Hans Trautlein[1]
Due: January 29, 2016

Book Problems

1. Indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify my answer.[2]

   (a) A flexible method would be appropriate, as you will get a better "fit" on the data with a more flexible approach and you have a large enough n to avoid the issues of using a flexible model with a small n.

   (b) You would want an inflexible method in this circumstance. If you have a small amount of variables you are more likely to have a model that is difficult to fit exactly, and the large amount of predictors might allow you to accidentally the model.

   (c) A more flexible model would have a better fit than a less flexible model. Since there are more "degrees of freedom" in a flexible model you could get a better fit on this highly non-linear model. While you might be able to get a decent fit with an inflexible model you would most likely get a better fit with a flexible one.

   (d) If the variance of the error terms, $\sigma^2 = Var(\epsilon)$, is extremely high I would expect a flexible model to be worse than an inflexible method because you might be more likely to overfit your model. In Nate Silver parlance your model would end up confusing the "signal" for the "noise."

2. (a) $n = 500$, $p = 3$. This a regression problem that is most interested in inference.

   (b) $n = 20$, $p = 13$. This is a classification problem that is most interested in prediction.

   (c) $n = 52$ (because of weekly data), $p = 3$. This is a regression problem that is most interested in prediction.

4. (a) Real-life applications of *classification*. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain.

   i. Response: Problem With Computer. Predictors: Can Computer Start, Operating System, Operating System Version, Run Antivirus, User's Familiarity With Computers, Is User A Gamer, Does User Pirate Anything, etc. Inference.

   ii. Response: Candidate Voter Supports. Predictors: Pro-Life/Choice, Church Visits/Week, Self-Identified Party, Past Voting Behavior, Income, Magazines Subscribed To, etc. Prediction.

   iii. Response: A student's favorite candy. Predictors: Hometown's Geographic Region, Interest in Organic Foods, Where They Shop, Income, Ethnicity, etc. Inference.

   (b) Real-life applications of *regression*. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain.

---

[1]I went to Andrew's office hours for help completing this problem set.

[2]Every answer in this Problem Set has referred frequently to Chapter 2 in James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*. New York: Springer, 2013. I have also gotten help from

i. Response: Apple Stock Price. Predictors: P/E Ratio, Previous Day's Opening, Previous Day's Close, Media Coverage, etc. Prediction.

ii. Response: Spouse's Age. Predictors: Gender, Gender of Spouse, Age, When They Met, Income, Spouse's Income, etc. Inference.

iii. Response: Candidates Total Vote Share Percentage. Predictors: State Polls, National Polls, Endorsements, Favorability Numbers, All of these variables for other candidates, Media Coverage, etc. Prediction.

(c) Real-life applications of *cluster analysis*.

i. Netflix movie recommendations! Maybe you like some intersection of thriller and action movies, or comedy and drama? I'm sure this is something that they do in some form or another (maybe not using cluster analysis, but it sounds like it'd be useful).

ii. What sort of ideology do you hold as a voter? Maybe you are a Republican, but are you more establishment or Tea Party? Or perhaps you have a more libertarian bent to your voting habits. On the Democratic side are you more of a neoliberal or a Marxist.

iii. Google probably also does this to try to target ads at you in the best way possible. What type of consumer are you? Are you shopping for flights to Jamaica and a bathing suit, or are you looking for tires for your car?

5. A very inflexible model for regression and classification has both advantages and disadvantages as well, as it often is easier to make inferences from more inflexible methods compared to flexible methods. For example, if inference is the overall goal then you might prefer a less flexible model, like "least squares" or "subset selection lasso" as opposed to "support vector machines." The restriction inherent in more inflexible models makes inference easier.

A more flexible model allows you to reduce "bias" and also fit models that are more complicated that might only be explainable by a simpler and more inflexible model. A more flexible approach might be preferred when prediction is weighed much heavier than inference in the model's goals. This is a slight oversimplification, but then the "why" takes a backseat to the "what," that is, the output that the model creator cares about is less a great understanding of the inner workings of the model and why it works that way but instead that the model has the best output possible. This might not be the best approach in the long run though - to at least a tiny degree you want to understand your models! But adding every variable that you have into the model might result in overfitting. This is further explained on page 26 of the text for this course in section 2.1.3, *The Trade-Off Between Prediction Accuracy and Model Interpretability.*

Overall, flexibility is preferred when you are looking for great prediction and inflexibility is prized when you are looking for easy inference.

6. Parametric approaches ask you to just estimate the variables and coefficients you are going to use because the form has already been chosen for you, whereas a non-parametric approach doesn not select the form for you at all.

The advantages of a parametric approach as opposed to a non-parametric approach are that you already have an assumed "form" for the model (linear) so you only have to estimate

the coefficients on the variables that you choose to use, instead of having to also figure out whether or not they are going to have some exponent added to them.

There are also disadvantages of a parametric approach however. Parametric approaches might oversimplify the form of the proper function for the best achievable model as the form might be different than the linear form assumed.

7. (a) I have placed my answers in the table below:

Obs.	X_1	X_2	X_3	Y	Distance
1	0	3	0	Red	3
2	2	0	0	Red	2
3	0	1	3	Red	$\sqrt{10}$
4	0	1	2	Green	$\sqrt{5}$
5	-1	0	1	Green	$\sqrt{2}$
6	1	1	1	Red	$\sqrt{3}$

(b) My prediction when $K = 1$ is that $Y = Green$ because of observation 5.[3]

(c) My prediction when $K = 3$ is that $Y = Red$ as observation 2 and 6 are the closest here.

(d) We would expect the *best* value for K to be small if the Bayes decision boundary in this problem is highly non-linear. K would be expected to be large if there were more points factored into the model and thus a more linear boundary would be drawn. A great image to go with this question is Figure 2.16 in the text for this course.

Additional exercise

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

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

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

Output: The associated actual ages of those 10 photos.

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

Input

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,64^2} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,64^2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{10,1} & x_{10,2} & \cdots & x_{10,64^2} \end{pmatrix}$$

[3]Note, that 7b-7c are not as well informed as the rest of this problem set. They are basically guesstimates.

Transformed Input

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,15} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,15} \\ \vdots & \vdots & \ddots & \vdots \\ x_{10,1} & x_{10,2} & \cdots & x_{10,15} \end{pmatrix}$$

Here it is suggested that there are fifteen different factors that we will pulled out of our input and will end up using in our model. For example, the first factor could be a dummy variable of whether or not the hair is more than 50% grey, the second factor could be nose length (cm), the third factor could be whether or not crows feet are visible, the fourth factor could be whether or not the person is bald, etc...this would correspond to the columns up above, with each factor having it's own column, and each factor represented once for each observation.

Output

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

Here each y will be an age, for example, if $y_3 = 43$ that would mean that the modeled age of the person in observation 3 would be forty-three years old.

Model

Here is an example model that could exist here:

$$Y = \beta_0 + \beta_1 X_{ears} + \beta_2 X_{hair} + \beta_3 X_{crows\_feet} + \beta_4 X_{bald} + \ldots + \beta_{15} X_{smiling} + \epsilon$$

Hopefully this model would work well!