MATH 390.4 / 650.2 Spring 2018 Homework #5t

Adina Bechhofer

Friday 18th 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.

Classification of "good" and "bad" lenders. He makes a point that historical data can be used to determine credit- worthiness of a person; whether or not they're likely to return a 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 is achieved by starting from a baseline of 670 and adding or subtracting to it to account for important factors such as employment and age. Out of 1,025 people with credit score of 700, only one is expected to default.

(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?

People with 4-8 credit cards are evaluated starting from the baseline of 670, in which 1 of 430 people is expected to default. People with 9 or more cards start from 18 points below the baseline, in which 1 out of 100 people is expected to default. This means that people with 9 cards or more are more than 4 times as likely to default as people with only 4-8 cards.

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

This term is used to describe the new kind of data. The new data sets are way too large to be interpreted by a human. It is an enormous trove of data that has low density of useful information . This means that in order to effectively use big data, one needs to use a lot of computing power and very smart algorithms.

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?

In regression, the response $\boldsymbol{y} \in \mathbb{R}$
In probability estimation, the response $p \in (0, 1)$

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

Regression, because we want to return a continuous value $\in (0, 1)$ for the probability estimation.

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

We don't know ahead of time what the response will be, therefore we think of it as a R.V. It needs to take on values of 0 and 1, so Bernoulli is appropriate for that.

(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, the δ is implicit in the idea that Y is a r.v.

(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?

yes, there is a t_{pr}; it outputs values in $\{0, 1\}$. There's no way of knowing what it looks like.

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

$$P\big(y_1, y_2, \ldots, y_n | x\big) = \prod_{i=1}^{n} P\big(y_i | x_i\big)$$

$$= \prod_{i=1}^{n} f_{pr}\big(x_i\big)^{y_i} \big(1 - f_{pr}\big(x\big)\big)^{1-y_i}$$

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

That the Y's are independent. If this assumption fails, we cannot express likelihood as a product.

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

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?

They return a probability estimate $\in (0, 1)$.
We don't want to return a 0, or 1 because we're never that sure about anything.

(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?

We like linear models because they're simple and monotonically increasing. This means that as the x's increase our probability increases. Also, the linear term is the log odds.

(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?

The link function, or the squishy function.

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

Logistic wing, probit (Inverse CDF of the normal), complementary log - log, 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.

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

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

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

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

$$= \begin{cases} (1 + e^{-\boldsymbol{w} \cdot x_i})^{-1} & \text{if } y_i = 1 \\ (1 + e^{\boldsymbol{x} \cdot x_i})^{-1} & \text{if } y_i = 0 \end{cases}$$

$$= \prod_{i=1}^{n} \left(1 + e^{(1-2y_i)\boldsymbol{w}\cdot x_i}\right)^{-1}$$

If we wish to minimize the likelihood, we can also minimize its log, because log is monotonic.

$$\arg\min \left\{ \ln \left(\prod_{i=1}^{n} \left(1 + e^{(1-2y_i)\boldsymbol{w}\cdot x_i}\right)^{-1} \right) \right\}$$

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

(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.

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

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".

Use numerical methods such as gradient descent.

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

The function g_{pr} which estimates f_{pr}, gives us a probability estimate rather than a definite answer.

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

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

(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.

The linear component $\boldsymbol{b}\cdot\boldsymbol{x}$ is the log odds $\left(\frac{\hat{p}}{1-\hat{p}}\right)$.

A 0 log odds corresponds to a 50% chance. Very large positive log- odds represent probabilities close to 1, very large negative log- odds represent probabilities close to 0.

(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.

In regression, we have y's to check the predictions against. In the logistic regression framework, we are given binary y's, but the prediction is a continuous probability estimate.
to validate we use scoring. 2 popular methods are:

1. Log scoring rule: $s_i = y_i ln(\hat{p}) + (1 - y_i) ln(1 - \hat{p})$
   Evaluate over $\frac{1}{n} \sum s_i$ = ave score.
2. Brier score: $s_i = -(y_i - \hat{p})^2$

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.

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

(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}.

$$\hat{\boldsymbol{y}} = \mathbb{1}_{\hat{p} \geq p_{th}}$$

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

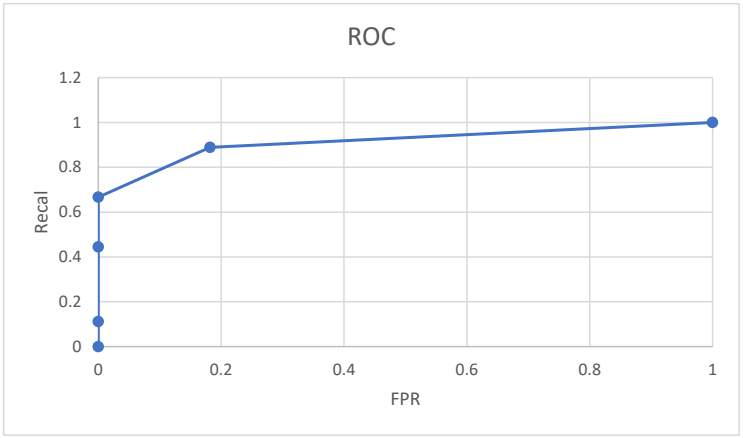
When p_{th} is low, many $\hat{\boldsymbol{y}}$'s will be 1's and only few will be 0. When p_{th} is high, many $\hat{\boldsymbol{y}}$'s will be 0's and only few will be 1.
We trade false positives for false negatives when we raise p_{th}

(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.

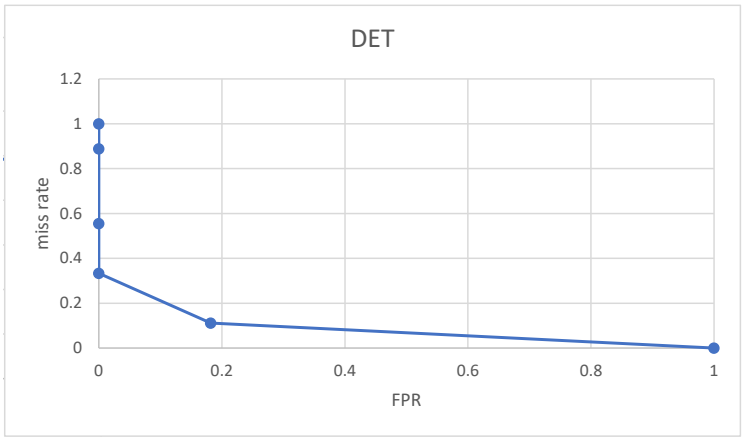
$\hat{p}$	y
0.35	>50K
0.49	>50K
0.73	>50K
0.91	>50K
0.01	<=50K
0.59	>50K
0.08	<=50K
0.07	<=50K
0.01	<=50K
0.76	>50K
0.32	<=50K
0.07	>50K
0.01	<=50K
0.00	<=50K
0.35	>50K
0.69	>50K
0.38	<=50K
0.07	<=50K
0.02	<=50K
0.00	<=50K

pth	TN	FP	FN	TP	FPR	Recall	miss rate
0	0	11	0	9	1	1	0
0.2	9	2	1	8	0.181818	0.888889	0.111111
0.4	11	0	3	6	0	0.666667	0.333333
0.6	11	0	5	4	0	0.444444	0.555556
0.8	11	0	8	1	0	0.111111	0.888889
1	11	0	9	0	0	0	1

(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.

Add an additional column of cost to the table. For each p_{th}, calculate the total cost as $C_{FP} \times FP + C_{FN} \times FN$. Select the p_{th} that minimizes the cost.

Problem 5

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

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

1.
$$E[Y|X = x] = f(x)$$

2.
$$Var(\Delta|X = x) = Var(\Delta) = \sigma^2$$

X has no effect on the variance of the irreducible error.

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

It's the expectation of Y conditional on X (by assumption 1).

(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.

$MSE = \sigma^2 + \mathbb{E}[var(g(x))] + \mathbb{E}[Bias(g(x))]^2$
σ^2 irreducible error due to ignorance.
$\mathbb{E}[var(g(x))]$ is the estimation error.
$\mathbb{E}[Bias(g(x))]^2$ misspecification error.

(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.

Since n is very large compared to p, the fitted lines will be very similar, and variance will be very small. Therefore, most 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.

Since trees are likely to overfit, the variance will be very large. However, the bias will be small since they capture complexity very well.

(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 $t_1 \neq t_2$.

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

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

$MSE = \sigma^2 + Bias[g]^2 + \rho Var[g]$

(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".

Bootstrap sampling. For this, for each g_i, sample n data points with replacement, and train the tree on that.

(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.

For each observation i we compute the average prediction of all the trees g_j that haven't had i in their "in bag". Since each of those models never saw i before, the prediction will be an OOS prediction.

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

In the CART algorithm, we use all of the data and find the splits that minimize the total error. In the Random Forests® algorithm, we create many trees where for each one, we randomly omit some features and use CART to train in on all the observations. This forces the splits for every tree to be different and random. Overall, we get trees that are much less correlated.

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

As shown above,
$MSE = \sigma^2 + Bias[g]^2 + \rho Var[g] + \frac{1-\rho}{T} Var[g]$
Randomly omitting features makes the trees less correlated, which brings down ρ.

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

Low bias and minimum variance.

Problem 6

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

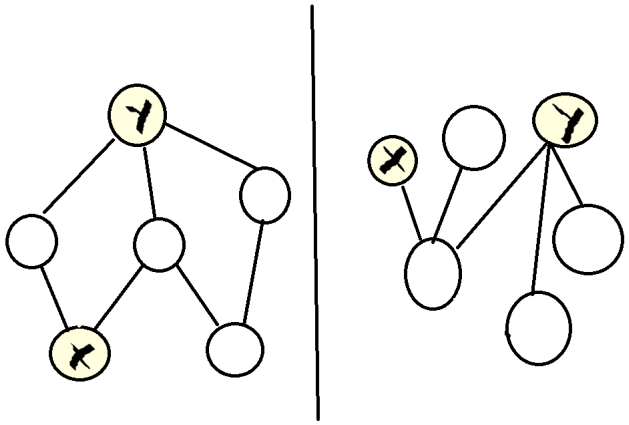
(a) [easy] You are provided with the responses measured from a phenomenon of interest $y_1, \ldots, y_n$ and associated measurements $x_1, \ldots, x_n$ where n is large. The sample correlation is estimated to be $r = 0.74$. Is $\boldsymbol{x}$ "correlated" with $\boldsymbol{y}$?

Yes, $r \neq 0$. If n is large enough, then you know that it is a real correlation.

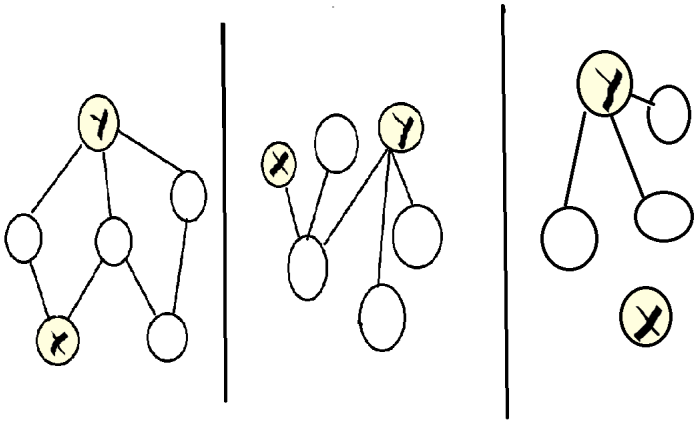
(b) [harder] Consider the case in (a), would $\boldsymbol{x}$ be a "causal" factor for $\boldsymbol{y}$? Explain.

No way to tell. Since the measurements were "naturally" observed, nothing can be said about causation.

(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 $\boldsymbol{y}$.

Create an experiment where all other nodes are held constant, manipulate x, and measure changes in 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.

Yes, for students with higher GPA you'd expect to see higher average income.

(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, GPA is probably a result of intelligence and dedication, which are causal factors for are average income.

(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.

The lurking variables are intelligence and dedication. They're lurking because they're unobservable. If observed, they will change the $cor(x, y)$.

(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?

b_x's value will be close to 0. This is because z is observed, so the $cor(x, y)$ goes to 0.

(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 2 diamonds A and B, sampled the same way as the data in $\mathbb{D}$ with the same color, clarity, cut, x, y, z, and table measures, and diamond A is measured to be one carat more than diamond B. Then, diamond A is predicted to cost on average \$6,500 more than 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 2 biopsies A and B, sampled the same way as the data in $\mathbb{R}$, that have all other features equal, and the clump thickness of A is one unit higher than B. Then, A is expected to have log- odds of being malignant which is 0.54 higher than B's log - odds of being malignant.