STATS216v – Introduction to Statistical Learning
Stanford University, Summer 2017

Midterm Exam (Solutions)

Duration: **1 hours**

Instructions:

- Remember the university honor code.
- Write your name and SUNet ID (ThisIsYourSUNetID@stanford.edu) on each page.
- There are 25 questions in total. 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. **Supervised vs. Unsupervised Learning**

   Your friend is doing a project in genetics. The goal is to identify major cell populations from a data set where each row represents a cell and each column is a measurement of gene expression level. Your friend uses a clustering algorithm on an unlabeled data set and claims that they have selected the best possible number of clusters by cross validation. Critique this claim.

   Your friend's claim does not make sense. Here we are looking at an unlabeled data set and there is no way to compute a cross-validation error estimate without training labels.

2. **Lasso for variable selection**

   Nasrin has a data set consisting of information of $n = 247$ patients and $p = 67$ measurments for each of them. She wants to fit a linear model that includes only the most important predictors. What is one way she can do this?

   She can use the Lasso. It will set many of the coefficients to zero, which will result in a more interpretable model. Alteranatively, she could use best subset or forward stepwise regression.

3. **LDA/QDA: flexible versus inflexible models**

   Suppose we are fitting a classifier based on n independent training observations with $p = 10$ predictors. If the number of observations n is very large, why do we expect QDA to have better predictive performance than LDA?

   QDA is more flexible than LDA, because it assumes differenct covariance matrices for each class. QDA fits a quadratic decision boundary, whereas LDA fits a linear decision boundary, so QDA can get closer to the best decision boundary. For large n, we will not be overfitting, so the extra flexibility will improve model performance. We can also state this in terms of the bias variance tradeoff: QDA has lower bias and for large n the variance of both LDA and QDA will be very small.

4. **Interpretting Linear Regression**

   A scientist is studying the relationship between two variables measuring human health: oxygen flow and muscle strength. From the scientist's past experience, he expects the two variables to be stongly positively correlated. He then fits a simple regression model with muscle strength as the response and oxygen flow as the predictor, on a data set of health measurements of Olympic athletes. To his surprise, the estimated slope is negative. What is a reasonable explanation for this outcome?

   Olympic atheletes can hardly represent the overall population in this problem. They have an overall much higher measurement in oxygen flow and muscle strength. Thus, it is reasonable that the simple regression estimates is dramatically different in this special group than in the whole population.

5. **Logistic Regression: Complete Seperation**

   We are trying to use logistic regression to classify crabs into two types. We have data about the width of 7 crabs in cm. There are 4 type 1 crabs with widths 5.8cm, 6.2cm, 7.7cm, 9.2cm and there are 3 type 2 crabs with widths 3.5cm, 5.2cm, 4.6cm. We run the glm function in R as follows:

```
glm(crab_type ~ width, family = binomial)
```

   We get the following output:

```
Call:
glm(formula = crab_type ~ width, family = binomial)

Deviance Residuals:
         1           2           3           4
-1.764e-05  -2.110e-08  -2.110e-08  -2.110e-08
         5           6           7
 2.110e-08   1.719e-05   2.110e-08

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    414.54  640578.21   0.001    0.999
width          -75.37  116131.19  -0.001    0.999

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9.5607e+00  on 6  degrees of freedom
Residual deviance: 6.0658e-10  on 5  degrees of freedom
AIC: 4

Number of Fisher Scoring iterations: 25
```

   Comment on the ouput.

   Here the two classes are completely separable and the maximum likelihood estimate of the logistic regression model goes to infinity. The software does not detect it and still gives an output after some iterations. The extraordinarily large standard deviation is a result of the complete seperation.

6. **Logistic Regression: Dummy Variables**

A researcher builds a logistic regression model to study the relationship between outcome of a automobile crashs (fatal, nonfatal) and a two qualitative features consisting of seat-belt usage (yes, no) and driver sex (male, female). After taking Stats 216, he decides to use dummy variables so that his logistic regression model is written as:

$$\log(\frac{\mathbb{P}(fatal)}{\mathbb{P}(nonfatal)}) = \theta + \alpha_1 x_1 + \alpha_2 x_2 + \beta_1 x_3 + \beta_2 x_4 \tag{1}$$

where x_1 is 1 if seat-belt usage is "yes", x_2 is 1 if seat-belt usage is "no", x_3 is 1 if sex is "male", x_4 is 1 if sex is "female", otherwise, they are 0. $\alpha_1, \alpha_2, \beta_1, \beta_2$ and θ are the coefficients to be estimated. He then realizes that there is an identifiability problem due to overparametrization. Explain the problem. How should he fix it?

Notice x_1 is 1 if and only if x_2 is 0 and similar relationship exists for x_3, x_4. There is no need to give two parameters for x_1 and x_2. The same goes for x_3 and x_4. We can delete the term with α_2 and β_2 in the model to fix it.

7. **Overfitting**

Alborz has just learned about model selection techniques and he wants to use it for selecting the best model for his chemistry experiment. He splits his data set into three parts: the training set, test set and validation set. After running 19 regression models on the training set and computing their errors on the test set, he choses the model which has the lowest prediction error on the test set. To his surprise, he then observes that the performance of this model is worse than many of the other models on the validation data set. Why might this be happening?

This is overfitting and selection bias (the "winner's curse"). Since he is choosing the best model from many different models, he may pick a model that simply got lucky on the test set, and performs worse on the validation data.

8. **Valid/Invalid CV**

Your supervisor gives you a project of predicting a real-valued response y based on 20 predictors $x_1, x_2, ..., x_{20}$. You built a ridge regression model and used 10-fold cross validation to select the tuning parameter λ. Just then, your supervisor told you that due to a technical error during the data collection, there are a lot of repeated measurements for several cases in the sample. Do you think this will be a problem for the model you just built? Explain.

Yes, this is a problem. Since the data contains many repeated measurements of the same cases, it can often happen in the cross validation step that the same data point is both in the training fold and the test fold. This will lead the CV error be too optimistic, which will lead to the wrong choice of λ.

9. **PCR and Linear Regression**

A researcher is trying to perform PCR on a data set with 10 predictors. In the first step, he picked the top 3 principal components, which together explain 95% of the variance in the data. He then fits the PCR with the selected principal components. His research assitant simultaneously runs a multiple linear regression with all the original predictors. This results in a higher R^2 than the PCR regression, which is confusing to the two researchers. Explain why this is in fact not surprising.

All the principal components are linear combinations of the original predictors and using all principal components will result in the same R^2 as using all the original predictors. Picking only the top 3 principal components will certainly have a lower R^2. PCR may have better predictive perfomance on new test data even if it has lower R^2; R^2 is computed only on the training data so higher R^2 may simply be overfitting.

10. **Boostrap Assumptions**

    A friend of yours is working on a stock price data set, trying to predict if the stock price will go up or down on a certain date using the stock prices in the days before. He fits a linear regression model with the past three days' stock price as predictors and the target day's stock price as the response. He then checks the distribution of the residuals and finds that they look quite different from a normal distribution. Because the residuals are not normal, he instead then applies the bootstrap on his data to get confidence intervals for his regression coefficients. Your friend's supervisor read his report and claims that his method is flawed. Do you agree with this assement? Why or why not?

    Your friend's supervisor is correct. Bootstrap assumes that the samples are independently identically distributed. Time series data is not independent and identically distributed, since oberservations are correlated with other observations close to them in time. Bootstrap sampling destroys this correlation, and so it is not a good approximation of the true data generating process. The bootstrap confidence intervals are not valid in this case.