7390–Machine Learning and Data Sciences

Northeastern University, Summer 2018

Final exam (Practice)

- The exam is open book.

Name: _____

1. For each of parts (a) through (d), indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.
   (a) The sample size n is extremely large, and the number of predictors p is small.
   (b) The number of predictors p is extremely large, and the number of observations n is small.
   (c) The relationship between the predictors and response is highly non-linear.
   (d) The variance of the error terms, i.e. $\sigma^2 = Var(\epsilon)$, is extremely high.

2. Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n (number of observations) and p (number of features).
   (a) We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.
   (b) We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

3. What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred?

4. Suppose we collect data for a group of students in a statistics class with variables X_1 = hours studied, X_2 = undergrad GPA, and Y = receive an A. We fit a logistic regression and produce estimated coefficient, $\beta_0 = 6$, $\beta_1 = 0.05$, $\beta_2 = 1$.
   (a) Estimate the probability that a student who studies for 40 h and has an undergrad GPA of 3.5 gets an A in the class.
   (b) How many hours would the student in part (a) need to study to have a 50 % chance of getting an A in the class?

① A flexible model will perform better in general. Because of the large sample size, we're likely to overfit even when using a more flexible model. Meanwhile a more flexible model tends to reduce bias.

② A inflexible model will perform better in general. A flexible model will cause overfitting because of the small sample size. This usually means bigger inflation in variance and small reduction in bias.

③ A flexible model will perform better in general because it'll be necessary to use a flexible model to find a non-linear effect.

④ An inflexible model will perform better. a flexible model will capture too much noise in the data due to the large variance of the errors

classification vs regression

main difference : the output variable in regression is numerical (or continuous) while that for classification is categorical.

inference : use the model to learn about the data generation process

Prediction : use the model to predict the outcomes for new data points

---

The advantages of a very flexible approach:

it may give a better fit for non-linear models and it decreases bias.

4

$$P(x) = \frac{e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}}{1 + e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}}$$

① 0.731 ② 50h.

The estimated probability that a student gets an A in the class on a test data is given in the following table.

ID	X_1	X_2	Y	Estimated Probability
1	50	3.9	A	0.60
2	45	3.7	A	0.49
3	42	3.9	A	0.5
4	40	3.5	not A	0.38
5	38	3.4	not A	0.33
6	36	3.2	not A	0.27
7	35	3.8	A	0.39
8	34	3.1	not A	0.23
9	32	3.0	not A	0.20
10	31	3.1	A	0.21
11	30	2.9	not A	0.17
12	28	3.7	A	0.29
13	28	3.3	not A	0.21
14	25	2.9	not A	0.16
15	20	2.8	not A	0.10

(c) Estimate the test AUC on this dataset.

(d) Calculate confusion matrix, FP and TP rates, accuracy, and balanced accuracy for threshold=0.3

5. Explain how k-fold cross-validation is implemented. What are the advantages and disadvantages of k-fold cross–validation relative to the validation set approach?

6. We perform best subset, forward stepwise, and backward stepwise selection on a single data set. For each approach, we obtain $p + 1$ models, containing $0, 1, 2, \ldots, p$ predictors. Explain your answers:

   (a) Which of the three models with k predictors has the smallest training RSS?

   (b) Which of the three models with k predictors has the smallest test RSS?

   (c) True or False:

   1. The predictors in the k-variable model identified by forward stepwise are a subset of the predictors in the $(k + 1)$-variable model identified by forward stepwise selection.
   2. The predictors in the k-variable model identified by back- ward stepwise are a subset of the predictors in the $(k + 1)$- variable model identified by backward stepwise selection.
   3. The predictors in the k-variable model identified by back- ward stepwise are a subset of the predictors in the $(k + 1)$- variable model identified by forward stepwise selection.
   4. The predictors in the k-variable model identified by forward stepwise are a subset of the predictors in the $(k + 1)$-variable model identified by backward stepwise selection.
   5. The predictors in the k-variable model identified by best subset are a subset of the predictors in the $(k + 1)$-variable model identified by best subset selection.

7. Indicate which of (a) through (d) is correct. Justify your answer. The lasso, relative to least squares, is:

   (a) More flexible and hence will give improved prediction accuracy when its increase in bias is less than its decrease in variance.

   (b) More flexible and hence will give improved prediction accuracy when its increase in variance is less than its decrease in bias.

   (c) Less flexible and hence will give improved prediction accuracy when its increase in bias is less than its decrease in variance.

   (d) Less flexible and hence will give improved prediction accuracy when its increase in variance is less than its decrease in bias.

estimate your AUC

$$AUC \approx P(score(x^+) \geq score(x^-))$$

$$= \frac{U}{|N^+| \cdot |N^-|}$$

$$\frac{9+9+9+9+4+7}{6 \times 9} = 54$$

TP: 4 FP: 2

FN: 2 TN: 7

Accuracy: $\frac{4+7}{15} = \frac{11}{15}$

Balanced accuracy: $\frac{1}{2}\left(\frac{TP}{N^+} + \frac{TN}{N^-}\right)$

$$2\left(\frac{4}{6} + \frac{7}{9}\right) = \frac{12+14}{18 \times 2}$$

$$= \frac{26}{36}\ \frac{13}{18} = \frac{13}{18}$$

FP = 2

TP Rate $= \frac{4}{6} = \frac{2}{3}$

FP Rate $= \frac{2}{7}$

TN Rate $= \frac{2}{9}$

FN Rate $= \frac{2}{6} = \frac{1}{3}$

explain the K-fold cross

The k-cross validation is implemented by taking the n observations and randomly splitting it into k non-overlapping groups of length of (approx.) n/k

8. Suppose we estimate the regression coefficients in a linear regression model by minimizing

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

for a particular value of λ. For parts (a) and (b), indicate which of 1 through 5 is correct. Justify your answer.

(a) As we increase λ from 0, the training RSS will:

1. Increase initially, and then eventually start decreasing in an inverted U shape.
2. Decrease initially, and then eventually start increasing in a U shape.
3. Steadily increase.
4. Steadily decrease.
5. Remain constant.

(b) Repeat (a) for test RSS.