STATS216v – Introduction to Statistical Learning
Stanford University, Summer 2017

Final Exam (Solutions)

Duration: **3 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. **Training data optimism**
   A student in the second week of stats 216v wishes to estimate the error of a linear regression model using cross-validation. The students fits the model and then splits the data set into 10 folds. On each fold, the student applies the model to each data point in the fold and computes the squared error. The student averages together these numbers to get an estimate of MSE for that particular fold. The student then averages the estimates from each of the 10 folds together to estimate the MSE of the model for future data. Explain why this estimate of MSE invalid and will be too low.

   The student is testing the model on the same data that was used to fit the model. This will lead to test error that is too small, because the model has already seen the test data. The student should instead re-fit the model within each step of CV without using the data in the test fold. This will give a fair estimate of the MSE.

2. **QDA, decision boundaries**
   A friend of yours is trying to predict the type abalones in a lake. There are known to be 5 types: A,B,C,D,E. From her biological expertise, she thinks that the decision boundary should be roughly quadratic. Name a method that would be appropriate for this task. If the true boundary is linear rather than quadratic, what will happened to the performance of this method?

   QDA,logistic regression on an enlarged set of features, or an SVM can fit a quadratic decision boundary. If the true boundary is linear, our method will be less data efficient since it will be more flexible than necessary (it has higher variance with the same bias as a linear classifier). With enough data, our method will still be a good classifier.

3. **Standardizing the inputs**
   Your friend is trying to apply PCA, regression tree analysis, and KNN classification analysis on a data set. Each predictor is measured in different units. Your friend has heard that standardizing the data is often a good idea for statistical learning tasks. For which of these methods will the results change if your friend standardizes the inputs?

   Standardizing will likely change the principal components and the KNN classification results since they are sensitive to scales. It won't change the result of regression trees since standardizing is a monotone transformation.

4. **Interpreting linear regression**
   You are a statistical consultant. A client from a social science department comes to you for help. Your client has just taken over a modeling project that was started by someone else who has recently left Stanford to join a start-up. Your client does not know where to begin, but has found a linear regression model that the previous researcher had built. Based on the output of the model shown below, what question do you think the researcher was investigating?

```
Call:
lm(formula = wage ~ age + education + race + education * race,
```

```
    data = Wage)

Residuals:
     Min       1Q   Median       3Q      Max
-111.583  -19.562   -3.682   15.013  223.409

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                                61.1697     3.3963  18.011  < 2e-16 ***
age                                         0.5726     0.0574   9.976  < 2e-16 ***
education2. HS Grad                        11.3784     2.7687   4.110 4.07e-05 ***
education3. Some College                   24.3673     2.9174   8.352  < 2e-16 ***
education4. College Grad                   39.4505     2.8891  13.655  < 2e-16 ***
education5. Advanced Degree                65.5441     3.1541  20.781  < 2e-16 ***
race2. Black                               -1.9513     6.9109  -0.282    0.778
race3. Asian                              -11.6063     9.5847  -1.211    0.226
race4. Other                               -2.9159    11.0901  -0.263    0.793
education2. HS Grad:race2. Black           -2.0005     7.8422  -0.255    0.799
education3. Some College:race2. Black      -5.2604     8.0119  -0.657    0.512
education4. College Grad:race2. Black     -12.3377     9.0598  -1.362    0.173
education5. Advanced Degree:race2. Black  -14.2217    10.1427  -1.402    0.161
education2. HS Grad:race3. Asian            0.6259    11.6164   0.054    0.957
education3. Some College:race3. Asian       1.7180    12.8717   0.133    0.894
education4. College Grad:race3. Asian      14.9252    10.6650   1.399    0.162
education5. Advanced Degree:race3. Asian    8.7520    10.8219   0.809    0.419
education2. HS Grad:race4. Other           -8.9230    14.9550  -0.597    0.551
education3. Some College:race4. Other      11.2658    16.9185   0.666    0.506
education4. College Grad:race4. Other     -35.4267    23.5492  -1.504    0.133
education5. Advanced Degree:race4. Other  -42.6169    27.7431  -1.536    0.125
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 35.86 on 2979 degrees of freedom
Multiple R-squared:  0.2665,Adjusted R-squared:  0.2616
F-statistic: 54.12 on 20 and 2979 DF,  p-value: < 2.2e-16
```

It appears that the researcher was looking at the relationship between income and age, education, and race. In particular, the researcher investigated the interaction effect between race and education. The researcher may have been interested in whether the effect of education changes for people of different races.

5. **Overfitting and regularization**
A genomicist is trying to predict the number of immune system cells using measurements of the expression of $p = 10,000$ genes. The genomicist has a large database

containing the gene expression levels for $n = 15,000$ patients, and the number of immune system cells for each patient has been measured. The researcher fits a linear regression, but finds that it has very poor predictive performance on the test set. What is a likely reason for this? What a better method for this task?

The model is probably overfitting. Here $\frac{n}{p} = \frac{3}{2}$, which is not very large. We should use ridge regression or lasso for instead to improve performance.

6. **Test and validation splitting**
Your friend is working on a classification problem with a data set containing 1000 samples. She is only interested in trying out SVM with radial kernel. To be cautious, she divided the data set into a 500-sample training set, a 250-sample test set and a 250-sample validation set. She trained SVM on the training set and then got the test error 0.18, validation error 0.17. Do you have any suggestions for her?

There is no need for a validation set since there is only one model concerned and no need for parameter selection. She can make better use of her data by just dividing the data set into training set and test set. She will then have a larger data set to fit for training. This will lead to a better estimate of the model performance of the final model.

7. **Bootstrap**
You are fitting a boosted decision tree model on $n = 500$ independent data points with $p = 100$ features each. You are are going to use your model to make a prediction $\hat{y}$ for a new data point x_{new}. You know the value of the features x_{new} but do not know the value of the response y of this new data point. How can you estimate the variance of the predicted value $\hat{y}$ of the new data point?

You can use the bootstrap.

Refit the boosted decision tree model many times on the bootstrapped sample version of the 500 data points. Each time, compute $\hat{y}$ for the data point x_{new}. We can use the variance of the $\hat{y}$ as an estimate of the variance of our prediction.

8. **Data transformations and linear regression**
Suppose we have a response y and two features x_1 and x_2. We think that the data comes from the following model:

$$y_i = e^{\epsilon_i} \beta_0 x_{i,1}^{\beta_1} x_{i,2}^{\beta_1}$$

with ϵ_i independent normal random variables with the same variance. How can we fit this model using linear regression?

We can consider $\log(y_i)$. After the transformation, this satisfies the assumptions of linear regression.

9. **Comparing nested models**
You are an engineer tasked with testing an experimental motor for a new electric car. You record the power output of the motor at different voltage inputs. You wish to know

if the the relationship between the power output and voltage input is purely linear or if it also has a component that is $\sqrt{\text{voltage}}$. How would you determine this using the measurements that you have collected?

We can fit a linear model containing and intercept, voltage, and $\sqrt{\text{voltage}}$. We can use the t-test to determine if the $\sqrt{\text{voltage}}$ component is statistically significant. If it is statistically significant, we can conclude that the relationship is not purely linear.

10. **Stepwise selection**
Aaron has run a forward stepwise procedure to select variables in a linear regression setting. After 3 steps, X_6 is in his model. He then runs the best subset selection procedure over all models with 3 predictors and is surprised because X_6 was **not** selected! What might have gone wrong?

Nothing has gone wrong. Forward stepwise is a greedy algorithm and in each step adds the predictor which reduces the RSS having the previous chosen predictors in the model.

11. **Interaction terms**
You and your teammates on the data science team at a car company are working on a project predicting the MPG(miles per gallon) of cars based on horsepower, acceleration and weight. Your supervisor expects the MPG to depend on the interaction among the three terms. Teammate A suggests using GAM model, and Teammate B suggests using a random forest model. Which model do you expect to perform better?

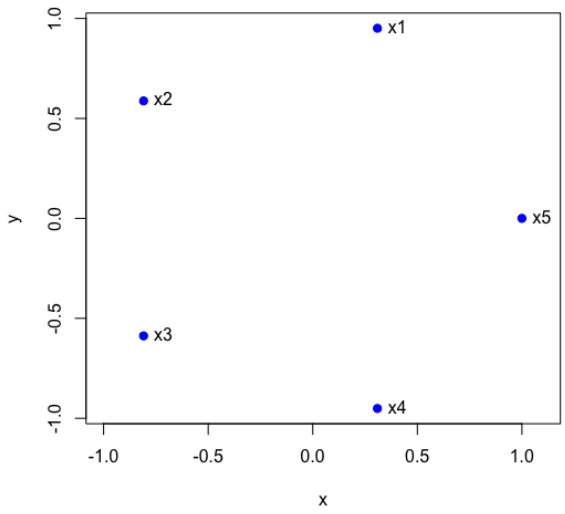
A GAM won't include the interaction between the predictors, but the random forest will, so we expect the random forest model to preform better.

12. **Random Forest OOB error**
You are fitting a random forest to a data set for a classification problem. At each split, the random forest chooses the best split from a random subset of p^α predictors, where p is the total number of predictors and $\alpha \in [0, 1]$. You intend to use cross validation to help you figure out the optimal value of the tuning parameter α to use. Two of your classmates approach you with their suggestions. Classmate A suggests that you can use out of bag error to select the optimal α instead of using cross validation and it will save a lot of computational effort. Classmate B says he agrees with A except that for $\alpha = 1$ there will be no out of bag error generated since each split is considering all p predictors. Which suggestion, if any, do you agree with?

Classmate A's suggestion is correct and helpful. However, classmate B's argument is wrong since the out of bag error is still produced when $\alpha = 1$ because random forest is bootstrapping the sample and it generates out of bag samples no matter how many predictors you are considering about in each split.

13. **Hierarchical clustering puzzle**
Suppose we have 5 vertices $x_1, x_2, \cdots, x_5$ of a regular pentagon (all of the sides have equal lengths) in a two dimensional space. Using hierarchical clustering, we form the clusters $(x_1, x_2), x_3, x_4, x_5$ in the first step and further cluster the points into

$(x_1, x_2, x_3), x_4, x_5$ in the second step. Consider using single or complete linkage hierarchical clustering. Which of theses methods, if any, could give the above clusters in steps 1 and 2?



Only single linkage.

Let d_1 be the distance between x_1 and x_2 and d_2 be the distance between x_1 and x_3. We have $d_1 < d_2$. Upon the second step, the complete and average linkage distance between (x_1, x_2) and x_3 is d_2 while the distance between x_3 and x_4 is d_1. Thus it is impossible for complete linkage to cluster (x_1, x_2, x_3) together in the second step. Using single linkage, we can cluster (x_1, x_2, x_3) together in the second step.

14. **GAM modeling**

A scientist is studying forest fires. She is investigating the effect of wind speed, average temperature and rainfall on the total area burned in a forest fire. She believes that the area burned should be a smooth function of the three predictors and the effect of the predictors have little interaction with each other. She would like a flexible model that allows the effect of each predictor to be very different in different regions of the input values. Propose a reasonable model for this setting.

Since there are little interaction between the predictors, we can use a Generalized Additive Model. For instance, we can try

$$area\ burned = \beta_0 + f_1(wind\ speed) + f_2(average\ temperature) + f_3(rainfall) + \epsilon$$

For each f_i, since the effect may behave differently across different ranges, a single polynomial function is not enough. Since we only want smoothness, using smoothing spline for each f_i is a reasonable answer. Using a regression spline is another option, but she would need to specify the knots in advance.

15. **Radial kernels**
    True or False: Perfect separation with SVM in two-class training data is always possible with distinct data points when using a radial kernel. Explain your answer.

    True. Provided no two points from different classes are ever exactly the same, the radial kernel's locality properties will lead to an arbitrarily flexible decision boundary. For sufficiently small kernel bandwidth in radial kernels, the decision boundary will look like you just drew little circles around the points whenever they are needed to separate the two classes.

16. **Regularization to fix perfect separation**
    A genomicist is trying to predict the risk of heart disease using measurements of the expression of $p = 5,000$ genes. She has a large database containing the gene expression levels for $n = 20,000$ patients, and for each patient it has been recorded whether or not they had heart disease. She fits a logistic regression, but finds that there is perfect separation so she knows that she cannot trust the fitted model. How could she modify logistic regression to fix this problem?

    She can add a regularization term, such as a lasso or ridge penalty.

17. **PCR**
    Ahmad has collected 1000 black and white 30x30 pixel pictures of his friends. He wants to predict how happy a person is on a scale of 1-100 based on their picture. For each observation in his training set, he knows how happy the person was at the time of the picture. He recently learned that since pictures of faces are usually similar, pictures of faces will only lie on a very small subspace of the 30x30 images. What is one method that will take advantage of this special structure?

    PCR. This will reduce the dimension of the input space and lead to a better model.

18. **Repeated measurements**
    A medical researcher wishes to build a classifier to classify tumors into two classes. She has $n = 10,000$ observations: 500 measurements each for 20 patients. 10 patients have a tumor of type A and 10 patients have a tumor of type B. For each observation, the researcher has measured $p = 100$ features. The researcher thinks that because she has such a large number of observations, she should fit a very flexible classifier like QDA or an SVM with a radial kernel. Do you agree? Explain your reasoning.

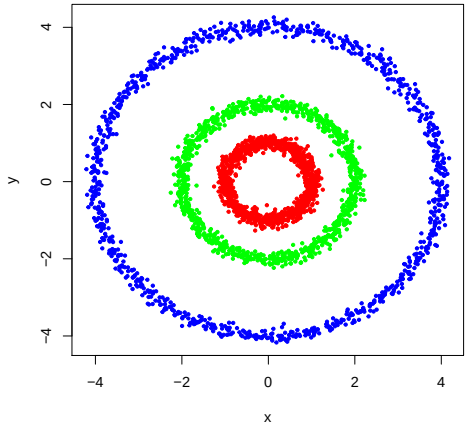
    We do not have $n = 10,000$ *independent* measurements, because many measurements belong to the same person. We only have observations from 20 distinct people. We should use a more inflexible method, to avoid overfitting.

19. **SVM support vectors**
    A data scientist is analyzing medical records. She is trying to predict whether or not a person will have to return to the hospital after surgery or not, based on $p = 20$ measurements taken during surgery. She fits a SVM model. After making some plots, she notices that there are some big outliers in her data set where the measurements were recorded improperly. These outliers appear to be classified correctly, but are very

far from the decision boundary of the SVM. Should she be worried about using the SVM model?

No, she does not need to worry. She can use the SVM model. The SVM fit depends only on support vectors: points which are close to the decision boundary. Since the outliers are not support vectors, the SVM fit is insensitive to these outliers.

20. **Hierarchical versus K-means clustering**
We have a data set of points coming from three different populations, marked by different colors in the following plot. Suppose we wish to recover the three populations without using the labels. Name one clustering method from the course that would work well on this data, and one clustering method that would not work well on this data.



Hierarchical clustering with single linkage and cutting the dendrogram for three clusters can help recover the three populations while k-means will not work well on such a data set.

21. **Random forests**
Lawrence is writing a paper for a machine learning conference. He has invented a 'new' version of random forests. To grow each tree of the random forest, he takes a sample of size n **without** replacement and then uses all features for every step of the tree fitting. Explain why Lawrence's method is not new.

If we take the samples without replacement, we will be using the same data set for each tree. Furthermore, if we use all of the features for each step of the fitting, then every tree we fit will be exactly the same. The final result will be the same as a single decision tree.

22. **Unsupervised data reduction**
You have a data set with the complete DNA sequence of $n = 5,000$ people. For each person, you know the base (A,C,G, or T) that is present at $p = 3,000,000$ locations of

the DNA sequence. You know that for each person, the base that is present at nearby locations along the DNA sequence are highly correlated. For future measurements, you know that you will not be able to measure every single feature, because it will be too expensive. How can you reduce the dimension so that for future measurements you will not need to measure every single feature, while preserving as much information as possible?

We can use hierarchical clustering on the features. For each cluster of features, we will only measure one representative feature from that cluster in the future. PCA will not work in this case, because we would still need to measure all of the features.

23. **Boosting and GAMs**
Explain why boosting decision trees with one split ("stumps") leads to a GAM:

$$f(x) = \sum_{j=1}^{p} f_j(x_j) \tag{1}$$

What kind of function is the resulting $f_i(x_i)$?

Boosting results in a linear combination of the base learners. If we are using decision stumps, the base learners are only a function of one variable. In this case $f_i(x_i)$ is the sum of piecewise constant functions, so it is piecewise constant function of x_i.

24. **Fitting complex model with CV**
A researcher fits a complex model to a training set of 1000 patients measured on 5000 features, to classify tumors as type A or type B. He does cross-validation and this tells him that a model with no parameters has the lowest error, and hence his model is not adding any predictive power. The researcher is puzzled, because he knows that one of his features, the concentration glucose, is known to be highly related to the type of tumor. Explain what may be happening.

If there are many features that do not have predictive power, then our fitting procedure may not be sensitive enough to find the few 'good' features. On each fold of the data, CV might detect that every model has poor performance due to overfitting. A better strategy may be for the researcher to build a simple model using a smaller number of features that he suspects are relevant. This model will likely have better predictive power.

25. **Cubic spline definition**
A researcher aims to fit a cubic spline model and writes it out as a piecewise cubic polynomial with continuous derivatives up to order 3 at each of the 10 knots $\xi_1, \cdots, \xi_{10}$. You realized that he has added to many constraints to his model. Why is this not a cubic spline, and what other model from stats 216v is this equivalent to? How can he modify his procedure to make this a cubic spline?

His model is equivalent to a degree 3 polynomial regression. He should only constrain the model to have continuous derivatives up to order 2.