STATS 415 – Practice Final Solution

Winter 2018

1. Circle all that apply below.

   (a) Which of the following apply to bagging?

      i. Bagging combines simple base classifiers by upweighting data points which are classified incorrectly.

      ii. Bagging builds different classifiers by training on repeated samples (with replacement) from the training data.

      iii. Bagging can convert a weak binary classifier with a slightly less than 50% error rate into a much better classifier by aggregating many such classifiers trained on different samples.

      iv. Bagging usually outperforms random forests.

      Answer: (ii), (iii)

   (b) Which of the following apply to K-means?

      i. K-means is a supervised learning method.

      ii. K-means is guaranteed to converge under some assumptions on the distribution of the data.

      iii. K-means is guaranteed to converge to the global minimum of some loss function, under some assumptions on the distribution of the data.

      iv. The output of K-means depends on initial values.

      Solution: (ii), (iv)

   (c) Which of the following apply to Principal Component Analysis (PCA)?

      i. PCA is a linear dimension reduction method.

      ii. We must always standardize the data before applying PCA.

      iii. The first principal component corresponds to a direction which maximizes the variance of the data points projected onto that direction.

      iv. Applying PCA to predictors X and then running regression is called Partial Least Squares.

      Solution: (i), (iii)

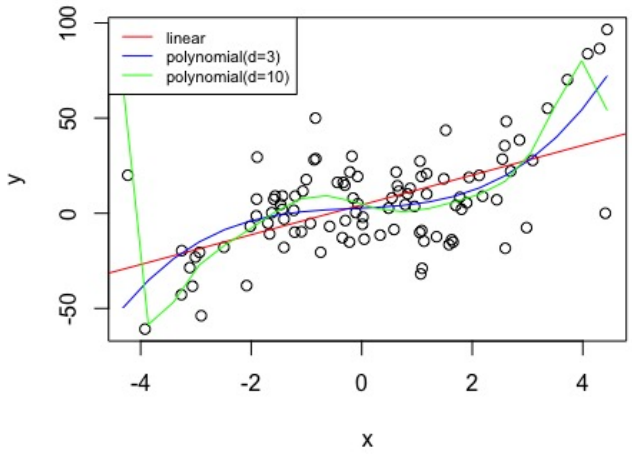
2. Suppose you have regression data generated by a polynomial of degree 3.

   (a) Characterize the bias-variance of the estimates of the following models on the data with respect to the true model by circling the appropriate entry. **Explain your answers**.
   Solution:

	Bias	Variance
Linear regression	high	low
Polynomial regression with degree 3	low	low
Polynomial regression with degree 10	low	high

   Polynomial of degree 3 is the true model so it will have the best bias and variance, both low. The most flexible model of the three, polynomial of degree 10, will follow the data closely, thus change a lot with a new random sample, thus have high variance, and low bias since it is so flexible. The least flexible of the three, linear regression, will behave in the opposite way and have low variance (a line can only follow the data so much) and high bias.

   (b) Sketch a hypothetical scatter plot of the data and the three fits listed above.
   Solution:

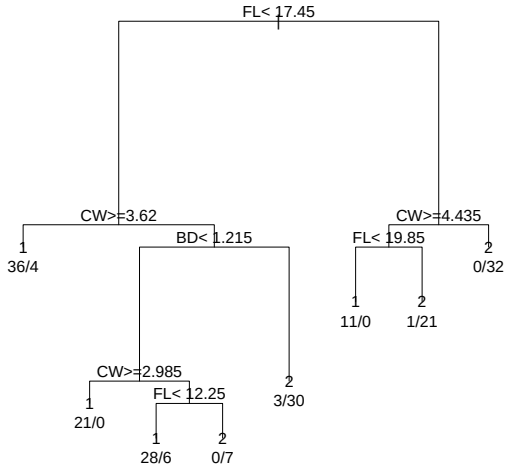


3. The crabs data set describes the morphological measurements of 200 crabs, of the species Leptograpsus variegatus collected in Australia. The variables are: **Species**: 1="blue crabs", 2 = "orange crabs"; **Sex**: 1=male, 2=female; **FL**: frontal lobe size measured in mm; **RW**: rear width in mm; **CL**: carapace length in cm **CW**: carapace width in cm; **BD**: body depth in cm.

   (a) A classification tree was constructed to predict the species of crab from all the other variables, pictured below. Use it to classify the crab with the following features:

```
Species Sex  FL  RW   CL  CW  BD
     ??   1 8.1 6.7 1.61 1.9 0.7
```

Reminder: the tree plot is R is set up so that if you answer "yes" to the split question, you go left, and if "no", you go right.



Solution: 1."blue crabs"
$FL < 17.45$? Yes $\rightarrow$ Left $\Rightarrow$ $CW > 3.62$? No $\rightarrow$ Right $\Rightarrow$ $BD < 1.215$? Yes $\rightarrow$ Left $\Rightarrow$ $CW > 2.985$? NO $\rightarrow$ Right $\Rightarrow$ $FL < 12.25$? Yes $\rightarrow$ Left $\Rightarrow$ 1

(b) Random forests were also run, and the following output was obtained:

	1	2	MeanDecreaseAccuracy	MeanDecreaseGini
Sex	0.00	0.00	0.00	1.33
FL	0.20	0.25	0.22	27.04
RW	0.02	0.06	0.04	10.58
CL	0.09	0.08	0.08	17.08
CW	0.11	0.12	0.11	21.28
BD	0.22	0.13	0.17	22.16

Write down the variables in the order of their importance according to the Gini index. For this dataset, does it matter whether you use classification error or the gini index to measure importance?

Solution: According to the Gini index : FL, BD, CW, CL, RW, Sex.
According to classification error : FL, BD, CW, CL, RW, Sex.
Both measures give the same ordering of the variables in terms of their importance, so it does not really matter what you use.

4. Many classifiers, in addition to a class label, output the probability the given observation has of being from each class. Suppose we have 10 bootstrapped samples used to train a classification tree that outputs the probability of a classification label (red or green in this example). Suppose for a value of X, we have the following 10 estimates of $P(\text{Class is Red}|X)$:

   0.1, 0.15, 0.2, 0.2, 0.55, 0.6, 0.6, 0.65, 0.7, and 0.75.

   (a) The approach we learned in class is to predict the label using each tree (as Red if the probability of Red is over 0.5, and Green otherwise), and take the majority vote. What label does this approach predict for X?

   Solution: The probabilities correspond to label assignments G, G, G, G, R, R, R, R, R, R - 4 Green, 6 Red. Thus the majority vote predicts Red.

   (b) Another common approach to combine these probabilities is to average them and then classify the observation based on the average probability (assigning to class Red if it's over 0.5, and to Green otherwise). What label does this approach predict for X?

   Solution: The average of the 10 probabilities is 0.45 ($<$ 0.5). Thus the average probability approach predicts Green.

   (c) If the labels obtained by these two different approaches agree, would you have more confidence in your prediction than if they disagree? Explain your answer.

   Solution: Yes. If they agree, the probabilities are more likely to be close to 0 and 1, giving you more confidence in predictions.

5. Suppose that for a particular data set, we perform hierarchical clustering using single linkage and complete linkage. We obtain two dendrograms.

   (a) At a certain point on the single linkage dendrogram, the clusters $\{1, 2, 3\}$ and $\{4, 5\}$ merge. On the complete linkage dendrogram, the clusters $\{1, 2, 3\}$ and $\{4, 5\}$ also merge at a certain point. Which merge on will occur higher on the tree, or will they be the same height, or is there not enough information to tell? Explain.

   Solution: Single linkage defines the distance between two sets of points as the distance between two closest points from the different dataset; it is the minimum of individual pairwise distances $\min\{d_{1,4}, d_{1,5}, d_{2,4}, d_{2,5}, d_{3,4}, d_{3,5}\}$, while the complete linkage uses the distances between two points farthest apart, i.e., $\max\{d_{1,4}, d_{1,5}, d_{2,4}, d_{2,5}, d_{3,4}, d_{3,5}\}$. Since the height of the merge is equal to the distance between the clusters, these two sets will fuse at a higher point (on the y-axis) on the complete linkage dendrogram.

   (b) At a certain point on the single linkage dendrogram, the clusters $\{6\}$ and $\{7\}$ merge. On the complete linkage dendrogram, the clusters $\{6\}$ and $\{7\}$ also merge at a certain point. Which merge on will occur higher on the tree, or will they be the same height, or is there not enough information to tell? Explain.

   Solution: Since there is only one point in each of the two sets, the minimum pairwise distance is the same as the maximum; thus the two sets $\{6\}$ and $\{7\}$ will merge at the same height in both the single linkage and complete linkage dendrograms.

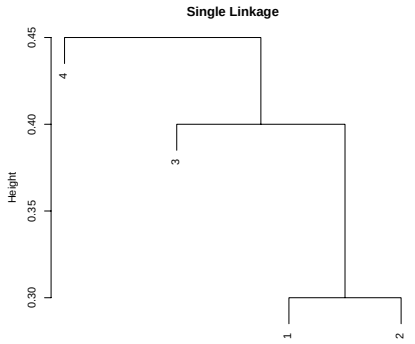
6. Suppose that we have four observations, for which we compute a distance matrix, given by

	P1	P2	P3	P4
P1	0	0.3	0.4	0.7
P2	0.3	0	0.5	0.8
P3	0.4	0.5	0	0.45
P4	0.7	0.8	0.45	0

For instance, the distance between the first and second observations is 0.3, and the distance between the second and fourth observations is 0.8.

(a) On the basis of this distance matrix, sketch the dendrogram that results from hierarchically clustering these four observations using single linkage. Be sure to indicate on the plot the height at which each merge occurs, as well as the observations corresponding to each leaf in the dendrogram. Show all your work.

Answer:



(b) Suppose that we cut the dendrogram obtained in (a) to obtain two clusters. Which observations are in each cluster?

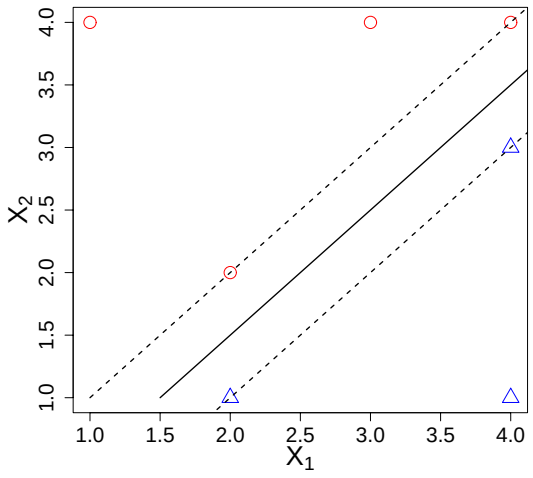
Solution: The two clusters will be $\{4\}$ and $\{1, 2, 3\}$.

7. Here we explore SVM on a toy data set.

   (a) We are given $n = 7$ observations in $p = 2$ dimensions. For each observation, there is an associated class label. Sketch the observations, using different symbols to indicate different classes.

Obs.	X_1	X_2	Y
1	3	4	Red
2	2	2	Red
3	4	4	Red
4	1	4	Red
5	2	1	Blue
6	4	3	Blue
7	4	1	Blue

   Solution: See the plot.



   (b) Sketch the optimal separating hyperplane and write down its equation.

   Solution: See the plot. The solid line is the optimal separating line, and the corresponding equation is $x_2 = x_1 - 1/2$ or $2x_1 - 2x_2 - 1 = 0$.

(c) Describe the classification rule corresponding to the hyperplane. It will have the form "Classify to Red if $\beta_0 + \beta_1 X_1 + \beta_2 X_2 > 0$, and classify to Blue otherwise." (provide specific values of β's).

Solution: Classify to "red" if Classify to "red" if $-2x_1 + 2x_2 - 1 > 0$; thus $\beta_0 = -1$, $\beta_1 = -2$ and $\beta_2 = 2$.

(d) On your sketch, indicate the margin for the classification rule you described. How wide is the margin?

Solution: The two dashed lines indicate the boundaries of the margin, and by basic geometry the width of the margin is $1/\sqrt{2}$.

(e) Indicate the support vectors.

Solution: The support vectors are observations that are on the margin boundaries, i.e., #2, #3, #5 and #6.

(f) Draw an additional observation on the plot so that the two classes are no longer separable by a hyperplane.

Solution: There are infinitely many possibilities, for example, draw a "blue" point at $(2, 3)$