Stat 202 **Final exam**

Your name: SOLUTION

The final exam is closed book and closed notes. No computers and calculators are allowed. Please show your work.

Scores:

1

2

3

4

5

6

7

8

9

10

Total

1. (3+3 points)
For each of the following two situations, state whether one should use k-means clustering or hierarchical clustering, and what choice of parameter for the clustering method to use (if any).

(a) You have height measurements of seventeen 12-year olds and thirteen 16-year olds and you want to cluster the height measurements into two groups.

(b) A geneticist has three dimensional measurements on a number of single cells. She believes that the cells should cluster into a very small number of groups based on similarity (Euclidean distance). She is not sure about the exact number of clusters, but she knows that no two cells in the same group will be more than a given distance D apart.

a K-means with K=2 (since we want to have 2 clusters and k-means is well suited for normally distributed clusters)

b She can perform hierarchical clustering with complete linkage and cut off the dendrogram at the value of D in order to select an appropriate number of clusters. (The use of complete linkage enforces the distance constraint on each cluster. Further, a clustering with a smaller number of clusters will violate the distance constraint, so this method results in a reasonable estimate of the smallest number of groups that satisfy the constraint. This is an appropriate approach since she knows that the number of groups is very small.

2. (5 points)
A data scientist fits a multivariate linear regression model $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_{20} X_{20}$ and finds all but three p-values are significant according to the t-test. He decides to drop those three variables and to keep all the remaining predictors. Comment.

This is not a good idea, for two reasons:

1. The t-test is a marginal test. If there is collinearity, then it is possible that after removing one predictor, a formerly non-significant predictor becomes significant.

2. Keeping all the remaining predictors is a problem because of the multiple testing fallacy. If no adjustment for multiple testing is made, then there may be several false positives.

3. (3+3 points)
For each of the following two classification problems, write down whether one should prefer LDA or logistic regression, and explain why:
(a) We want to predict gender based on height and weight. The training set consists of heights and weights for 82 men and 63 women.

(b) We want to predict gender based on annual income and weekly working hours. The training set consists of annual income and weekly working hours for 770 men and 820 women.

a LDA is preferable as heights and weights are normally distributed with the same SD for men and women.

b logistic regression is preferable as incomes do not follow the normal curve but have a long right tail.

4. (3+3 points)
(a) If the Bayes decision boundary is linear, do you expect LDA or QDA to perform better on the test set? Explain.
(b) What is your answer if we compare the performance on the training set?

a We expect LDA to do better as it can fit a linear boundary without bias. QDA can do that also but has higher variance because it is more flexible.

b QDA will have a lower training error due to overfitting, as it is more flexible.

5. (2+4 points)
We want to build a regression model and have many observations and many predictors.

(a) From a computational point of view, which of these two model building algorithms is preferable: best subset selection or forward stepwise selection?

(b) True or false and explain: Best subset selection will result in a smaller prediction error than forward stepwise selection because every model that is considered in forward stepwise selection is also considered in best subset selection.

a Forward stepwise selection: It is a greedy algorithm that looks at much fewer than all possible subsets, whereas best subset selection does ~~just that~~ the latter.

b False: Comparing all possible models increases the chance of overfitting. The greedy ~~algorithm of forward~~ progression of forward stepwise restricts the search space, hence reduces variance. This may result in a smaller prediction error.

6. (4+2+2 points)

(a) You want to apply a k-nearest neighbor classifier to a certain data set. Explain the various steps required to choose the value of k with 10-fold cross-validation.

(b) What are the advantages and disadvantages of k-fold cross-validation relative to the validation set approach?

(c) What are the advantages and disadvantages of k-fold cross-validation relative to LOOCV?

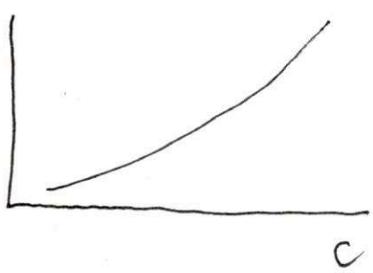
a Divide the data into 10 folds at random.
For $i=1,\dots 10$: Take the ith fold as test set,
combine the other 9 folds into a training set.
For $K=1$ to a large number (say 100): Predict the label of each test observation with the K-nearest neighbor classifier, i.e. by a majority vote among the labels of the K nearest training data. Compute the proportion of misclassified test data.
Average these proportions over $i=1,\dots 10$, resulting in a misclassification rate for each K.
Pick the K with the smallest misclassification rate.

b K-fold CV has more variance but less bias than the validation set approach. (for $K>2$)

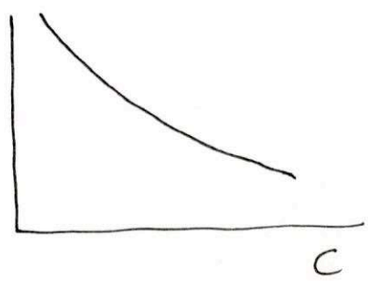
c K-fold CV has less variance but more bias than leave-one-out CV (for $k<n$)

7. (2+2+2+2 points) A support vector classifier comes with a 'budget' C, which is a bound on the distance of the observations to the margin over all observations that are on the wrong side of the margin. Draw sketches of
(a) the bias
(b) the variance
(c) the training error
(d) the test error
as a function of C.

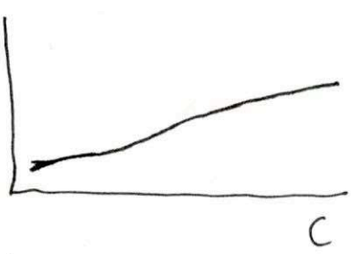
a



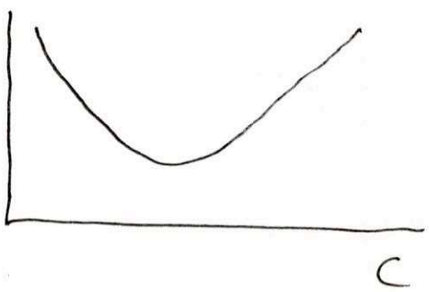
b



c



d



(The parameter C controls the bias-variance trade-off, so these are the usual curves we get for such situations. (For small values of C, the support vector classifier may not have a solution.))

8. (3+3 points)

(a) In the Kaggle competition, the public leaderboard is calculated with 30% of the test data. The final result will be based on the other 70% of the test data. As of Dec 5, team 'instinct.ai' and team 'giraffeccc' are ranked 3 and 4 in the Kaggle competition with an almost identical score on the public leaderboard. One team uses a much simpler model than the other. Which team do you expect to do better on the final result: the one with the simpler model or the one with the more elaborate model. Give a brief justification.

(b) In the famous competition for the Netflix Prize, a number of teams decided to merge before the competition finished. Explain briefly what strategy the teams likely used to combine their algorithms, and why one can expect that this strategy will improve the prediction.

a The simpler model is expected to do better: Overfitting is a bigger problem for the more elaborate model, and since multiple submissions are allowed on the 30% of the test data, we expect that the elaborate model is overfit more than the simpler model, so there will be a bigger deterioration when going to the 70% test data. Since both models did equally well on the 30% part, the bigger deterioration of the elaborate model means the simpler model should perform better on the 30% part.

b They likely averaged their predictions. Averaging reduces variance by the square-root law, so this should improve performance.

9. (3+3 points)
A data scientist divides his dataset into two parts (A and B), trains a variety of methods on part A, evaluates the performance of each method on part B, and finds that degree 2 polynomial regression results in the best part B performance.
(a) The data scientist suspects that the part B error of his selected method is biased, so he uses cross validation on the entire dataset to produce a final estimate of prediction accuracy of degree 2 polynomial regression. Can she trust the cross validation error estimate? Explain.

(b) The data scientist would like to get estimates of the standard error of the regression coefficient of the quadratic term of the degree 2 polynomial. State a method that allows to do this and explain the general idea behind this method (no formulas required).

a No. Since she selected degree 2 polynomial regression using a portion (A) of the data used for CV (A and B), she may have overfit the data set in the selection phase. This introduces a bias in the CV estimate.

b The bootstrap: The SE can be simulated by drawing repeatedly a sample of the same size from the population, fitting the model, and then computing the SD of the resulting coefficient. While we cannot draw from the population, the bootstrap idea is that we can as well resample from the data.

10. (2+2+2+3 points)
We have n observations $y_1, \ldots, y_n$ and p predictors $x_{ij}, 1 \leq i \leq n, 1 \leq j \leq p$.
(a) Write down the expression that one needs minimize in order to get the parameter estimates of the Lasso.

(b) Write down the expression that one needs minimize in order to get the parameter estimates of ridge regression.

(c) Comparing the models that these two methods produce, which is typically more interpretable and why?

(d) Is the prediction error of the Lasso always smaller than that of ridge regression? If yes, why? If no, what is a typical situation where ridge regression will do better?

a $\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$

b $\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$

c Lasso is more interpretable because it shrinks more coefficients to 0, hence results in a simpler model.

d No. Ridge regression will do better when the response is a function of many predictors (with all coefficients of roughly the same size).