STAT 422/722 Spring 2016 Homework #1
Limited Solutions

Professor Adam Kapelner

Sunday 26th February, 2017

Problem 1

These questions are about prediction and modeling theory.

(a) [easy] Give three examples of "predictions".

(1) predicting selling prices of homes in Queens (2) predicting quality ratings of white wines (3) predicting if customers are going to cancel their subscriptions.

(b) [easy] Considering the etymological definitions, do we "predict" or do we "forecast"? Explain your answer.

forecast. A predictions (if they are true to the etymological root) uses voodoo, divination, prophecy or otherwise means that are non-scientifically validated (at the time of this writing) to foretell the future.

(c) [easy] Explain what each row in a dataframe represents. Give every synonym for the rows in a dataframe. Write a sentence as to why each of the particular vocabulary words were employed here.

observation	Each row is "observed"
unit	The rows are the individual "units" in analysis.
record	The rows are "recorded" in a database.
object	The rows consist of information about an individual "object".
subject	The rows detail a "subject" (usually a person).

(d) [easy] Explain what each column in a dataframe represents. Give every synonym for the column. Write a sentence as to why each of the particular vocabulary words were employed here. In the models in this class, there will be *one* special column. Explain what its called and give definitions for all its synonyms.

features	Information is considered to be “features” of an object.
attributes	“
characteristics	“
variables	These are numbers that vary subject-subject.
measurements	Information is “measured” on an object.

(e) [easy] Explain why theories are mathematical (generally speaking).

It is only way to make statements precise so they can be shared universally and validated through measurements.

(f) [harder] Below is an excerpt from Box and Draper (1987, page 424) and contains the famous line “all models are wrong, but some are useful”. Explain what this means.

Models are “wrong” in the sense that they do not accurately represent the measurements precisely but they are “useful” in the sense that they provide approximations that can be good enough to rely on to make decisions or engineer systems.

(g) [harder] What did we call $\delta(\xi)$ in class (ibid, Equation 13.1.7)?

Misspecification error.

(h) [E.C.] If the true model were to be found and estimated from the a finite sample dataframe, would it be better for inference than a simple model? Yes / no explain. Would it be better for prediction? Yes / no explain.

(i) [harder] In many sciences there is a belief in the so-called “tapering effect” which means that there are large effects, then small effects, then really small effects. How can this be used to explain the success of linear regression in predicting responses with likely myriad inputs (a la slide 32, Lecture 2)?

As long as the predictors are responsible for the large effects, linear regression can be quite a useful approximation for predicting the response.

(j) [E.C.] Why will “full reality always remain elusive in the biological sciences”? (Burnham and Anderson, 1998) What does that say about the softer sciences such as economics, psychology, sociology, etc?

(k) [harder] Explain why non-parametric models can give you lower misspecification error but possibly higher model estimation error.

This should have been marked as extra credit. Ignore.

(l) [harder] I got a phone call from a startup founder who described the idea for a company that predicts startup success. The founder told me their model was that they assign 5 points to a startup for having two or more C-level founders, 10 points for closing an angel round, 5 points for the first employee, etc. What kind of model is this? Why would you trust / not trust its predictions?

This is not a data-driven model. There is no historical dataframe nor validation against a gold-standard. It is a made-up model. However, if the person making it up is doing so through experience, then there could be some value in it. However, these types of models are not ones humans are good at making.

(m) [easy] Look at slide 32 in lecture 2 but not slide 33. Write down all observations you have about this picture.

(1) We never can figure out the true response function. (2) Many variables are non-causal and only correlational. (3) Some variables are unrelated.

(n) [E.C.] If you find a confounding / lurking variable Z, does that mean Z is causal? If yes, explain; if not, provide a counterexample.

(o) [harder] Given a model with one response and 10 variables where the 10 variables are realized simultaneously but the response is realized afterwards, how many models can be posited? Ignore the fact that each functional relationship can be different. See slide 35 lecture 2.

$$4^{\binom{10}{2}}2^{10}$$

(p) [easy] What advantage does a "real" correlation have over a *spurious* correlation?

Real correlations can be put to use for prediction.

(q) [harder] My friend has six children, all born on Wednesday. Is this "significant"? Discuss.

If this is the only data you have ever seen or thought about in your life, yes. Chances are though, you look at data in some form all day everyday so after you correct for multiple tests, this is likely not significant.

Problem 2

THIS PROBLEM IS OPTIONAL. Here we will be analyzing the theory that "skiing is dangerous".

(a) [easy] Define the response(s) and the predictors(s) in this model.

(b) [harder] Mathematize this model. Explain clearly what you are measuring and how it is measured.

(c) [easy] If you were to use a data-driven approach, what would the dataframe look like? What are the datatypes of each variable? Will the eventual model be a regression? Classification? Something else?

(d) [harder] Explain why a deterministic model for your response variable is absurd.

(e) [harder] Create a stochastic (statistical) model for the response. Pay attention to which letters are lowercase/uppercase.

(f) [harder] In this model, would the error term $\mathcal{E}$ be large or small? Explain.

(g) [harder] If you were given leeway to collect a multidimensional representation of "skiing" (i.e. a more natural, raw representation), what would you collect?

(h) [difficult] Build a causal model (using bubbles and arrows) for the response.

(i) [difficult] Does skiing *cause* the response? If so, is it a major contributor? Explain using your diagram from (h).

Problem 3

We will discuss confounding here. The prevailing data on wage inequality says that women are paid 90 cents on the dollar that men earn for comparable work.

(a) [easy] If you were to do a regression of y : earnings on x_1 : employee height, what would the results look like and why? Report the p-values on each $\hat{\beta}_j$'s and the omnibus F statistic's p value.

The $\hat{\beta}$ would be positive, the t test and the F test would be significant (and have the same values).

(b) [easy] Build a causal model for y: earnings and x_1: employee height and x_2 : employee gender. No need to include unknown variables.

There are two models students considered here.

The first links gender to both earnings and height (thus assuming that height is non-causal and that gender is the lurking variable).

The second is the same as the first except height also links to earnings.

(c) [easy] If you were to do a regression of y : earnings on x_1 : employee height and x_2 gender, what would the results look like? Report the p-values on each $\hat{\beta}_j$'s and the omnibus F statistic. Also say if the $\hat{\beta}_j$ values changed and which direction vs. the regression in part (a).

In model 1, there would be no effect of height once gender is entered into the model. The coefficient would be near zero with an insignificant p value. The omnibus F test is still significant at around the same level as (a).

In model 2, height is still causal but once gender is controlled for the coefficient shrinks closer to zero but remains significant. The F test would be more significant than (a).

Problem 4

A few questions about likelihood. Imagine a simple model where you flip the same coin three times. You are modeling the response "flipping a head" with a statistical model and make a parametric assumption that the event is a Bernoulli r.v.

(a) [easy] What does θ represent in this model?

The probability the coin flips heads.

(b) [harder] Find the joint probability density / mass function for these three events.

$$\mathbb{P}(Y_1, Y_2, Y_3) = \theta^{x_1+x_2+x_3}(1-\theta)^{3-(x_1+x_2+x_3)}$$

(c) [easy] Find the likelihood function.

Same as in (b).

(d) [E.C.] Find the maximum likelihood estimator for θ.

(e) [easy] If your data was heads, heads, tails (in the image above). What is the maximum likelihood estimate? This will allow you to locate the best possible model given your parametric assumptions.

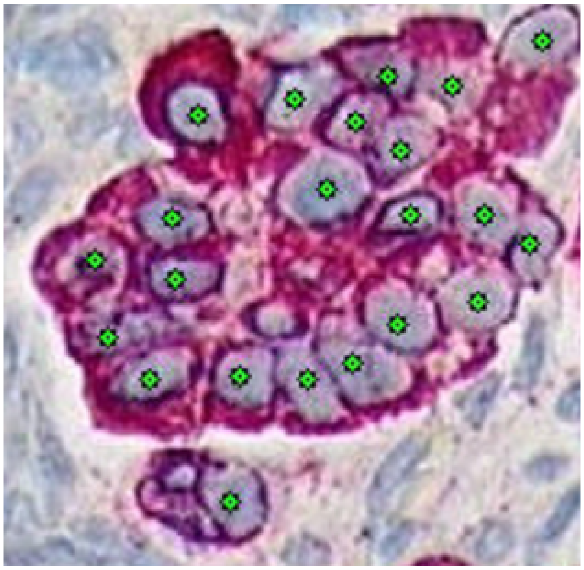
2/3

(f) [E.C.] If your maximum likelihood estimate in (e) was indeed the value of θ, what data would be most probable? Note: this is the inverse question of likelihood and it is meant to trick you.

Problem 5

Here we will be considering different types of AI. Imagine you have the following problem: you are tasked with finding the centers of cancer cells in microscopic images. Images are composed of pixels which encode a color. Typical coloring schemes for computer graphics are "true color" and use about 16.8 milion different colors per pixel. Here's a typical image:

All cells are stained using immunohistochemistry using a compound which appears blue. But cancer cells in this type of immunohistochemical staining appear to have a red membrane due to a surface marker. (In the example above a green dot is placed in the center of every cancer cell to indicate to you what you are looking for; this dot is *not part of the original image*). Assume we are using a data-driven approach to solve this problem.



(a) [easy] What type of AI would work the best here? No need to discuss.

Deep learning.

(b) [easy] What is the raw data representation?

The pixel values as represented inside the computer.

(c) [harder] What is the unit of analysis?

The pixel.

(d) [harder] Why is it easy for you to find the cell centers but difficult for a computer?

We are good at building such models using our senses that both know what information is important and can filter out all irrelevant information. The computer needs an explicit algorithm to do so.

(e) [difficult] Consider the situation where you employ classic machine learning. What features would you collect on the units of analysis? Enumerate and describe these features.

(f) [difficult] How would you sample to build a dataframe (collect historical data)? Explain the procedure and the goals of this step. This is known as "supervised machine learning".

(g) [harder] Once you built the dataframe, would a human be able to use that dataframe to create a predictive model? Yes / no and discuss.

No. Humans are poor at building models from numerical features of high dimension.

(h) [easy] Considering you selected features in (e) and sampled in (f), would this entire enterprise be considered “good machine learning”? Explain.

Yes. Thought was given to features that are important in determining whether or not the pixel is pictured in a cancer cell. Thought was also given to which examples are entered in. There will be many diverse examples of cancer cells and many diverse examples of non-cancer cells.

(i) [easy] Now you have the dataframe. Given the problem context, which worldview would you select — the parametric or the nonparametric and why?

Likely your bottom line is accuracy in finding cancer cells and you don’t care which of your features is driving the predictions, then the non-parametric modeling approach would be employed.

(j) [difficult] Assume you went the parametric model route and you built a vanilla linear model. Explain where it would be wrong. Be explicit by referencing your predictors in (e).

Problem 6

These exercises will dicsuss the linear model and linear regressions.

(a) [easy] Think of three loss functions $L(e_1, \ldots, e_n)$. Do not list any that we did in class.

There are many. Simple ones can be taken by using sums of even powers.

(b) [harder] You are building a data-driven model and choose to use the linear parametric assumption but not necessarily the other three OLS assumptions. Describe a situation where fitting this model using $L = SSE$ is *not* a good idea because it does not accuractely reflect the loss function in your situation at hand.

Imagine the typical utility function where if you act on a prediction that losing $1 is a higher loss than making $1. SSE will fit the best function where these two scenarios are equal.

(c) [E.C.] Prove that the MLE of the β’s is the same solution as minimizing SSE.

(d) [E.C.] Assume that you have proven the above and plugged in those estimates to the likelihood expression. Now $\sum_{i=1}^{n} \mathcal{E}_i^2 = \sum_{i=1}^{n} e_i^2 = SSE$. Prove that $\hat{\sigma^2}_{MLE} = MSE$.

Problem 7

We will now analyze the baseball data (`baseball.csv`). You can use any software package you wish to answer these questions.

(a) [easy] Fit a linear model with reponse variable y : salary in thousands. Use all available predictors. Provide a valid interpretation on $\hat{\beta}_j$ for the feature “number of RBI’s”.

If RBIs increase by one unit, the salary will increase by $17,415 for another naturally-observed baseball player with all other characteristics being the same.

(b) [easy] Is this interpretation reasonable given what you know about number of RBI’s and how it is related to other predictors? You may need to ask someone who knows a bit about baseball.

This interpretation is probably not realistic as there is a large degree of collinearity between RBIs and some of the other predictors.

(c) [easy] Does a causal additive model for number of RBI’s make sense? Yes / no.

The answer here can be both yes/no.

(d) [easy] Would you be able to make a randomized experiment to find the additive causal effect of number of RBI’s? Yes / no.

No.

(e) [easy] Some of these variables may be significant because we dredged. Why is this likely *not* the case?

After making a Bonferroni adjustment, only one previously significant variable is no longer significant.

(f) [E.C.] Use a likelihood ratio test to test the effect of number of RBI’s and Number of Walks.