Clara's view changes the result in (a). Remember to show your calculations.

8. Cross validation

Here is a very small, simple data set.

Y	X
A	1.0
A	1.8
B	3.2
A	4
B	5
B	5.8

You are evaluating a k-nearest neighbors classifer, $Y = \hat{f}(X)$.

Find the cross-validated confusion matrix with $k = 3$ using leave-one-out cross validation.