Stat 202 **Final exam**

Your name:

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.

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.

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.

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?

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.

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?

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.

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.

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).

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?