STATS216v – Introduction to Statistical Learning
Stanford University, Summer 2017

Practice Midterm(Solutions)

Duration: **1 hour**

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

- Remember the university honor code.
- Write your name and SUNet ID (ThisIsYourSUNetID@stanford.edu) on each page.
- There are 10 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.** All answers should be written in the space provided between questions.
- You may not access the internet during the exam.
- 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. Your lab friend and you are working on different experiments, and you each end up fitting a regression to predict a relevant outcome (but a different outcome in each case). He gets a R^2 of 78%, while you only get 42%. He declares himself the winner. Comment.

   If the R^2 is on training data, then he may simply have overfit his data. So let's focus instead on test data. His prediction task might be much easier than yours, with the inherent noise in the system much lower; i.e. the SNR or "signal-to-noise-ratio" is higher for his problem than yours. You may have done a much better job in modeling, and achieving 42% might be more impressive on this tougher task than 78% on his easier task.

2. Marketing consultants are hired by a company to estimate how spending in magazine ads affect their profits. The company is interested in not only a pointwise estimate of the coefficient β_1 associated with magazine advertising expenditure, but also an interval that they could be 90% sure to contain the true coefficient β_1. Besides the least squares coefficient $\hat{\beta}_1$, what other information would the marketing consultants need to find such an interval?

   Since they are interested in a 90% confidence interval, they would also need the standard error of $\hat{\beta}_1$. The confidence interval would be of the form $[\hat{\beta}_1 - c\hat{\sigma}, \hat{\beta}_1 + c\hat{\sigma}]$.

3. In linear regression of Y on a set of p variables, are the following statements TRUE or FALSE?

   (a) When adding a second variable to a regression (after inclusion of the first), the RSS always decreases.
   FALSE. The second variable could be a copy of the first, or perfectly correlated with it. This can occur in non-obvious ways, like a factor with two levels can be *nested* in a factor with three levels.

   (b) Variables with coefficients having smaller standard errors are more relevant than those with larger.
   FALSE. The standard error of a coefficient depends on the scale; it is the ratio of coefficient to standard error that is important.

4. An ecologist would like to estimate the number of fish in a lake given the previous month's profit from the local fishery. You first consider applying a linear regression, but then the ecologist tells you that he expects the relationship between the predictor and response to be cubic, not linear. Can linear regression still be made appropriate for the task at hand? Explain.

   Yes. If we include not only the profit from the fishery but also its square and cube as regressors, linear regression should work fine.

5. The owners of a clothing store would like to decide who will receive the "employee of the month" award. Unfortunately, they have no data on the individual sales for each employee. For any given hour, the store owners know only the total number of pieces of clothing that were sold, and exactly which 7 employees were working. (The store is set up so that at any given hour there are always exactly 7 employees working.) Using linear regression, suggest a reasonable way to decide which employees contributed the most to the month's sales figure.

   I would set up a linear regression model using the hourly sales as response, and dummy variables for each employees as predictors, indicating whether they were working at hour i or not (important: should not include all the dummies in the regression). I would pick the employee with the largest associated regression coefficient to receive the award.

6. You read an article saying that the success of a movie can be reasonably represented as a linear function of the number of famous actors in the movie, its genre, and its budget. However, your friend, who is a movie producer, posits that the director's experience and the release date are also crucial in determining future profits. Assuming the profits follow a normal distribution, suggest a way to determine whether or not your friend is right.

   One way to determine who is right is to set up a linear regression for the model suggested by the article (the submodel), as well as a linear regression for the model suggested by your friend (the full model). Then we can do an ANOVA F-test to see if your friend is right.

7. A colleague suspects (but does not know for sure) that there are two distinct forms of red panda lung cancer, distinguishable by gene expression patterns in red panda cells. He has collected gene expression data from 10,000 red pandas with lung cancer and wants to use a classifier to discover the two cancer subgroups. Explain why a classifier is an inappropriate tool for this task.

   A classifier is an inappropriate tool, because this is an unsupervised learning task (and more specifically a clustering task). There are no labels to distinguish one hypothesized form of cancer from the other, so a classifier cannot be trained.

8. An outreach program is tasked with estimating the annual income of a household given its school district, neighborhood crime rate, proximity to clean drinking water, and roof type (there are three types: thatched, tin, or tile). The program only collected roof type information for a (uniformly random) 90% of households in its dataset and wants to impute (that is, fill in a best guess for) the missing roof type values prior to carrying out subsequent analyses. Suggest a reasonable way to impute the missing roof type values in a manner that makes use of the other features collected for each household.

   We could treat the imputation problem as a classification problem! First we could train any classifier that supports more than two classes (e.g., linear discriminant analysis) on those datapoints without missing data to predict the outcome (roof type) based on the other predictors. Then, we can use the predictions of that classifier on the remaining datapoints as our best guesses for the missing values.

9. You are interested in using K-fold cross validation to select the regularization parameter λ for a Lasso regression. However, you are not sure which value of K you should use for K-fold CV. Describe how you could use your data to select an appropriate value of K for K-fold CV for this problem.

   One reasonable option is to view this as a model selection problem and to use the validation set approach to select an appropriate value for K-fold CV. That is, withhold a fraction of your original dataset as a validation set and consider the rest to be your training set. Now, for each value of K, use K-fold CV to pick a regularization parameter λ_K based on the training set, fit Lasso regression using λ_K to the training set, and evaluate its error on the validation set. Finally, pick the value of K that yielded the smallest validation set error.

10. Suppose we run a forward stepwise linear regression procedure on a set of 12 predictor variables. We see that variable 3 enters first because it causes the biggest drop in RSS (over the mean). After adding (one-by-one) the next 5 variables, we pause to see which variable, if dropped, would increase the RSS the least. Could this be variable 3?

    Yes. Variable 3 might be redundant at this stage, even though in the beginning it was the best representative for the entire team of predictors.