MATH 390.4 / 650.2 Spring 2018 Homework #5t

Professor Adam Kapelner

Due 11:59PM Monday, May 14, 2018 under the door of KY604

(this document last updated Thursday 10th May, 2018 at 7:17pm)

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. This is your responsibility to supplement in-class with your own readings.

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 Finlay's book, chapters ... For all parts in this question, answer using notation from class (i.e. $t, f, g, h^*, \delta, \epsilon, e, t, z_1, \ldots, z_t, \mathbb{D}, \mathcal{H}, \mathcal{A}, \mathcal{X}, \mathcal{Y}, X, y, n, p, x_{\cdot 1}, \ldots, x_{\cdot p}$, $x_{1\cdot}, \ldots, x_{n\cdot}$, $f_{pr}, h^*_{pr}, g_{pr}, p_{th}$, and terminology about correlation and causation.

(a) [easy] ...

Problem 2

This question is about probability estimation. We limit our discussion to estimating one event.

(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 $\boldsymbol{w} \cdot \boldsymbol{x}$. Why do we want our 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.

(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 $\boldsymbol{b} \cdot \boldsymbol{x}$?

(l) [difficult] Using your answer to (k) and what we learned about causality in lectures 24 and 25, what is the proper interpretation of the jth component of $\boldsymbol{b}$ i.e. b_j?

(m) [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) [harder] Consider the $c_{FP} = \$5$ and $c_{FN} = \$1,000$. Find the probabilistic classifier model that minimizes cost among the p_{th} values you considered in your performance table in (d) and write it explicitly below.

Problem 5

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

(a) [easy] ...

Problem 6

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

(a) [easy] ...