Statistical Learning Problem Set 1

Jay Lee

8/30/2017

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.

- A flexible method would be better for this sample, because changing any one point out of a large sample will not have as much influence on the model. In addition, any random changes in the pattern will only affect the model if they are systemic.

(b) The number of predictors p is extremely large, and the number of observations n is small.

- An inflexible model would be better here, because any variance picked up by the model is more likely to be due to ϵ rather than f, because of such a small sample size. In addition, fewer predictors means more guessing in a flexible model.

(c) The relationship between the predictors and response is highly non-linear.

- A flexible model would be better here, because a highly non-linear relationship would cause high bias if we impose an inflexible model.

(d) The variance of the error terms, i.e. $\sigma^2 = Var(\epsilon)$, is extremely high.

- We should fit an inflexible model, because a flexible model would pick up too much variance to fit test data well.

---

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 and p.

(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.

- Regression (numeric y), inference, n = 500, p = 3.

(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.

- Classification (categorical y), prediction, n = 20, p = 13.

(c) We are interested in predicting the % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.

- Regression (numeric y), prediction, n = 52, p = 3.

---

4. You will now think of some real-life applications for statistical learning.

(a) Describe three real-life applications in which classification might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

- A model whose response is whether a given person will vote in an election. Predictors include factors like voting history, age, address (state laws differ), registration status, eligibility, income, education level, and competitiveness of races. The goal of this model is both inference (for researchers/political scientists) and prediction (for campaigns/parties).
- A model where the response is how a Reed student will travel to class on a given day (bus/MAX, bike, car, walk). Predictors include factors like weather, class schedule, distance to campus, auxiliary errands (i.e. grocery shopping), and carpool availability. The goal of this model is inference.
- A model where the response is the form of dinner a person consumes (homemade, delivery, takeout, eating out, frozen/microwave). Predictors include factors such as income, prices for each option locally, transportation methods available, hours worked that day/week (proxy for time available), number of people in household, distance to options, and dietary restrictions. The goal of this model is inference.

(b) Describe three real-life applications in which regression might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

- A model where the response is the mpg a car gets. Predictors include make (brand), number of cylinders, engine volume, and car weight. The goal of this model is both inference (for manufacturers) and prediction (for consumers).
- A model where the response is how many people ride the MAX on any given day. Predictors include whether it's a weekday or weekend, weather, previous traffic patterns, and events going on.
- A model where the response is how far people travel to get to work. Predictors include income, transportation available, and distance to a major city.

(c) Describe three real-life applications in which cluster analysis might be useful.

- Amazon could (or probably already does) group customers by their purchases, then use these clusters to recommend new things somebody in a cluster bought to other people in that cluster.
- Dimension reduction. Given a dataset, you could cluster the data and reduce to the most salient dimensions (i.e., what do all these observations share?).
- Grouping organisms into taxonomy classes. Given features and DNA, cluster the organisms and assign each cluster a group name.

---

5. 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?

- A very flexible model can fit your training data set very well, but at the expense of having too much variance to fit the test data. A more flexible approach would be good when the data has a clear non-linear pattern. A less flexible approach would be good when a pattern exists, but the data has high variability.

---

6. Describe the differences between a parametric and a non-parametric statistical learning approach. What are the advantages of a parametric approach to regression or classification (as opposed to a non-parametric approach)? What are its disadvantages?

- A parametric approach assumes a form that the data follows then assigns coefficients based on the data, where a non-parametric approach just "fits a curve" based entirely on the response value of the points around the area in question. Parametric approaches are simpler and less computationally expensive, but if the assumed form does not match the actual data, the model will have high bias.

---

7. The table below provides a training data set containing six observations, three predictors, and one qualitative response variable.

Obs.	X1	X2	X3	Y
1	0	3	0	Red
2	2	0	0	Red
3	0	1	3	Red
4	0	1	2	Green
5	-1	0	1	Green
6	1	1	1	Red

Suppose we wish to use this data set to make a prediction for Y when $X1 = X2 = X3 = 0$ using K-nearest neighbors.

(a) Compute the Euclidean distance between each observation and the test point, $X1 = X2 = X3 = 0$.

$$
\begin{aligned}
|\hat{o}_1 - \hat{x}_t| &= \sqrt{(0-0)^2 + (3-0)^2 + (0-0)^2} = 3 \\
|\hat{o}_2 - \hat{x}_t| &= \sqrt{(2-0)^2 + (0-0)^2 + (0-0)^2} = 2 \\
|\hat{o}_3 - \hat{x}_t| &= \sqrt{(0-0)^2 + (1-0)^2 + (3-0)^2} = \sqrt{10} \\
|\hat{o}_4 - \hat{x}_t| &= \sqrt{(0-0)^2 + (1-0)^2 + (2-0)^2} = \sqrt{5} \\
|\hat{o}_5 - \hat{x}_t| &= \sqrt{(1-0)^2 + (0-0)^2 + (1-0)^2} = \sqrt{2} \\
|\hat{o}_6 - \hat{x}_t| &= \sqrt{(1-0)^2 + (1-0)^2 + (1-0)^2} = \sqrt{3}
\end{aligned}
$$

(b) What is our prediction with K = 1? Why?

- Green, because the closest point is $\hat{o}_5$.

(c) What is our prediction with K = 3? Why?

- Red, because the three closest points $(\hat{o}_5, \hat{o}_6, \hat{o}_2)$ have y values of Green, Red, Red (respectively).

(d) If the Bayes decision boundary in this problem is highly non-linear, then would we expect the best value for K to be large or small? Why?

- We would expect the best value for K to be small, because that allows for a more flexible model which could more closely follow the Bayes decision boundary.

---

8. Using the notation standards described at the end of chapter one, please provide notation for the following objects:

(a) Input: The 10 photos that we looked at on the first day, as if they were scanned at 64 x 64 pixel resolution.

$$
\mathbf{X} = \begin{pmatrix}
x_{1,1} & x_{1,2} & \cdots & x_{1,4096} \\
x_{2,1} & x_{2,2} & \cdots & x_{2,4096} \\
\vdots & \vdots & \ddots & \vdots \\
x_{10,1} & x_{10,2} & \cdots & x_{10,4096}
\end{pmatrix}
$$

(b) Transformed Input: The 10 photos, after a small number of features have been identified.

$$
\mathbf{X}_{trans} = \begin{pmatrix}
x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\
x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\
\vdots & \vdots & \ddots & \vdots \\
x_{10,1} & x_{10,2} & \cdots & x_{10,p}
\end{pmatrix}
$$

(c) Output: The associated actual ages of those 10 photos.

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{10} \end{pmatrix}$$

(d) Model: Provide a guess at what f might look like (there is no single right answer here).

$$f = 30 * grayHair + 5 * wrinkleLevel$$