MATH 390.4 / 650.2 Spring 2018 Homework #5t

Professor Adam Kapelner

Due 11:59PM Friday, May 18, 2018 under the door of KY604

(this document last updated Thursday 10th May, 2018 at 11:44pm)

Instructions and Philosophy

The path to success in this class is to do many problems. Unlike other courses, exclusively doing reading(s) will not help. Coming to lecture is akin to watching workout videos; thinking about and solving problems on your own is the actual "working out." Feel free to "work out" with others; **I want you to work on this in groups.**

Reading is still *required*. For this homework set, read about all the concepts introduced in class online e.g. probabilistic classification, the logistic link function, performance characteristics of binary classification, asymmetric cost / reward classifiers, bias-variance decomposition, bagged models, RandomForests®, correlation, causation, lurking variables and graphical depictions of causal models. It is your responsibility to supplement the notes with your own readings. Also, read the introduction in Finlay.

The problems below are color coded: green problems are considered *easy* and marked "[easy]"; yellow problems are considered *intermediate* and marked "[harder]", red problems are considered *difficult* and marked "[difficult]" and purple problems are extra credit. The *easy* problems are intended to be "giveaways" if you went to class. Do as much as you can of the others; I expect you to at least attempt the *difficult* problems.

This homework is worth 100 points but the point distribution will not be determined until after the due date. See syllabus for the policy on late homework.

Up to 10 points are given as a bonus if the homework is typed using LaTeX. Links to instaling LaTeX and program for compiling LaTeX is found on the syllabus. You are encouraged to use `overleaf.com`. If you are handing in homework this way, read the comments in the code; there are two lines to comment out and you should replace my name with yours and write your section. The easiest way to use overleaf is to copy the raw text from hwxx.tex *and* preamble.tex into two new overleaf tex files with the same name. If you are asked to make drawings, you can take a picture of your handwritten drawing and insert them as figures or leave space using the "\vspace" command and draw them in after printing or attach them stapled.

The document is available with spaces for you to write your answers. If not using LaTeX, print this document and write in your answers. I do not accept homeworks which are *not* on this printout. Keep this first page printed for your records.

NAME: _____

Problem 1

These are questions about the Finlay's introduction to his book.

(a) [easy] Finlay introduces predictive analytics by using the case study of what supervised learning problem? Explain.

(b) [difficult] What does a credit score of 700 mean? Use figure 1.2 on page 5 when answering this question.

(c) [difficult] How much more likely is someone to default if that have 9 or more credit cards than someone with 4-8 credit cards?

(d) [easy] Summarize Finlay's conception of "big data".

Problem 2

This question is about probability estimation. We limit our discussion to estimating the probability that a single event occurs.

(a) [easy] What is the difference between the regression framework and the probability estimation framework?

(b) [easy] Is probability estimation more similar to regression or classification and why?

(c) [difficult] Why was it necessary to think of the response Y as a random variable and why in particular the Bernoulli random variable?

(d) [difficult] If we use the Bernoulli r.v. for Y, are there any error terms (i.e. δ, ϵ, e) anymore? Yes/no.

(e) [easy] What is the difference between f in the regression framework and f_{pr} in the probabilistic classification framework?

(f) [difficult] Is there a t_{pr}? If so, what does it look like?

(g) [easy] Write out the likelihood as a function of f_{pr}, the $\boldsymbol{x}_i$'s and the y_i's.

(h) [difficult] What assumption did you have to make and what would happen if you didn't make this assumption?

(i) [easy] Is f_{pr} knowable? Yes/no.

Problem 3

This question continues the discussion of probability estimation for one event via the logistic regression approach.

(a) [harder] As before, if we are to get anywhere at all, we need to approximate the true function f_{pr} with a function in a hypothesis set, $\mathcal{H}_{pr}$. Let us examine the range of all elements in $\mathcal{H}_{pr}$. What values can these functions return and why?

(b) [difficult] We would also feel warm and fuzzy inside if the elements of $\mathcal{H}_{pr}$ contained the term $\boldsymbol{w} \cdot \boldsymbol{x}$. What is the main reason we would like our prediction functions to contain this linear component?

(c) [easy] The problem is $\boldsymbol{w} \cdot \boldsymbol{x} \in \mathbb{R}$ but in (a) there is a special range of allowable functions. We need a way to transform $\boldsymbol{w} \cdot \boldsymbol{x}$ into the range from (a). What is this function called?

(d) [easy] Give some examples of such functions.

(e) [easy] We will choose the logistic function. Write the likelihood again from 2(g) but replace f_{pr} with the element from $\mathcal{H}_{pr}$ that uses the logistic function.

(f) [difficult] Simplify your answer from (e) so that you arrive at:

$$\sum_{i=1}^{n} \ln\left(1 + e^{(1-2y_i)\boldsymbol{w}\cdot\boldsymbol{x}_i}\right)$$

(g) [E.C.] We will now maximize this likelihood w.r.t to $\boldsymbol{w}$ to find $\boldsymbol{b}$, the best fitting solution which will be used within g_{pr} i.e.

$$\boldsymbol{b} = \underset{\boldsymbol{w}\in\mathbb{R}^{p+1}}{\arg\max} \left\{ \sum_{i=1}^{n} \ln\left(1 + e^{(1-2y_i)\boldsymbol{w}\cdot\boldsymbol{x}_i}\right) \right\}$$

to do so, we should find the derivative and set it equal to zero i.e.

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{w}} \left[\sum_{i=1}^{n} \ln\left(1 + e^{(1-2y_i)\boldsymbol{w}\cdot\boldsymbol{x}_i}\right) \right] \overset{\text{set}}{=} 0$$

Try to find the derivate and solve. Get as far as you can. Do so on a separate page

(h) [easy] If you attempted the last problem, you found that there is no closed form solution. What type of methods are used to approximate $\boldsymbol{b}$? Note: once you use such methods and arrive at a $\boldsymbol{b}$, that is called "running a logistic regression".

(i) [easy] In class we used the notation $\hat{p} = g_{pr}$. Why?

(j) [easy] Write down $\hat{p}$ as a function of $\boldsymbol{b}$ and $\boldsymbol{x}$.

(k) [harder] What is the interpration of the linear component $\boldsymbol{b} \cdot \boldsymbol{x}$? What does it mean for $\hat{p}$? No need to give the full, careful interpretation.

(l) [difficult] How does one go about *validating* a logistic regression model? What is the fundamental problem with doing so that you didn't have to face with regression or classification? Discuss.

Problem 4

This question is about probabilistic classification i.e. using probability estimation to classify. We limit our discussion to binary classification.

(a) [easy] How do you use a probability estimation model to classify. Provide the formula which provides $\hat{y}(\hat{p})$ i.e. the estimate of whether the event of interest occurs as a function of the probability estimate of the event occuring. Use the "default" rule.

(b) [easy] In the formula from (a), there is an option to be made, write the formula again below with this option denoted p_{th}.

(c) [harder] What happens when p_{th} is low and what happens when p_{th} is high? What is the tradeoff being made?

(d) [difficult] Below is the first 20 rows of in-sample prediction results from a logistic regression whose reponse is $> 50K$ (the positive class) or $\leq 50K$ (the negative class). You have the $\hat{p}_i$'s and the y_i's. Create a performance table that includes the four numbers in the confusion table as well as FPR and recall. Leave some room for one

additional column we will compute later in the question. The rows in the table should be indexed by $p_{th} \in \{0, 0.2, \ldots, 0.8, 1\}$ which you should use as the first column. Hint: you may want to sort by $\hat{p}$ and convert y to binary before you begin.

$\hat{p}$	y
0.35	>50K
0.49	>50K
0.73	>50K
0.91	>50K
0.01	<=50K
0.59	>50K
0.08	<=50K
0.07	<=50K
0.01	<=50K
0.76	>50K
0.32	<=50K
0.07	>50K
0.01	<=50K
0.00	<=50K
0.35	>50K
0.69	>50K
0.38	<=50K
0.07	<=50K
0.02	<=50K
0.00	<=50K

(e) [harder] Using the performance table from (d), trace out an approximate ROC curve.

(f) [harder] Using the performance table from (d), trace out an approximate DET curve.

(g) [easy] Consider the $c_{FP} = \$5$ and $c_{FN} = \$1,000$. Explain how you would find the probabilistic classifier model that minimizes cost among the p_{th} values you considered in your performance table in (d) but do not do any computations.

Problem 5

These are questions related to bias-variance decomposition, bagging and random forests.

(a) [easy] List the assumptions for the bias-variance decomposition.

(b) [harder] Why is $f(\boldsymbol{x})$ called the "conditional expectation function"?

(c) [easy] Provide an expression for the bias-variance decomposition formula for the average MSE over the distribution $\mathbb{P}(\boldsymbol{X})$ for $y = g + (f - g) + \delta$. You should have three terms in the expression. Make sure you explain conceptually each term in English.

(d) [E.C.] Rederive the bias-variance decomposition formula for the average MSE over the distribution $\mathbb{P}(\boldsymbol{X})$ for $y = g + (h^* - g) + (f - h^*) + \delta$. You should group the final expression into *four* terms where two will be the same as the expression found in (c), one will be similar to a term found in (c) and one will be new. Make sure you explain conceptually each term in English. Do so on an additional page.

(e) [harder] Assume a $\mathbb{D}$ where n is large and p is small and you fit a linear model g to all features. Your in-sample R^2 is low. In the expression from (c), indicate term(s) are likely large, which term(s) are likely small and explain why.

(f) [harder] Assume a $\mathbb{D}$ where n is large and p is small and you fit a tree model g to all features. Your in-sample R^2 is low. In the expression from (c), indicate term(s) are likely large, which term(s) are likely small and explain why.

(g) [easy] Provide an expression for the bias-variance decomposition formula for the average MSE over the distribution $\mathbb{P}(\boldsymbol{X})$ for $y = g + (f - g) + \delta$ where g now represents the average taken over constituent models $g_1, g_2, \ldots, g_T$. (This is known as "model averaging" or "ensemble learning"). You can assume that $\rho := \text{Corr}\left[g_{t_1}, g_{t_2}\right]$ is the same for all $t_1 \neq t_2$.

(h) [easy] If $T \to \infty$, rewrite the bias-variance decomposition you found in (k).

(i) [easy] If $g_1, g_2, \ldots, g_T$ are built with the same data $\mathbb{D}$ and $\mathcal{A}$ is not random, then $g_1 = g_2 = \ldots = g_T$. What would ρ be in this case?

(j) [easy] Even though each of the constituent models $g_1, g_2, \ldots, g_T$ are built with the same data $\mathbb{D}$, what idea can you use to induce $\rho < 1$? This idea is called "bagging" which is a whimsical portmanteau of the words "bootstrap aggregation".

(k) [easy] Explain how examining predictions averaged on the out of bag (oob) data for each $g_1, g_2, \ldots, g_T$ can constitute model validation for the bagged model.

(l) [easy] Explain how the Random Forests® algorithm differs from the CART (classification and regression trees) algorithm.

(m) [easy] Explain why the MSE for the Random Forests® algorithm expected to be better than a bag of CART models.

(n) [easy] List the three major advantages of Random Forests® for supervised learning / machine learning.

Problem 6

These are questions related to correlation, causation and the interpretation of coefficients in linear models / logistic regression.

(a) [easy] You are provided with the responses measured from a phenomenon of interest $y_1, \ldots, y_n$ and associated measurements $x_1, \ldots, x_n$ where n is large. The sample correlation is estimated to be $r = 0.74$. Is $\boldsymbol{x}$ "correlated" with $\boldsymbol{y}$?

(b) [harder] Consider the case in (a), would $\boldsymbol{x}$ be a "causal" factor for $\boldsymbol{y}$? Explain.

(c) [harder] Consider the case in (a) and create two plausible causal models using the graphical depiction style used in class (nodes representing variables and lines represent causal contribution where node A below node B means node A is measured before node B). Your model has to include x and y but is not limited to only those variables.

(d) [harder] Consider the case in (a) but now n is small. Create a third plausible causal model (in addition to the two you created in the last problem) using the same graphical depiction style. Your model has to include x and y but is not limited to only those variables.

(e) [easy] Explain briefly how you would prove beyond a reasonable doubt that $\boldsymbol{x}$ is not only correlated with $\boldsymbol{y}$ but that $\boldsymbol{x}$ is a causal factor of $\boldsymbol{y}$.

(f) [easy] Consider $\boldsymbol{x}$ is college GPA and $\boldsymbol{y}$ is career average income. Is $\boldsymbol{x}$ correlated with $\boldsymbol{y}$? Do not lookup data online, I want you to answer conceptually using your own argument.

(g) [harder] Consider $\boldsymbol{x}$ is college GPA and $\boldsymbol{y}$ is career average income. Is $\boldsymbol{x}$ a causal factor of $\boldsymbol{y}$? Do not lookup data online, I want you to answer conceptually using your own argument.

(h) [harder] Consider $\boldsymbol{x}$ is college GPA and $\boldsymbol{y}$ is career average income. Can you think of a $\boldsymbol{z}$ which is a lurking variable? Explain the variable and why you believe it fits the description of a lurking variable.

(i) [harder] If you fit a linear model for $\boldsymbol{y}$, $g = b_0 + b_x x + b_z z$, what would the b_x value be close to? Why?

(j) [E.C.] Create a causal model using the same graphical depiction style that justifies the four linear regression assumptions. Do so on a different page.

(k) [harder] When running a regression of `price` on all variables in the `diamonds` dataset, the coefficient for `carat` is about \$6,500. Interpret this value as best as you can.

(l) [harder] When running a logistic regression of class `malignant` on all variables in the `biopsy` dataset, the coefficient for `V1` (which measures clump thickness) is about 0.54. Interpret this value as best as you can.