Stats 202 Practice Problems

Stats 202 Course Staff

Fall 2013 and 2014

1. Perform K-means clustering on the following 2-dimensional observations with $K = 3$ and initial labels $(1, 1, 2, 3, 3, 2)$. Use the *Manhattan* distance between a pair of points: $d(A, B) = |X_{1A} - X_{1B}| + |X_{2A} - X_{2B}|$, instead of the Euclidean distance. With this distance, the centroid of a cluster is obtained by taking the median of the samples in each dimension.

Obs.	X_1	X_2
A	1	4
B	1	3
C	3	4
D	5	2
E	3	2
F	3	0

2. Perform single-linkage hierarchical clustering on the data of problem 1, using the Manhattan distance.

3. Every single linkage hierarchical clustering satisfies the following property:

   Let ℓ be the smallest single-linkage distance between any pair of clusters. For any pair of samples i and j in the same cluster, we can find a chain of samples within the same cluster connecting i and j such that the distance between two consecutive samples is at most ℓ.

   Provide a proof for this fact.

4. We fit a linear regression model $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$ to some data. Suppose we change the units of the predictors X_i, to obtain a new set of predictors $Z_i = cX_i$. Then, we fit the same data to the model: $Y = \alpha_0 + \alpha_1 Z_1 + \ldots + \alpha_p Z_p$.

   (a) What is the relationship between the least squares coefficients $\hat{\alpha} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_p)$ and $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)$? Provide a proof.

   (b) What is the relationship between the fitted values in the two models?

5. Your colleague fitted a multivariate linear regression model $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$ and found all but three p-values are significant in the t-test. He decides to drop those three variables and keep all the remaining predictors. What do you think of your colleague's method?

6. Explain the purpose of an F-test for multiple linear regression.

7. True or false: The variance of a regression estimator $\hat{f}$ in the bias-variance decomposition can be written:

$$\frac{1}{n-1}\sum_{i=1}^{n}(\hat{f}(x_i) - m)^2,$$

where

$$m = \frac{1}{n}\sum_{i=1}^{n}\hat{f}(x_i),$$

and $x_1, \ldots, x_n$ are the inputs of the training data.

8. Suppose the data $(\mathbf{X}, \mathbf{y})$ are well fit by a linear model, how would you diagnose if the data point (x_i, y_i) is an outlier or a high leverage point?

9. Suppose we have a dataset with N observations, and each observation consists of three values:

   - y: binary variable that is 1 if a student passed and 0 if a student failed the exam
   - x_1: the number of hours spent studying for the exam
   - x_2: a binary variable indicating whether or not the student passed the previous exam.

   Suppose upon fitting a logistic regression of the y on x_1, x_2, and an intercept, the estimates for $\beta = (\beta_0, \beta_1, \beta_2)$ are

$$\begin{aligned}\hat{\beta}_0 &= -1.2\\ \hat{\beta}_1 &= 0.3\\ \hat{\beta}_2 &= 1.2\end{aligned}$$

   Now suppose instead of using the number of hours spent studying, we used the number of minutes spent using for the exam. Can you identity what the new β_0, β_1, and β_2 would be? Why or why not? Justify your claim.

10. Suppose we have a classification problem with a binary response Y and a p-dimensional predictor variable $X = (X_1, \ldots, X_p)$. Logistic regression is fitted to a set of n samples. Then, logistic regression is fitted again to the same observations, where we include one additional predictor, such that:

$$X = (X_1, \ldots, X_p, X_{p+1}).$$

   Explain how the training error, test error, and coefficients change in each of the following cases:

   (a) $X_{p+1} = X_1 + 2X_p$.

   (b) X_{p+1} is a random variable independent of Y.

11. A scientist performs a ridge regression in R and estimates the optimal parameter λ by 10-fold cross-validation. She uses the following code:

```
> library(glmnet)
> X = as.matrix(data[,-1])
> Y = as.vector(data[,1])
> cv.out = cv.glmnet(X,Y,alpha = 0)
> cv.out$lambda.min
[1] 0.5
> min(cv.out$cvm)
[1] 485.1199
```

The scientist concludes that the test MSE of ridge regression with $\lambda = 0.5$ is approximately 485.119. Explain the problem with this estimate and suggest a better way to estimate the test MSE.

12. On the make believe island of Statlantis, there's a volcano which periodically erupts. It's known the times between eruptions are random variables $X_1, X_2, \ldots,$ which are independent and uniformly distributed in the interval $[0, \theta]$, but θ is unknown. Suppose we observe n of the times between eruptions $x_1, x_2, \ldots, x_n$, and we wish estimate θ using the estimator

$$\hat{\theta} = \frac{2}{n} \sum_{i=1}^{n} x_i,$$

that is, the average of the observations multiplied by 2.

(a) We can estimate the standard error of $\hat{\theta}$ using the bootstrap. In this example, it is possible to compute the boostrap error analytically (the error we would estimate from an infinite number of bootstrap resamplings). Find an expression for this standard error as a function of the sample variance $\hat{\sigma}^2$ of the observed times $x_1, \ldots, x_n$.

(b) Find the standard error of $\hat{\theta}$ analytically. Express your answer as a function of the variance of the time between eruptions σ^2.

(c) Argue that the bootstrap standard error will be similar to the true standard error when n is large.

Hint: The variance of a sum of independent random variables is the sum of their variances.

13. What are the key differences between LOOCV and k-fold cross validation for estimating the true test error? Is one better than the other, and if so, why?

14. Suppose we are trying to model a classification problem with two labels: 'sick' and 'healthy.' For the purpose of this test, we consider a positive result to be testing sick and a negative to test as healthy. After fitting the model with LDA in R, we compare predicted values from the actual values, as shown below:

```
> lda.fit = lda(test.result ~ x1 + x2, data=mydata)
> lda.pred = predict(lda.fit)
> lda.class=lda.pred$class
> table(lda.class, mydata$test.result)
          mydata
lda.pred sick healthy
sick     40   32
healthy  25   121
```

What is the misclassification rate for the model above? How can we decrease the rate of false positives to false negatives in LDA? Why might we want to do that and, assuming LDA is a good model for the data, how is this likely to affect the misclassification rate?

15. Consider the dataset

x	y
-2	‘slow’
5	‘fast’
-1	‘slow’
10	‘fast’
4	‘fast’

Suppose we used logistic regression to fit this model: that is, if y is a binary variable that is either ‘fast’ or ‘slow’, we wish to fit the model

$$\mathbb{P}(y_i = \text{fast}) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_i)}} \qquad \mathbb{P}(y_i = \text{slow}) = \frac{e^{-(\beta_0+\beta_1 x_i)}}{1+e^{-(\beta_0+\beta_1 x_i)}}$$

for all $i = 1, \dots, 5$. What value(s) of β would maximize the likelihood (and thus be the estimates returned from fitting this model)?

16. Suppose that $\mathbf{X}$ is an $n \times p$ matrix of predictors and $\mathbf{y}$ is a quantitative response. Suppose that $p > n \geq 1000$. You want to fit a linear model that helps you make predictions with new data. Explain which of the following methods could be applied, and the advantages of each one.

    (a) Least squares
    (b) Lasso
    (c) Ridge regression
    (d) Backward stepwise selection
    (e) Forward stepwise selection
    (f) Best subset selection

17. Consider selecting subsets of predictors in the standard linear model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

We run best subset selection, forward stepwise selection, and backward stepwise selection using RSS as the criterion and get the list of models:

$$\begin{aligned}
&\text{Best subset selection}: \mathcal{M}_0^{(bs)}, \mathcal{M}_1^{(bs)}, \cdots \mathcal{M}_p^{(bs)}\\
&\text{Forward stepwise selection}: \mathcal{M}_0^{(forward)}, \mathcal{M}_1^{(forward)}, \cdots, \mathcal{M}_p^{(forward)}\\
&\text{Backward stepwise selection}: \mathcal{M}_p^{(backward)}, \mathcal{M}_{p-1}^{(backward)}, \cdots, \mathcal{M}_0^{(backward)}
\end{aligned}$$

For example, $\mathcal{M}_k^{(bs)}$ is the 'best' model among all the models with k predictors and $\mathcal{M}_k^{(forward)}$ is the 'best' model among the $p-k$ candidate models after choosing $\mathcal{M}_{k-1}^{(forward)}$.
Let $RSS_{\mathcal{M}}$ the training error fitted by model $\mathcal{M}$. Show that

(a) $RSS_{\mathcal{M}_p^{(forward)}} = RSS_{\mathcal{M}_p^{(backward)}}$.

(b) $RSS_{\mathcal{M}_1^{(forward)}} \leq RSS_{\mathcal{M}_1^{(backward)}}$.

18. Ridge regression solves the following optimization problem:

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

Suppose we fit a ridge regression model for a single predictor X_1 and intercept with parameter $\lambda = \delta > 0$ and get $\hat{\beta}_0 = a$ and $\hat{\beta}_1 = b$. Now we include additional predictor X_2 which satisfies $X_2 = X_1$ and fit a ridge regression model with parameter 2δ. Express the fitted model in terms of a and b and compare with the previous fitted model.

Hint: $2(\beta_1^2 + \beta_2^2) \geq (\beta_1 + \beta_2)^2$ with equality only when $\beta_1 = \beta_2$.

19. In this problem, we compare the lasso estimator with the best subset estimator. Lasso regression solves the following optimization problem:

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

Let $\hat{\beta}^{lasso}$ the solution of this problem with s nonzero $\hat{\beta}_j^{lasso}$s. i.e., $\sum_{j=1}^{p} 1(\hat{\beta}_j^{lasso} \neq 0) = s$. Consider best subset method of size s which solves the following optimization problem:

$$\min_\beta \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2$$

$$\text{subject to } \sum_{j=1}^{p} 1(\beta_j \neq 0) = s$$

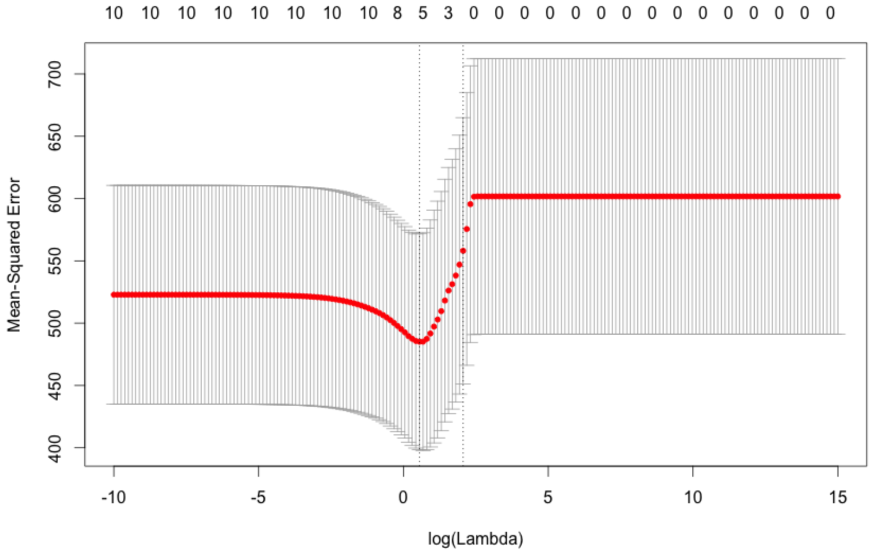
Let $\hat{\beta}^{bs}$ the solution of this problem. Prove that

$$\sum_{j=1}^{p} |\hat{\beta}_j^{lasso}| \leq \sum_{j=1}^{p} |\hat{\beta}_j^{bs}|$$

Can we say that the lasso estimator is a shrinkage estimator of the best subset estimator?

20. When the number of observations and predictors are large, best subset selection can be computationally intensive. However, if we ignore this issue, best subset selection might be preferred to forward stepwise selection because every model that appears in forward stepwise selection algorithm is considered in best subset selection. Explain the reason that forward selection can still be preferred.

21. The plot below displays the cross-validation errors of lasso regression computed on a range of tuning parameter λ. There are $p = 10$ parameters in the model, and the training set and test set has $n = 100$ observations each.



Cross-validation error plot for lasso regression

```
> cv.out$cvm[1]
[1] 601.7324
> cv.out$cvm[200]
[1] 522.8358
> min(cv.out$cvm)
[1] 485.1199
```

What are the two models marked with dashed vertical lines in the plot? Would you expect an appropriate shrinkage method to give a smaller test MSE than the test MSE of multiple linear regression for this data?

22. What is the main advantage of the Lasso with respect to Ridge regression?

23. Assume 2 predictors and a quantitative outcome have a linear relationship $y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \varepsilon_i$, where ε_i is i.i.d. noise of unit variance. You have a training set $(\mathbf{X}, \mathbf{y})$, where $\mathbf{X}$ has full column rank and

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} d_1 & 0 \\ 0 & d_2 \end{bmatrix}.$$

Let $\hat{y}_\lambda$ be the prediction of ridge regression at a test point with predictors (x_{01}, x_{02}). It can be shown that

$$\begin{aligned} \text{Bias}(\hat{y}_\lambda) &= \lambda\left(\frac{\beta_1 x_{01}}{d_1+\lambda} + \frac{\beta_2 x_{02}}{d_2+\lambda}\right) \\ \text{Var}(\hat{y}_\lambda) &= \left[\frac{d_1 x_{01}^2}{(d_1+\lambda)^2} + \frac{d_2 x_{02}^2}{(d_2+\lambda)^2}\right]. \end{aligned}$$

Prove that ridge regression achieves a strictly lower test MSE at (x_{01}, x_{02}) than least squares regression for a positive value of λ.

24. A new startup has developed a cheap way to measure the expression of antibodies in infants. They hope to use these data to classify infants according to whether they will develop atopic syndrome. The company has collected data from 500 infants and followed them for 5 years to evaluate the incidence of atopy. They ran an initial trial in which they measured just 1,000 antibodies, and then a second trial which considered a more complete library of 100,000 antibodies. In each case, they performed variable selection using forward stepwise selection and produced a classification by logistic regression.

    They found that the classification quality was worse in the second trial. Puzzled by this outcome, they decided to consult with you. How would you explain this observation?

25. Variable Y is generated by the following formula:

$$Y = 1 + X + X^2 + \epsilon$$

where $X, \epsilon \sim N(0,1)$ and X, ϵ are independent. 100 samples are generated.

```
set.seed(1)
x=rnorm(100)
e=rnorm(100)
y = 1+ x + x^2 + e
```

We apply Principal Component Regression and Partial Least Squares using predictors X, X^2, X^3, X^4 in order to predict Y variable.

```
> pcr.fit<-pcr(y~.,data=x_data,scale=TRUE,validation="CV")
> summary(pcr.fit)
Data:  X dimension: 100 4
Y dimension: 100 1
Fit method: svdpc
Number of components considered: 4

VALIDATION: RMSEP
Cross-validated using 10 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps
CV           1.733    1.226    1.018   0.9560   0.9780
adjCV        1.733    1.153    1.016   0.9544   0.9754

TRAINING: % variance explained
   1 comps  2 comps  3 comps  4 comps
X    56.37    93.90    98.44    100.0
y    66.77    66.88    70.29     70.3
> pls.fit<-plsr(y~.,data=x_data,scale=TRUE,validation="CV")
> summary(pls.fit)
Data:  X dimension: 100 4
Y dimension: 100 1
Fit method: kernelpls
Number of components considered: 4

VALIDATION: RMSEP
Cross-validated using 10 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps
CV           1.733    1.003   0.9750   0.9686   0.9868
adjCV        1.733    1.002   0.9718   0.9669   0.9838
```

```
TRAINING: % variance explained
   1 comps  2 comps  3 comps  4 comps
X    56.34    68.63    98.43    100.0
y    67.40    69.77    70.30     70.3
```

We observe that when the same number of components are used, the percentage of variance explained in the predictors in PCR are larger or equal than that in PLS, but the percentage of variance explained in the response in PCR are always smaller or equal than that in PLS. Briefly explain the reasons.

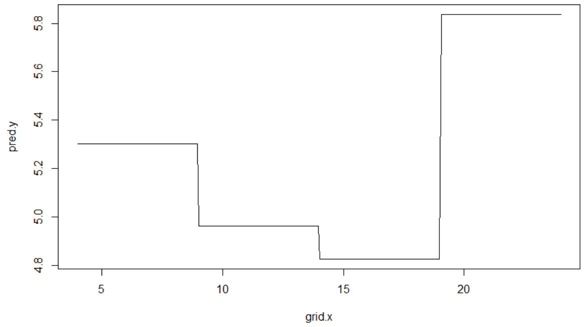
26. We fit a step function regression on a dataset with a single predictor X. 3 cutpoints c_1, c_2, c_3 in the range of X are selected. Construct 4 variables

$$C_0(X) = I(X < c_1),\ C_1(X) = I(c_1 \leq X < c_2),\ C_2(X) = I(c_2 \leq X < c_3), C_3(X) = I(c_3 \leq X)$$

where $I(\cdot)$ is an indicator funtion. Remark that the linear model using $C_0(X), \cdots, C_3(X)$ as predictors is:

$$Y = \beta_0 + \beta_1 C_1(X) + \beta_2 C_2(X) + \beta_3 C_3(X) + \epsilon$$

The plot below displays the fitted step function on a dataset.



Now we fit a lasso regression using the same predictors. Using the tuning parameter $\lambda = 0.17$, we get 3 nonzero β_is out of 4; $\hat{\beta}_1 = 0$. Sketch the lasso-fitted step function on the plot above.

27. Consider two curves, $\hat{g}_1$ and $\hat{g}_2$, defined by

$$\hat{g}_1 = \arg\min_g \left(\sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int [g^{(3)}(x)]^2 dx \right)$$

$$\hat{g}_2 = \arg\min_g \left(\sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int [g^{(4)}(x)]^2 dx \right)$$

where $g^{(m)}$ represents the mth derivative of g. For each of the following questions, either provide an answer or state that there is not enough information to answer.

(a) As $\lambda \to \infty$, which of $\hat{g}_1$ or $\hat{g}_2$ has the smaller training RSS?

(b) As we increase λ, how do the bias and the variance of $\hat{g}_1$ change?

(c) As $\lambda \to \infty$, which of $\hat{g}_1$ or $\hat{g}_2$ has the smaller test RSS?

28. We can estimate a natural cubic spline with K knots by fixing the locations of the knots at quantiles of X and finding the coefficients $\beta_0, \ldots, \beta_{K+1}$ which minimize the RSS. True or false: Smoothing splines are obtained in the same way by setting K to the number of samples n. Explain your answer.

29. A cubic spline with K knots is a function that:

    - is a cubic polynomial between each pair of knots,
    - is continuous at the knots, and
    - has continuous first and second derivatives at each knot.

    Without using the definition of a cubic spline in terms of basis functions, explain why the spline has $K + 4$ free parameters or degrees of freedom.

30. (Locally weighted regression) Often a linear regression is inadequate, and we turn to more flexible forms of regression. One such form is locally weighted regression. Consider the case of a single predictor x, and suppose the model is $y = f(x) + \epsilon$, with $\epsilon \sim N(0, \sigma^2)$. Local weighted regression approximates $f(x)$ at x_0 by $\hat{f}(x_0) = a_0 + b_0 x_0$, where a_0 and b_0 depend on x_0 and are chosen to minimize

$$\sum_{i=1}^{n} W\left(\frac{|x_i - x_0|}{\omega}\right)(y_i - a_0 - b_0 x_i)^2.$$

$W(r)$ is a positive weight function. Two options for weight functions are, the uniform weight function:

$$W_u(r) = \begin{cases} 1 & \text{if } |r| \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

and the tri-cube weight function, invented by John Tukey:

$$W_{tc}(r) = \begin{cases} (1 - |r|^3)^3 & \text{if } |r| \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

The parameter ω is known as the *window size* or bandwidth. Since x_0 is arbitrary, we have an approximation $\hat{f}$ at any value of x.

(a) Explain in words how this method works and why it makes sense.

(b) Explain the role of ω, and describe a principled way of selecting ω.

(c) What would be the main difference in the function $\hat{f}$ when we use the uniform weight function vs. the tri-cube weight function?

31. Consider a natural cubic spline and a polynomial regression with the same degrees of freedom on the same data set. Which is likely to be more stable for extreme values of the predictor?

32. List 2 regression methods for which it is possible to compute the leave one out cross validation (LOOCV) error analytically without performing n fits.

33. In class, we studied local linear regression, which solves a separate least squares problem at each target point x_0:

$$\min_{\alpha(x_0),\beta(x_0)} \sum_{i=1}^{N} K(x_0, x_i)[y_i - \alpha(x_0) - \beta(x_0)x_i]^2.$$

The estimate is then $\hat{f}(x_0) = \hat{\alpha}(x_0) + \hat{\beta}(x_0)x_0$.

Following this line, we can obtain local quadratic fits by solving the following least squares problems:

$$\min_{\alpha(x_0),\beta(x_0),j=1,\ldots,d} \sum_{i=1}^{N} K(x_0, x_i) \left[y_i - \alpha(x_0) - \beta_1(x_0)x_i - \beta_2(x_0)x_i^2\right]^2$$

with solution $\hat{f}(x_0) = \hat{\alpha}(x_0) + \hat{\beta}_1(x_0)x_0 + \hat{\beta}_2(x_0)x_0^2$.

Please list pros and cons of these two methods.

34. One-dimensional spline models can be extended to multidimensional inputs. Suppose we have two predictors X_1 and X_2, and the functions $h_k,\ k = 1, \ldots, M$ form a basis for a cubic spline with fixed knots $\xi_1, \ldots, \xi_{M-4}$. We can define two dimensional basis functions by the tensor product

$$g_{jk}(X) = h_j(X_1)h_k(X_2),\ j, k = 1, \ldots, M,$$

and a two-dimensional function can be modeled as a linear combination of these:

$$g(X) = \sum_{j_1=1}^{M} \sum_{j_2=1}^{M} \theta_{jk} g_{jk}(X).$$

How does this model compare in terms of bias and variance with a GAM which represents $g(X)$ as a sum of a cubic spline in X_1 and a cubic spline in X_2 with knots at $\xi_1, \ldots, \xi_{M-4}$ in each case?

35. Imagine starting to grow a 2-class classification tree. We have 80 points total, with 3 possible values of x. In class 1, 30 of the points are at $x = 0$, 10 points are at $x = 1$, and no points are at $x = 2$. In class 2, 10 of the points are at $x = 0$, 10 points are at $x = 1$, and 20 points are at $x = 2$. There are two potential splits on x: split between $x = 0$ and $x = 1$, or split between $x = 1$ and $x = 2$.

Compute the the misclassification error and the Gini index for the two splits. Which criterion produces a pure region?

36. Suppose you are fitting a regression tree and have a categorical variable with K classes. When considering splits on this variable, do you have to consider each of the $2^K - 2$ possible splits? Explain.

37. Describe the procedure for computing the out of bag error from Bagging a regression tree.

38. How does Bagging a regression tree differ from a Random Forest? Why might Random Forests be preferable?

39. We build a classification tree using the predictors $X_1, X_2, \ldots, X_p$. We build a second tree using the predictors $f(X_1), f(X_2), \ldots, f(X_p)$, where f is monotone:

$$f(x) \geq f(y) \text{ if and only if } x \geq y.$$

Prove that the two trees produce the same partition of the training data.

40. The standard method for fitting a decision tree involves:

    - Growing the tree split by split. We maximize the reduction of the training error at each step until there are at most 5 samples per region.
    - Pruning the tree to obtain a sequence of trees of decreasing size.
    - Selecting the optimal size by cross-validation.

    Consider the following alternative approach. Grow the tree split by split until the reduction in the training error produced by the next split is smaller than some threshold. This approach may lead to bad results because it is possible to make a split which does not decrease the error by much, and then make a second split which reduces the error significantly.

    Draw an example dataset where this happens with 2 predictors X_1 and X_2, and a binary categorical response.

41. Describe what the following code does. The `shrinkage` parameter in `gbm` controls what was called λ or the learning rate in class, `n.trees` controls the number of iterations to perform (called B in class), and `interaction.depth` controls the depth of the trees used (called d in class).

```
library(gbm)
library(MASS)
data(Boston)
lambda.grid = seq(.001, .2, length.out = 100)
test.error = rep(0, length(lambda.grid))
train.idx = sample(nrow(Boston), round(nrow(Boston) * .7))
for(i in 1:length(lambda.grid)) {
   lambda = lambda.grid[i]
   boost = gbm(medv ~ ., data = Boston[train.idx,],
               distribution = "gaussian", n.trees = 5000,
               interaction.depth = 1, shrinkage = lambda)
   yhat.boost = predict(boost, newdata = Boston[-train.idx,],
                        n.trees = 5000)
   test.error[i] = mean((yhat.boost - Boston$medv[-train.idx])^2)
}
```

42. When we apply Bagging to decision trees, or when we construct Random Forests, each tree fit to a different bootstrap replicate of the data can be grown “deep”, i.e. to a level where there are very few training samples per leaf. The pruning step is skipped. Explain why this doesn’t lead to overfitting.

43. We apply Bagging to a 1-nearest neighbor regression problem. To make a prediction at x_0, we average the predictions of 1-nearest neighbor regression fit to M Bootstrap replicates of the training data. Prove that the bagging prediction for M large is equivalent to a weighted average of k-nearest neighbors regression predictions with different values of k.

    Hint: Suppose we take n samples with replacement from the set $\{1, 2, \ldots, n\}$. Let Z be the smallest of the samples. Then,

$$\mathbb{P}(Z = i) = \left(1 - \frac{i-1}{n}\right)^n - \left(1 - \frac{i}{n}\right)^n,$$

    which is monotone decreasing in i.

44. Explain why the maximal margin classifier cannot be applied in the case of non-separable data.

45. Consider the maximal margin classifier:

$$\begin{array}{ll} \text{maximize} & M \\ \text{subject to} & \sum_{j=1}^{p} \beta_j^2 = 1 \\ & y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}) \geq M \ \forall \ i = 1, ..., n \end{array}$$

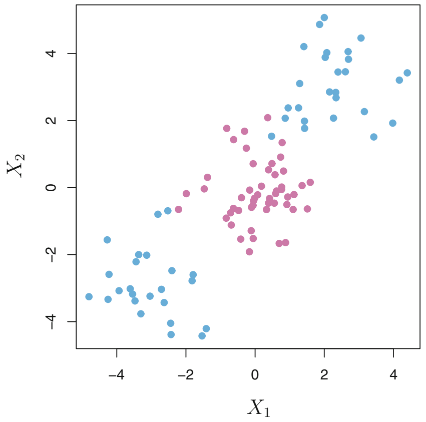
    and the support vector classifier:

$$\begin{array}{ll} \text{maximize} & M \\ \text{subject to} & \sum_{j=1}^{p} \beta_j^2 = 1 \\ & y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}) \geq M(1 - \epsilon_i) \ \forall \ i = 1, ..., n \\ & \epsilon_i \geq 0, \ \ \sum_{i=1}^{n} \epsilon_i \leq C \end{array}$$

    Note that the support vector classifier introduces a tuning parameter C, and if C is chosen to be zero, then the support vector classifier is equivalent to the maximal margin classifier.

    Suppose that a data set is separable, so that the maximal margin classifier is applicable. Describe the relationship between the choice of C and the robustness of the support vector classifier to individual observations.

46. The figure below, taken from *Introduction to Statistical Learning*, depicts a data set of two classes, represented by the colors blue and purple. Explain why the support vector machine would be much preferable to the support vector classifier in this instance.



47. Describe two methods for extending support vector machines from two classes to an arbitrary number (say, K) of classes.

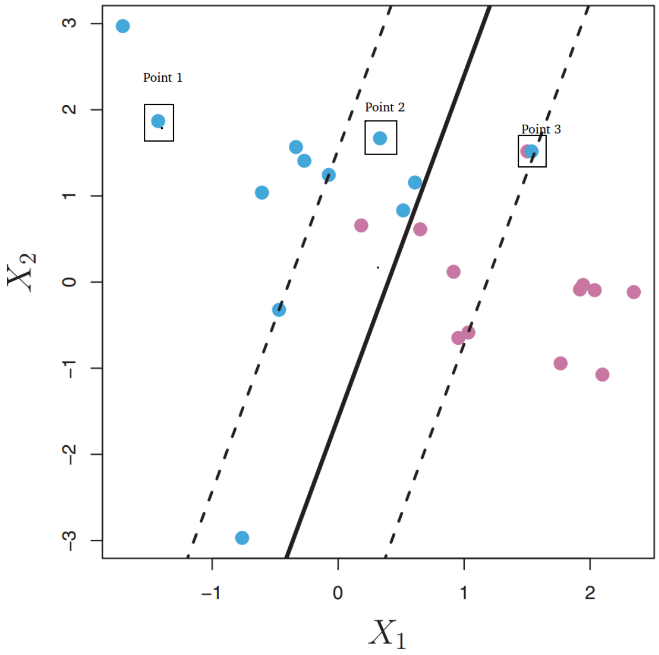
48. Recall an equivalent formulation of the support vector **classifier**:

$$\text{minimize} \left\{ \sum_{i=1}^{n} \max[0, 1 - y_i(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip})] + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

where λ is some tuning parameter. Describe the relationship between λ and the tuning parameter C from the definition of the support vector classifier above. What does this relationship imply about the choice of C as it pertains to bias-variance tradeoff?

49. The plot below, from ISLR, shows the separating hyperplane (solid line) and margin (dotted lines) resulting from fitting the SVM optimization

$$\begin{aligned}
&\text{maximize } M && (1)\\
&\text{over } \beta, \epsilon \text{ such that,} && (2)\\
&\beta_0 + \beta_1^2 + \beta_2^2 = 1 && (3)\\
&y_i\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}\right) \geq M\left(1 - \epsilon_i\right) && (4)\\
&\epsilon_i \geq 0 && (5)\\
&\sum_{i=1}^{n} \epsilon_i \leq C. && (6)
\end{aligned}$$

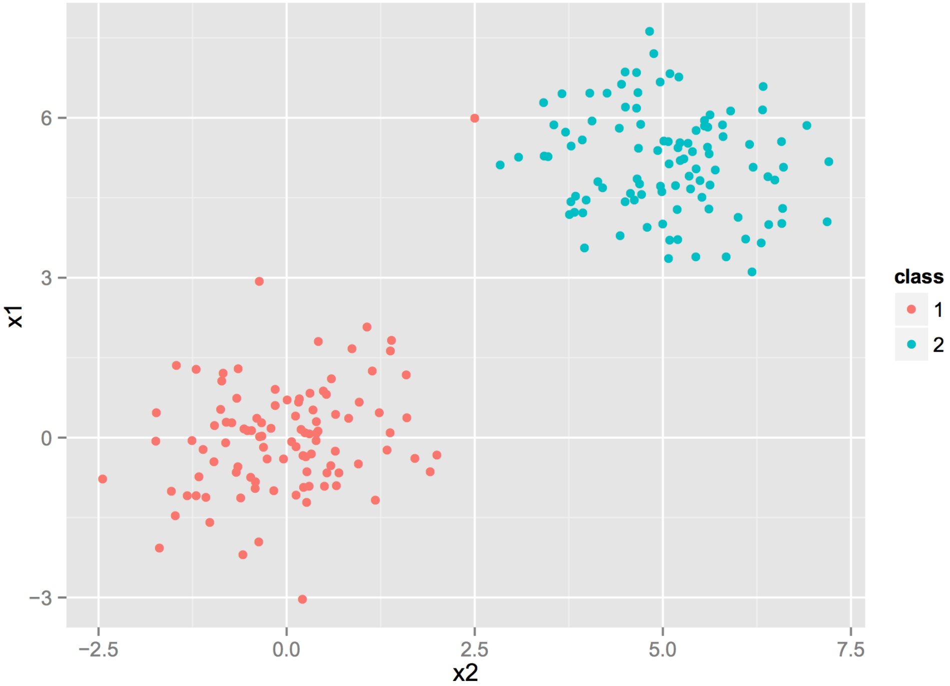


This problem is about the interpretation of these parameters.

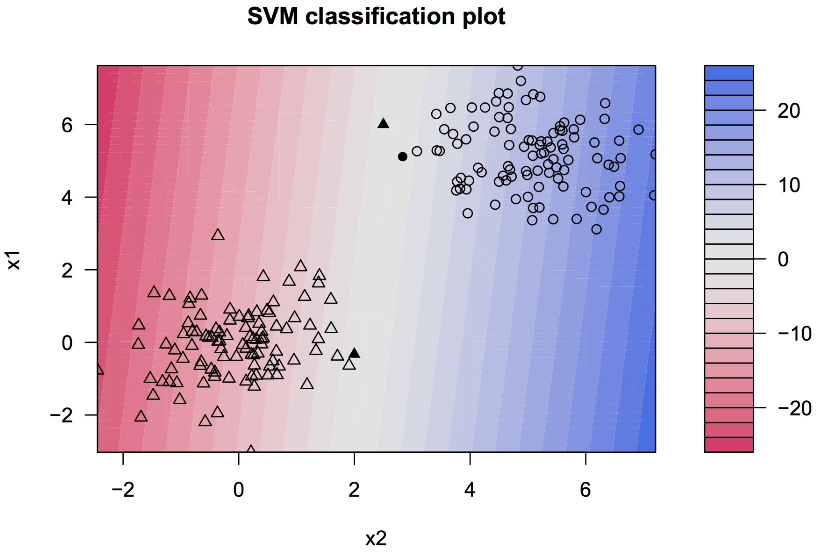
(a) Suppose we decrease C. What happens to the separating hyperplane? What happens to the margin?

(b) For each of the boxed blue points, specify whether $\epsilon_i = 0, \epsilon_i \in (0, 1]$, or $\epsilon_i > 1$.

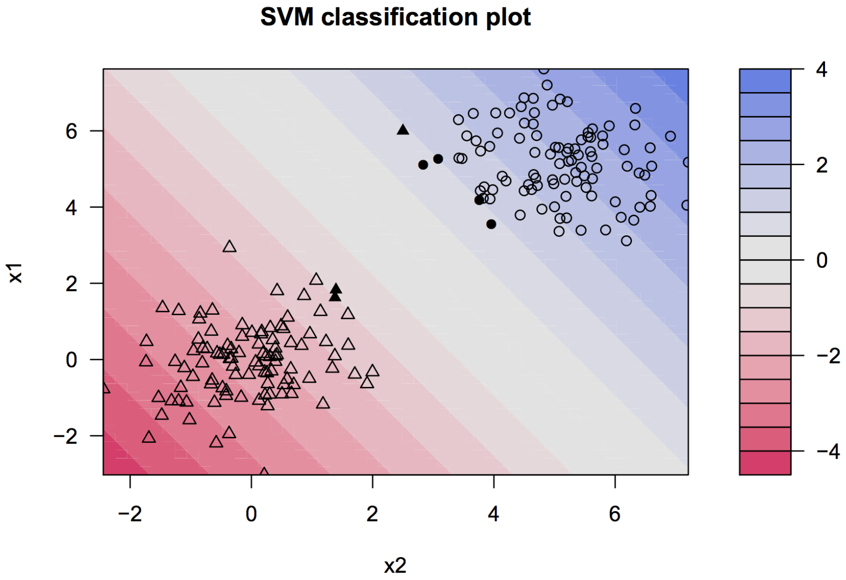
50. You have collected a data set with two clearly distinguishable classes, but with one outlier, as plotted below.



You decide to build an linear-kernel SVM to distinguish between these two classes. You tinker with different parameters, and plot the resulting separating hyperplanes. For this problem, consider the following two fits, resulting from model 1,



and model 2,



The gradient indicates the distance from the hyperplane, which is positive in the direction of the circles and negative in the direction of the triangles. The filled points are the support vectors.

(a) You forgot which model has a larger value of C, the constraint on the sum of the slack variables in the usual SVM optimization.

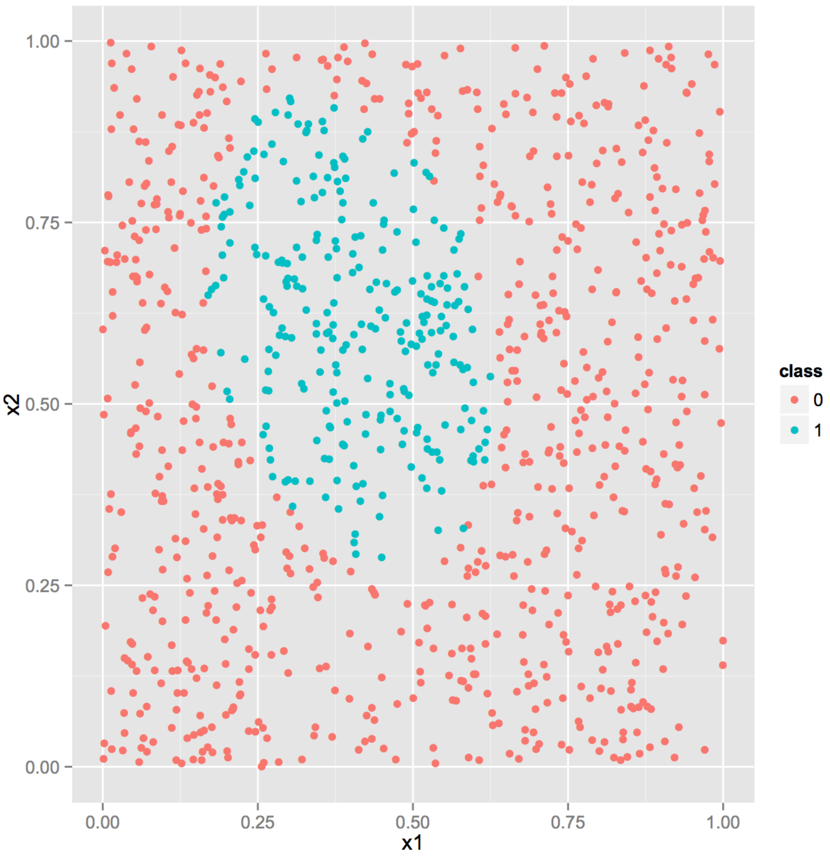
$$\begin{aligned} &\text{maximize } M \\ &\quad \text{over } \beta, \epsilon \text{ such that} \\ \beta_0 + \beta_1^2 + \beta_2^2 &= 1 \\ y_i \left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}\right) &\geq M\left(1 - \epsilon_i\right) \\ \epsilon_i &\geq 0 \\ \sum_{i=1}^{n} \epsilon_i &\leq C. \end{aligned}$$

Can you deduce this from the given plots?

(b) Which should you use?

(c) In general, state if increasing the parameter C increases or decreases the following quantities: size of margin, number of discovered support vectors, variance of the fitted model.

51. You have a data set where one class falls in the interior of an ellipse, and the other falls outside, as in the plot below.



You would like to build an SVM to distinguish between the two classes.

(a) Recall the linear kernel,

$$K\left(x_i, x_{i'}\right) = \sum_{j=1}^{2} x_{ij} x_{i'j}, \tag{7}$$

the $d-$degree polynomial kernel,

$$K\left(x_i, x_{i'}\right) = \left(1 + \sum_{j=1}^{2} x_{ij} x_{i'j}\right)^d, \tag{8}$$

and the radial kernel,

$$K\left(x_i, x_{i'}\right) = \exp\left(-\gamma \sum_{j=1}^{2} \left(x_{ij} - x_{i'j}\right)^2\right). \tag{9}$$

Which of these kernels is most appropriate for your problem?

Hint: In two dimensions, an ellipse centered at $x_0 = \begin{pmatrix} x_{01} \\ x_{02} \end{pmatrix}$ with rotation $K = \begin{pmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{pmatrix}$ is the set of vectors $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ satisfying

$$\left(x_1 - x_{01}\right)^2 K_{11} + \left(x_2 - x_{01}\right)^2 K_{22} + 2\left(x_2 - x_{02}\right)\left(x_1 - x_{01}\right) \leq C, \tag{10}$$

for some C.

(b) For each of the polynomial and radial kernels above, explain the impact of increasing d and γ, respectively, on the bias and variance of the fit.