Optional practice problems

Stat 151A, Fall 2017

November 21, 2017

1. For the following statements determine whether they are true of false. In each case provide a reason behind your choice.

   (a) Let the model m have fitted coefficients $\hat{\boldsymbol{\beta}}$ from fitting the model to the full dataset. To perform K-fold cross validation, you would divide the data into K parts, use $\hat{\boldsymbol{\beta}}$ to get $\hat{\boldsymbol{y}}$, and then calculation the prediction error is

   $$\sum_i (y_i - \hat{y}_i)^2$$

   Wrong, we wanna calculate beta_hat excluding ith fold and then calculate predition error on ith fold.

   .

   (b) In a logistic regression model that has an intercept, the null deviance is always greater than or equal than the residual deviance. yes

   (c) Null deviance can be used to compute AIC and BIC for a general logistic regression model. no

2. From the book 14.2, 14.3, 14.4, 14.9, 15.6, 15.7.

3. Show that adjusted $\tilde{R}^2$ is less than or equal to one.

4. Let $Y_1, Y_2, \cdots, Y_n$ be independent Poisson(μ_i), where

   $$\log \mu_i = \beta_1 + \beta_2 X_i.$$

   Set up the generalized linear model problem for estimation of β_1 and β_2. What is the link function? Describe how would you find the maximum likelihood estimates of β_1 and β_2.

5. Let $Y_1, Y_2, \ldots, Y_n$ be independent Bernoulli(π_i), where

   $$\pi_i = \Phi(\beta_1 + \beta_2 X_i)$$

   and Φ is the cdf of the standard normal distribution.

   (a) Set up the generalized linear model problem for estimation of β_1 and β_2. What is the link function g?

(b) Describe how you would find the mle estimates of β_1 and β_2.

6. With the notation from Handout "Regression and Classification Trees", show that for the first split, the quantity $RSS(j, c)$ is always smaller than or equal to TSS for all j and c.

7. Recall, for ridge regression, we seek to find $\tilde{\beta}$ that, for some fixed λ, minimizes

$$||Y - X\beta||^2 + \lambda||\beta||^2$$

which turns out to be

$$\tilde{\beta} = (X'X + \lambda I)^{-1}X'Y$$

(a) In the constant predictor case where $Y \sim N(\beta, \sigma^2)$, state the design matrix X and show that this reduces to

$$\tilde{\beta} = \frac{n}{n+\lambda}\bar{Y}$$

(b) Find $E(\tilde{\beta})$ and $var(\tilde{\beta})$.

(c) Verify that $MSE(\tilde{\beta}) = (1-\alpha)^2\beta^2 + \alpha^2\sigma^2/n$, where $\alpha = n/(n+\lambda)$, and that it is smaller than that of the OLS estimator $\bar{Y}$ if and only if $\frac{\beta^2}{\sigma^2} < \frac{1+\alpha}{n(1-\alpha)}$.

8. Consider the dataset *titanic* which consists of 891 passengers who were aboard titanic. The response variable is *Survived* which takes the value 1 if the passenger survived and 0 otherwise. Consider fitting a classification tree to this dataset with the response variable being *Survived* and the explanatory variables being *Pclass* (this is a proxy for the class in which the passenger travelled; has three levels 1, 2 and 3), *Sex* (gender), *SibSp* (number of siblings/spouses aboard), *Parch* (number of Parents/Children aboard), *Fare* (ticket fare) and *Embarked* (port of embarkation; has three levels: C for Cherbourg, Q for Queenstown and S for Southampton).

(a) I first fit a classification tree to this dataset for the response variable *Survived* using only *Fare* the explanatory variable. This gave me the following tree.

```
> rt1 = rpart(Survived ~ Fare, method = "class", data = titanic)
> rt1
n= 891

node), split, n, loss, yval, (yprob)
* denotes terminal node

1) root 891 342 0 (0.6161616 0.3838384)
2) Fare< 10.48125 339  67 0 (0.8023599 0.1976401) *
3) Fare>=10.48125 552 275 0 (0.5018116 0.4981884)
6) Fare< 74.375 455 XXX XX (0.5582418 0.4417582)
```

```
12) Fare>=69.425 15   2 0 (0.8666667 0.1333333) *
13) Fare< 69.425 440 199 0 (0.5477273 0.4522727)
26) Fare< 52.2771 403 171 0 (XXXXXXXXX XXXXXXXX) *
27) Fare>=52.2771 37   9 1 (0.2432432 0.7567568) *
7) Fare>=74.375 XX  23 1 (0.2371134 0.7628866) *
```

Fill the five missing values in the R output above with proper reasoning.

(b) I tried to plot this regression tree via

```
plot(rt1)
```

and this resulted in the plot in Figure 1. Label this plot manually so that it corresponds to the R function `text(rt1)`.

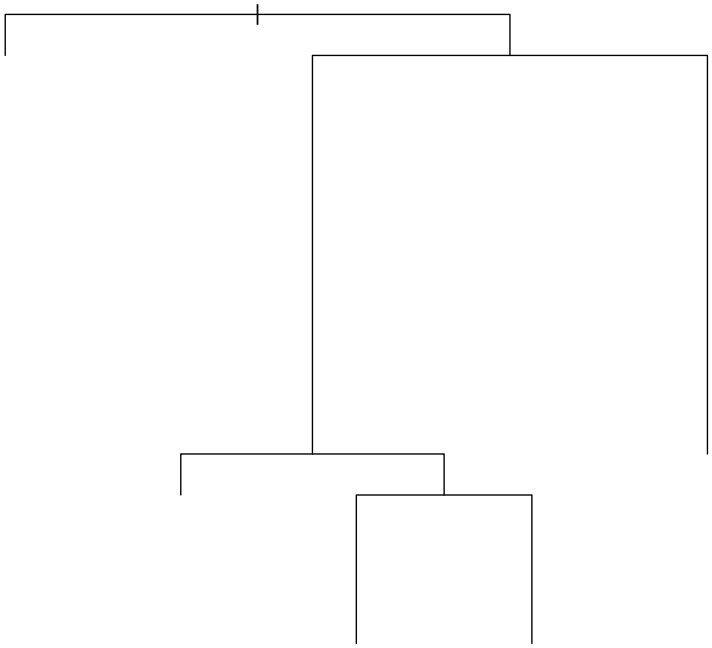

Figure 1: The tree *rt1*

.

(c) Suppose a passenger travelled in Titanic with a ticket fare of 50, what would be the predicted probability of his survival according to the classification tree $rt1$?.

(d) Calculate the precision and recall of the classification tree $rt1$.

(e) I next fit a classification tree to this dataset more explanatory variables as follows.

```
> rt = rpart(Survived ~ as.factor(Pclass)+ Sex + SibSp + Parch + Fare
+ Embarked,  method="class", data=titanic)
> rt
n= 891

node), split, n, loss, yval, (yprob)
* denotes terminal node

1) root 891 342 0 (0.61616162 0.38383838)
2) Sex=male 577 109 0 (0.81109185 0.18890815) *
3) Sex=female 314  81 1 (0.25796178 0.74203822)
6) as.factor(Pclass)=3 144  72 0 (0.50000000 0.50000000)
12) Fare>=23.35 27   3 0 (0.88888889 0.11111111) *
13) Fare< 23.35 117  48 1 (0.41025641 0.58974359)
26) Embarked=S 63  31 0 (0.50793651 0.49206349)
52) Fare< 10.825 37  15 0 (0.59459459 0.40540541) *
53) Fare>=10.825 26  10 1 (0.38461538 0.61538462)
106) Fare>=17.6 10   3 0 (0.70000000 0.30000000) *
107) Fare< 17.6 16   3 1 (0.18750000 0.81250000) *
27) Embarked=C,Q 54  16 1 (0.29629630 0.70370370) *
7) as.factor(Pclass)=1,2 170   9 1 (0.05294118 0.94705882) *
```

Based on the tree rt, what would be the predicted probability of survival for a female passenger who travelled in *Pclass* 3 with a fare of 15, who embarked in Southampton and had 2 siblings (and no spouse) aboard?.

(f) Again based on the tree rt, what would be the predicted probability of survival for a female passenger who travelled in *Pclass* 1 with a fare of 15, who embarked in Southampton and had a spouse (and no siblings) aboard?.