

CSE 417T: Homework 4

Due: November 8 (Wednesday), 2017 at 10PM

Notes:

- Please check the submission instructions for Gradescope and SVN provided on the course website. You must follow those instructions exactly.
- You have to submit your written solutions (including reports for programming problems) to Gradescope. SVN is used to submit code only. **Write-ups submitted via SVN will not be accepted or graded.**
- Homework is due **by 10 PM on the due date**. Remember that you may not use more than 2 late days on any one homework, and you only have a budget of 5 in total.
- Please keep in mind the collaboration policy as specified in the course syllabus. If you discuss questions with others you **must** write their names on your submission, and if you use any outside resources you **must** reference them. **Do not look at each others' writeups, including code.**
- Please comment your code properly.
- There are 3 problems on 2 pages in this homework.

Problems:

1. (60 points) For this problem, you will be doing LFD Problem 4.4 parts (a) through (d) with some changes / help / instructions / requirements. First, you can find headers for all the code you need to implement in your SVN repository for the class. There is also a matlab script called `run_expts.m` which you can use as an example for how to run your code to return the results we want. **Second, read Problem 4.3 carefully. You can (and will need to) use the recurrence defined there as well as the formula in 4.3(e).**
 - (a) In addition to answering the question about why we need to normalize f , also prove that the term to normalize by is $\sqrt{\sum_{q=0}^Q \frac{1}{2q+1}}$ (hint: use the formula in 4.3(e)).
 - (b) Answer the question. For your implementation, we suggest you use `glmfit` with the additional options `'normal'`, `'constant'`, `'off'`.
 - (c) Answer the question (hint: use the formula in 4.3(e)).
 - (d) Implement the framework and **answer the questions**, with the modification that **you only need to look at** $Q_f \in \{5, 10, 15, 20\}$, $N \in \{40, 80, 120\}$, $\sigma^2 \in \{0, 0.5, 1.0, 1.5, 2.0\}$. Compute **both the median and the mean of the overfit measure** applied to many (at least 500) different datasets for each choice of parameters, and **report how these measures vary as a function of the** complexity of the true hypothesis, the number of training

examples, and the level of stochastic noise (use line graphs). Explain your observations, and also comment on the differences you observe between the mean and median measures.

Here are some potentially useful notes and hints for this:

- You will be graded on your writeup. **Correctness of the code in itself does not count for credit**, but we may look at and examine your code manually if needed.
 - You should use your judgment in selecting which graphs to show in support of your answers and explanations. There are different acceptable ways to do this. For example, you could include 3-6 graphs, selected to show what you think is most interesting/relevant. For each one, you could hold one variable constant, and plot different lines for a second variable, while putting the third one on the X axis. Alternatively you could explore heatmaps/colormaps/colorbars.
 - Do **not** use the Matlab built-in functions related to Legendre polynomials – those compute something different from what we are looking for.
 - You may use or modify `run_expts.m` as you see fit. It's meant to provide an example of how you could do things, not to be the last word on the issue.
2. (20 points) LFD Problem 5.4
3. (20 points) You have been hired by a biologist to learn a decision tree to determine whether a mushroom is poisonous. You have been given the following data:

Color	Stripes	Texture	Poisonous?
Purple	No	Smooth	No
Purple	No	Rough	No
Red	Yes	Smooth	No
Purple	Yes	Rough	Yes
Purple	Yes	Smooth	Yes

Use ID3 to learn a decision tree from the data (this is a written exercise – no need to code it up):

- (a) What is the root attribute of the tree? Show the computations.
- (b) Draw the decision tree obtained using ID3.