MATH 390.4 / 650.2 Spring 2018 Homework #5t

Professor Adam Kapelner

Tuesday 22nd May, 2018

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.

Finlay introduces predictive analytics with a credit scoring model. A persons credit score is calculated from a scorecard of features. Each feature has a range levels. For example, in the employment status feature, you add 28 points to your score if you have a full time job. However, if you're unemployed, you must subtract 42 points from your score. A high credit score indicates that a person is likely to pay back their loan.

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

A credit score of 700 corresponds to 1024:1 odds. This implies that out of 1025 people who take out a loan with a credit score of 700, 1024 will repay their loans. In other words, a person with a credit score of 700 has approximately a 0.1% of defaulting.

(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?

Everyone starts with a score of 670 according to the table. Someone if 9 or more credit cards subtracts 18, bringing their score down to 652≈224:1 odds. Someone with 4-8 cards adds 0 to their score, remaining at 670≈320:1 odds. So a person with +9 credits is approxiemetly 30% more likely to defualt than someone with 4-8 credit cards.

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

According to Finlay there are 4 important features of 'Big Data'

(a) Volume-usually contains at least one terabyte of data
(b) Variety-structured/unstrcutred, text, numbers..
(c) Volatility-usually dynamic data, not static
(d) Multi-sourced-data is gathered from all over.

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?

The outputs in regression are assigned real values, $\boldsymbol{y} \in \mathbb{R}$. Outputs in probability estimation are assigned probabilities, $\hat{p} \in (0, 1)$

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

Probability estimation is more similar to classification than regression. Most likely, we will take the probability estimates calculated from the model to sort our data into classes: it will happen, it might happen, it wont happen.

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

Y is predetermined, it is either a 0 or 1. We can never be certain about what Y truly is, we can only estimate it. Bernoulli random variable can be used as a binary classifier for Y, since it returns one of two values based on the $\hat{p}$ input.

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

No. We can only estimate a random variable so the error is implied when we use $\approx$ instead of $=$.

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

f in the regression framework assigns a real value to its response. f_{pr} assigns the probability that its inputs will be a 1 .

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

$$t_{pr}(z_1, \cdots, z_t) = t(z_1, \cdots, z_t)$$

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

$$P(Y_1, \cdots, Y_n) = \prod_{i=1}^{n} f_{pr}(\vec{x_i})(1 - f_{pr}(\vec{x_i}))$$

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

We assumed that $Y_1, \cdots, Y_n$ were independent. If we didn't make this assumption, then we don't know the dependence structure of $Y_1, \cdots, Y_n$ and cannot use the Π and neat formula.

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

No, f_{pr} is not knowable.

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?

These functions can take on values between 0 and 1. The values can never be exactly 0 or 1 because we can never be 100% sure.

(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?

It's monotonically increasing and smooth, just like probabilities.

(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?

Link Function

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

$$\phi(u) = \frac{e^u}{1+e^u} = \frac{1}{1+e^{-u}} \tag{1}$$

$$\phi(u) = 1 - e^{-e^u} \tag{2}$$

$$\phi(u) = \tanh(u) = \frac{e^u - e^{-u}}{e^u + e^{-u}} \tag{3}$$

(1)Logistic (2)Complementary log-log (3) Hyperbolic tangent

(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.

$$P(Y_1, \cdots, Y_n) = \prod_{i=1}^{n} (\frac{\boldsymbol{e}^{\vec{w}\cdot\vec{x}}}{1+\boldsymbol{e}^{\vec{w}\cdot\vec{x}}})^{y_i} (1 - \frac{\boldsymbol{e}^{\vec{w}\cdot\vec{x}}}{1+\boldsymbol{e}^{\vec{w}\cdot\vec{x}}})^{1-y_i}$$

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

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

$$argmax\{\prod_{i=1}^{n} (\frac{1}{1+\boldsymbol{e}^{-\boldsymbol{w}\cdot\boldsymbol{x}_i}})^{y_i} (\frac{1}{1+\boldsymbol{e}^{\boldsymbol{w}\cdot\boldsymbol{x}_i}})^{1-y_i}\}$$

Then if

$$y_i = 1 \Rightarrow \prod_{i=1}^{n} (1+\boldsymbol{e}^{-\boldsymbol{w}\cdot\boldsymbol{x}_i})^{-1} \qquad y_i = 0 \Rightarrow \prod_{i=1}^{n} (1+\boldsymbol{e}^{\boldsymbol{w}\cdot\boldsymbol{x}_i})^{-1}$$

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

To get rid of the negative sign we can take the argmin instead of the argmax

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

(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".

Numerical methods such as gradient descent.

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

Since g_{pr} is what we are using to approximate f_{pr}, it also returns a probability.

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

$$\hat{p} = (1 + e^{-\boldsymbol{b}\cdot\boldsymbol{x}})^{-1}$$

(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.

$$\boldsymbol{b} \cdot \boldsymbol{x} = \ln\left(\frac{\hat{p}}{1-\hat{p}}\right)$$

It represents logodds$(\boldsymbol{Y}|\boldsymbol{x})$

(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 occurring. Use the "default" rule.

$$\hat{y} = \mathbb{1}_{\hat{p} \leq 0.5}$$

(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}.

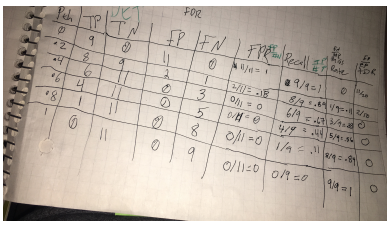
$$\hat{y}_i = \mathbb{1}_{\hat{p}_i \leq p_{th}}$$

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

The trade off will be between FP and FN. Low p_{th} implies a high rate of False Positives and low rate of False Negatives. High p_{th} implies high rate of False Negatives and low rate of False Positives.

(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

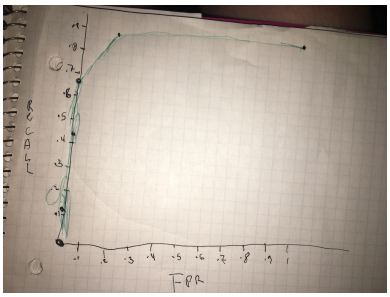


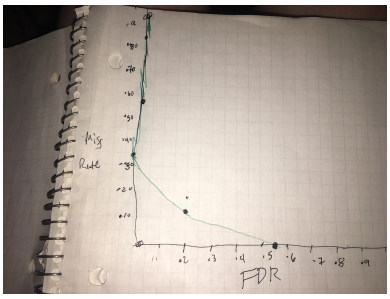
(e) [harder] Using the performance table from (d), trace out an approximate ROC curve.

(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.

Use the performance table to find the p_{th} value with the lowest FN rate, that also has a reasonably low FP rate.





Problem 5

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

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

$$E[Y|X = x] = f(x) \tag{1}$$
$$Var[\Delta|X = x] = Var[\Delta] = \sigma^2 \tag{2}$$

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

It approximates the expected value of the function Y given x

(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.

$$MSE = \sigma^2 + E_{\mathcal{X}}[Var[g(\vec{x})]] + E_{\mathcal{X}}[Bias[g(\vec{x})]^2]$$

σ^2 is the irreducible error. $E_{\mathcal{X}}[Bias[g(\vec{x})]^2]$ is how far off the g is from f, on average. $E_{\mathcal{X}}[Var[g(\vec{x})]]$ is how much

(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.

Variance will be small due to large n and the rest of the reducible error will be in the Bias term.

(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.

Trees keep bias low but have very high variance from tree to tree.

(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 $\rho := \text{Corr}[g_{t_1}, g_{t_2}]$ is the same for all $t_1 \neq t_2$.

$$MSE = \sigma^2 + \rho Var[g_t] + \frac{1-\rho}{T} Var[g_t] + Bias[g_t]^2$$

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

$$\sigma^2 + \rho Var[g_t] + Bias[g_t]^2$$

(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?

$$\rho = 1$$

(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".

We can build each $\mathbb{D}_{train_t}, \mathbb{D}_{test_t}$ by sampling $\mathbb{D}$ with replacement. g_{bag} is then the average of all models $g_1, \cdots, g_T$.

(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.

We can validate each model individually by testing it on the data that was left "out of the bag" when we were training the model. We can average all of these 'oob' error statistics together to get oob_{bag} and use this to validate g_{bag}

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

The CART algorithm uses all features/splits, Random Forests use a randomized subset of features.?

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

It's more random.

(n) [easy] List the three major advantages of Random Forests® for supervised learning / machine learning.
You get something for nothing

Problem 6

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

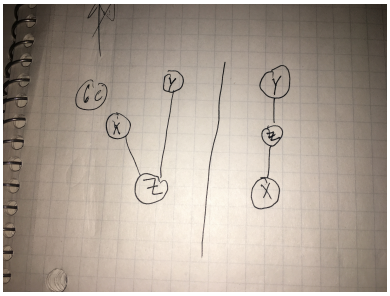
(a) [easy] You are provided with the responses measured from a phenomenon of interest $y_1, \ldots, y_n$ and associated measurements $x_1, \ldots, x_n$ where n is large. The sample correlation is estimated to be $r = 0.74$. Is $\boldsymbol{x}$ "correlated" with $\boldsymbol{y}$?

Yes?

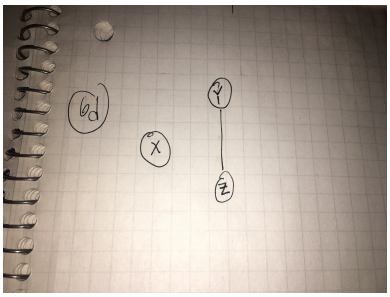
(b) [harder] Consider the case in (a), would $\boldsymbol{x}$ be a "causal" factor for $\boldsymbol{y}$? Explain.

It's impossible to tell.

(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 $\boldsymbol{y}$.

We can keep everything constant in the system, manipulate x and if we see a chance in $\boldsymbol{y}$, then we can say 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.

x is slightly correlated with GPA. Most universities have a minimum threshold GPA score students must maintain in order to graduate. If a students GPA is below this threshold, even though the score is > 0, the student might not be allowed to graduate. A person with a GPA of 1.0 indicates more laziness than a person GPA of 0.0. There could be a person who never went to college, so their GPA=0, but they have a learned trade(plumbing). The plumber probably makes more money than the college dropout.

(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.

No,if we kept everything constant in the system and increaed a persons GPA by a certain amount, their average income would remain the same.

(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.

Personality type. Having a high GPA is not about IQ, you must care about certain things and follow certain rules(handing in assignments on time, attendance). The drive behind a persons desire to achieve and MAINTAINE a high GPA won't dissapear after they graduate college. They will likely still feel the need to perform well in the eyes of others, and take an office job somewhere to wither and die.

(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.

For two diamonds A and B that were observed in the same way as the diamonds dataset, if A has a value of carat one unit larger than the carat value of B, then the price of diamond A is predicted to be \$6500 more than the price of diamond B.

(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.

For two observations A and B that were sampled in the same way as the data in the biopsy data set, if A's clump thickness is greater than Bs by a value equal to V1, then A is .54 times more likely to be malignant than B,