MATH 390.4 / 650.2 Spring 2018 Homework #3t

Professor Adam Kapelner

Due 11:59PM Monday, May 7, 2018 under the door of KY604

(this document last updated Sunday 29th April, 2018 at 5:08pm)

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. multivariate least squares linear modeling, orthogonal projections, non-linear linear regression, stepwise linear regression, non-parametric regression, regression trees, classification trees, performance characteristics of binary classification, etc. This is your responsibility to supplement in-class with your own readings. Also, read ch 3–6 in Silver.

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 Silver's book, chapters ... For all parts in this question, answer using notation from class (i.e. $t, f, g, h^*, \delta, \epsilon, e, t, z_1, \ldots, z_t, \mathbb{D}, \mathcal{H}, \mathcal{A}, \mathcal{X}, \mathcal{Y}, X, y, n, p, x_{\cdot 1}, \ldots, x_{\cdot p}$, $x_{1\cdot}, \ldots, x_{n\cdot}$, etc. and also we now have $f_{pr}, h^*_{pr}, g_{pr}, p_{th}$, etc from probabilistic classification as well as different types of validation schemes).

(a) [easy] What algorithm that we studied in class is PECOTA most similar to?

(b) [easy] Is baseball performance as a function of age a linear model? Discuss.

(c) [harder] How can baseball scouts do better than a prediction system like PECOTA?

(d) [harder] Why hasn't anyone (at the time of the writing of Silver's book) taken advantage of Pitch f/x data to predict future success?

(e) [difficult] Chapter 4 is all about predicting weather. Broadly speaking, what is the problem with weather predictions? Make sure you use the framework and notation from class. This is not an easy question and we will discuss in class. Do your best.

(f) [easy] Why does the weatherman lie about the chance of rain? And where should you go if you want honest forecasts?

(g) [difficult] Chapter 5 is all about predicting earthquakes. Broadly speaking, what is the problem with earthquake predictions? It is *not* the same as the problem of predicting weather. Read page 162 a few times. Make sure you use the framework and notation from class.

(h) [easy] Silver has quite a whimsical explanation of overfitting on page 163 but it is really educational! What is the nonsense predictor in the model he describes?

(i) [easy] John von Neumann was credited with saying that "with four parameters I can fit an elephant and with five I can make him wiggle his trunk". What did he mean by that and what is the message to you, the budding data scientist?

(j) [difficult] Chapter 6 is all about predicting unemployment, an index of macroeconomic performance of a country. Broadly speaking, what is the problem with unemployment predictions? It is *not* the same as the problem of predicting weather or earthquakes. Make sure you use the framework and notation from class.

(k) [E.C.] Many times in this chapter Silver says something on the order of "you need to have theories about how things function in order to make good predictions." Do you agree? Discuss.

Problem 2

This question is about validation for the supervised learning problem with one fixed $\mathbb{D}$.

(a) [easy] For one fixed $\mathcal{H}$ and $\mathcal{A}$ (i.e. one model), write below the steps to do a simple validation and include the final step which is shipping the final g.

(b) [easy] For one fixed $\mathcal{H}$ and $\mathcal{A}$ (i.e. one model), write below the steps to do a K-fold cross validation and include the final step which is shipping the final g.

(c) [harder] For one fixed $\mathcal{H}$ and $\mathcal{A}$ (i.e. one model), write below the steps to do a bootstrap validation and include the final step which is shipping the final g.

(d) [harder] For one fixed $\mathcal{H}_1, \ldots \mathcal{H}_M$ and $\mathcal{A}$ (i.e. M different models), write below the steps to do a simple validation and include the final step which is shipping the final g.

(e) [difficult] For one fixed $\mathcal{H}_1, \dots \mathcal{H}_M$ and $\mathcal{A}$ (i.e. M different models), write below the steps to do a K-fold cross validation and include the final step which is shipping the final g. This is not an easy problem! There are a lot of steps and a lot to keep track of...

Problem 3

This question is about ridge regression — an alternative to OLS.

(a) [harder] Imagine we are in the "Luis situation" where we have $\boldsymbol{X}$ with dimension $n \times (p+1)$ but $p+1 > n$ and we still want to do OLS. Why would the OLS solution we found previously break down in this case?

(b) [harder] We will embark now to provide a solution for this case. The solution will also give nice results for other situations besides the Luis situation as well. First, assume λ is a positive constant and demonstrate that the expression $\lambda \left|\left|\boldsymbol{w}\right|\right|^2 = \boldsymbol{w}^\top(\lambda \boldsymbol{I})\boldsymbol{w}$ i.e. it can be expressed as a quadratic form where $\lambda \boldsymbol{I}$ is the determining matrix. We will call this term $\lambda \left|\left|\boldsymbol{w}\right|\right|^2$ the "ridge penalty".

(c) [easy] Write the $\mathcal{H}$ for OLS below where there parameter is the $\boldsymbol{w}$ vector. $\boldsymbol{w} \in$?

(d) [easy] Write the error objective function that OLS minimizes using vectors, then expand the terms similar to the previous homework assignment.

(e) [easy] Now add the ridge penalty $\lambda \left|\left|\boldsymbol{w}\right|\right|^2$ to the expanded form you just found and write it below. We will term this two-part error function the "ridge objective".

(f) [easy] Note that the ridge objective looks a bit like the hinge loss we spoke about when we were learning about support vector machines. There are two pieces of this error function in counterbalance. When this is minimized, describe conceptually what is going on.

(g) [harder] Now, the ridge penalty term as a quadratic form can be combined with the last term in the least squares error from OLS. Do this, then use the rules of vector derivatives we learned to take $d/d\boldsymbol{w}$ and write the answer below.

(h) [easy] Now set that derivative equal to zero. What matrix needs to be invertible to solve?

(i) [difficult] There's a theorem that says *positive definite* matrices are invertible. A matrix is said to be positive definite if every quadratic form is positive for all vectors i.e. if $\forall \boldsymbol{z} \neq \mathbf{0} \quad \boldsymbol{z}^\top A \boldsymbol{z} > 0$ then A is positive definite. Prove this matrix from the previous question is positive definite.

(j) [easy] Now that it's positive definite (and thus invertible), solve for the $\boldsymbol{w}$ that is the argmin of the ridge objective, call it $\boldsymbol{b}_{ridge}$. Note that this is called the "ridge estimator" and computing it is called "ridge regression" and it was invented by Hoerl and Kennard in 1970.

(k) [easy] Did we just figure out a way out of Luis's situation? Explain.

(l) [harder] It turns out in the Luis situation, many of the values of the entries of $\boldsymbol{b}_{ridge}$ are close to 0. Why should that be? Can you explain now conceptually how ridge regression works?

(m) [easy] Find $\hat{\boldsymbol{y}}$ as a function of $\boldsymbol{y}$ using $\boldsymbol{b}_{ridge}$. Is $\hat{\boldsymbol{y}}$ an orthogonal projection of $\boldsymbol{y}$ onto the column space of $\boldsymbol{X}$?

(n) [E.C.] Show that this $\hat{\boldsymbol{y}}$ is an orthogonal projection of $\boldsymbol{y}$ onto the column space of some matrix $\boldsymbol{X}_{ridge}$ (which is not $\boldsymbol{X}$!) and explain how to construct $\boldsymbol{X}_{ridge}$ on a separate page.

(o) [easy] Is the $\mathcal{H}$ for OLS the same as the $\mathcal{H}$ for ridge regression? Yes/no.
Is the $\mathcal{A}$ for OLS the same as the $\mathcal{A}$ for ridge regression? Yes/no.

(p) [harder] What is a good way to pick the value of λ, the hyperparameter of the $\mathcal{A}$ = ridge?

(q) [easy] In classification via $\mathcal{A}$ = support vector machines with hinge loss, how should we pick the value of λ? Hint: same as previous question!

(r) [E.C.] Besides the Luis situation, in what other situations will ridge regression save the day?

(s) [difficult] The ridge penalty is beautiful because you were able to take the derivative and get an analytical solution. Consider the following algorithm:

$$\boldsymbol{b}_{lasso} = \underset{\boldsymbol{w} \in \mathbb{R}^{p+1}}{\arg\min} \left\{ (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) + \lambda ||\boldsymbol{w}||^1 \right\}$$

This penalty is called the "lasso penalty" and it is different from the ridge penalty in that it is not the norm of $\boldsymbol{w}$ squared but just the norm of $\boldsymbol{w}$. It turns out this algorithm (even though it has no closed form analytic solution and must be solved numerically a la the SVM) is very useful! In "lasso regression" the values of $\boldsymbol{b}_{lasso}$ are not shrunk *towards* 0 they are harshly punished *directly to* 0! How do you think lasso regression would be useful in data science? Feel free to look at the Internet and write a few sentences below.

(t) [easy] Is the $\mathcal{H}$ for OLS the same as the $\mathcal{H}$ for lasso regression? Yes/no.
Is the $\mathcal{A}$ for OLS the same as the $\mathcal{A}$ for lasso regression? Yes/no.

Problem 4

These are questions about non-parametric regression.

(a) [easy] In problem 1, we talked about schemes to validate algorithms which tried M different prespecified models. Where did these models come from?

(b) [harder] What is the weakness in using M pre-specified models?

(c) [difficult] Explain the steps clearly in forward stepwise linear regression.

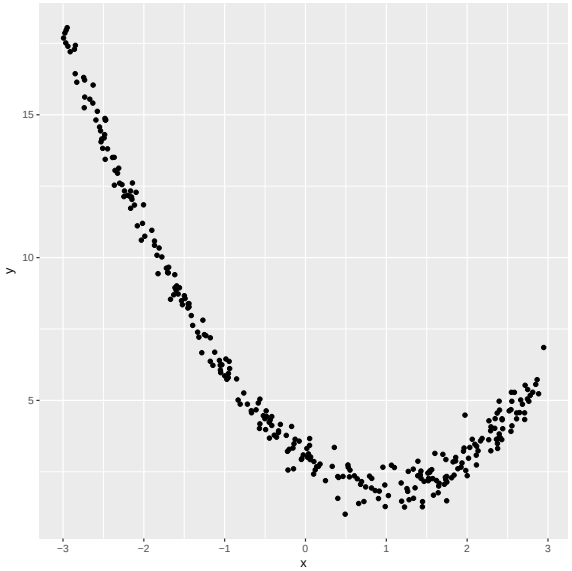
(d) [difficult] Explain the steps clearly in *backwards* stepwise linear regression.

(e) [harder] What is the weakness(es) in this stepwise procedure?

(f) [easy] Define "non-parametric regression". What problem(s) does it solve? What are its goals? Discuss.

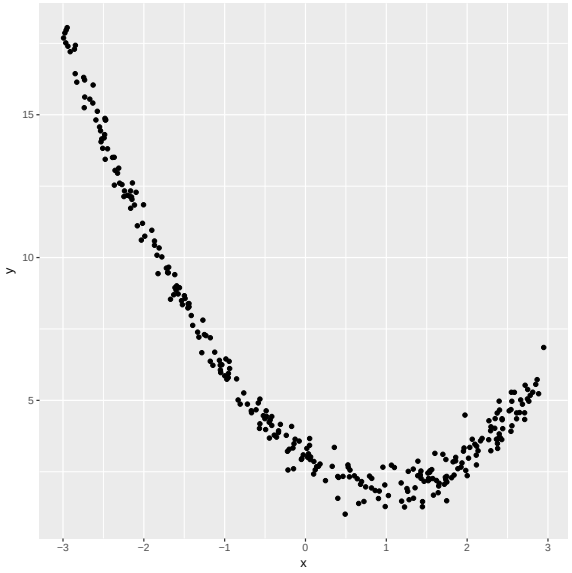
(g) [harder] Provide the steps for the regression tree (the one algorithm we discussed in class) below.

(h) [easy] Consider the following data

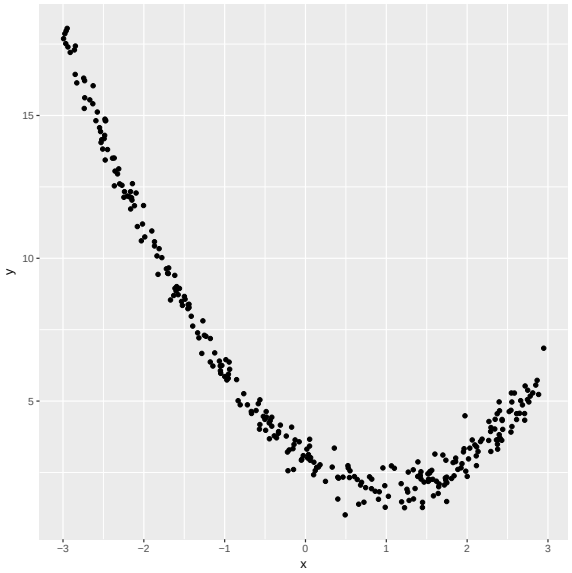


Create a tree with maximum depth 1 (i.e one split at the root node) and plot g above.

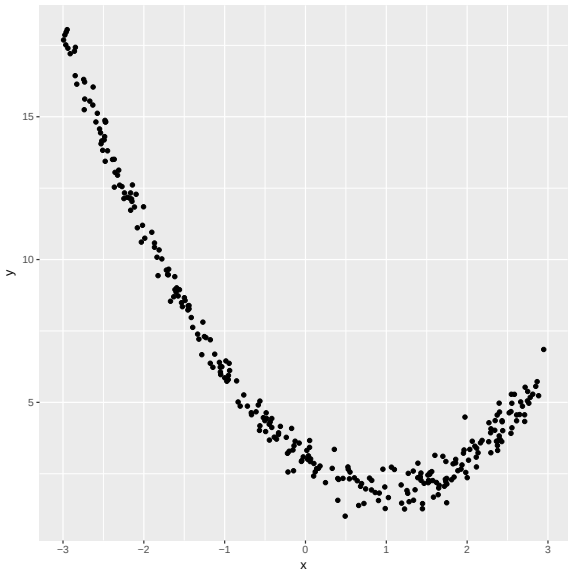
(i) [easy] Now add a second split to the tree and plot g below.



(j) [easy] Now add a third split to the tree and plot g below.

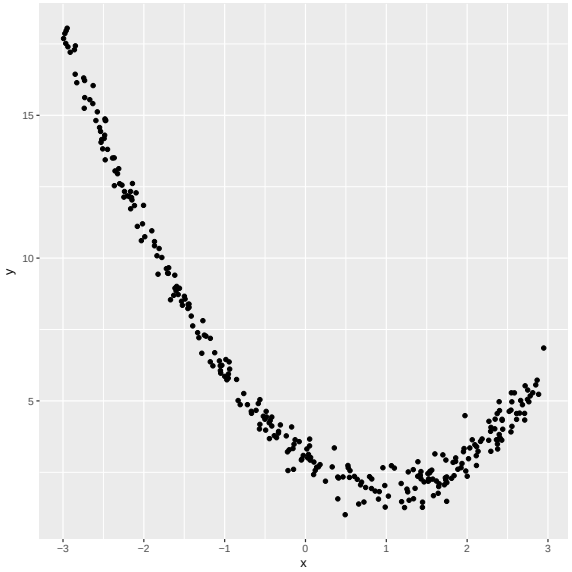


(k) [easy] Now add a fourth split to the tree and plot g below.



(l) [easy] Draw a tree diagram of g below indicating which nodes are the root, inner nodes and leaves. Indicate split rules and leaf values clearly.

(m) [easy] Plot g below for the mature tree with the default $N_0 = \texttt{nodesize}$ hyperparameter.



(n) [easy] If $N_0 = 1$, what would likely go wrong?

(o) [easy] How should you pick the $N_0 = \texttt{nodesize}$ hyperparameter in practice?

Problem 5

These are questions about classification trees.

(a) [easy] How are classification trees different than regression trees?

(b) [harder] What are the steps in the classification tree algorithm?

Problem 6

These are questions about measuring performance of a classifier.

(a) [easy] What is a confusion table?

Consider the following in-sample confusion table where "> 50K" is the positive class:

```
       y_hats_train
y_train <=50K >50K
  <=50K  3475  262
  >50K    471  792
```

(b) [easy] Calculate the following: n (sample size) =

FP (false positives) =

TP (true positives) =

FN (false negatives) =

TN (true negatives) =

$\#P$ (number positive) =

$\#N$ (number negative) =

$\#PP$ (number predicted positive) =

$\#PN$ (number predicted negative) =

$\#P/n$ (prevalence / marginal rate / base rate) =

$(FP + FN)/n$ (misclassification error) =

$(TP + TN)/n$ (accuracy) =

$TP/\#PP$ (precision) =

$TP/\#P$ (recall, sensitivity, true positive rate, TPR) =

$2/(\text{recall}^{-1} + \text{precision}^{-1})$ (F1 score) =

$FP/\#PP$ (false discovery rate, FDR) =

$FP/\#N$ (false positive rate, FPR) =

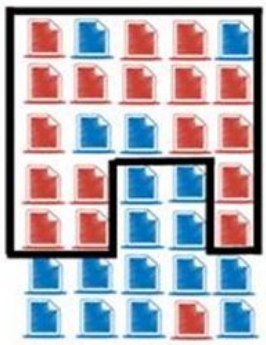
$FN/\#PN$ (false omission rate, FOR) =

$FN/\#P$ (false negative rate, FNR) =

(c) [easy] Why is FPR also called the "false alarm rate"?

(d) [easy] Why is FNR also called the "miss rate"?

(e) [easy] Below let the red icons be the positive class and the blue icons be the negative class.



The icons included inside the black border are those that have $\hat{y} = 1$. Compute both precision and recall.

(f) [harder] There is always a tradeoff of FP vs FN. However, in some situations, you will look at FPR vs. FNR. Describe such a classification scenario. It does not have to be this income amount classification problem, it can be any problem you can think of.

(g) [harder] There is always a tradeoff of FP vs FN. However, in some situations, you will look at FDR vs. FOR. Describe such a classification scenario. It does not have to be this income amount classification problem, it can be any problem you can think of.

(h) [harder] There is always a tradeoff of FP vs FN. However, in some situations, you will look at precision vs. recall. Describe such a classification scenario. It does not have to be this income amount classification problem, it can be any problem you can think of.

(i) [harder] There is always a tradeoff of FP vs FN. However, in some situations, you will look only at an overall metric such as accuracy (or $F1$). Describe such a classification scenario. It does not have to be this income amount classification problem, it can be any problem you can think of.