

ISTA 410/510 Homework III

For contribution to the final grade, due dates, current late policy, and instructions for handing the assignment in, see the assignment web page.

Create a PDF document with your answers and/or the results of any programs that you write. You should also hand in your programs, but I won't necessarily look at them. Your PDF should be named <first_name>-<last_name>-<assignment>.pdf (e.g., kobus-barnard-hw3.pdf).

You should explain, in your PDF, where the results come from (e.g., "These plots are the result of running the program hw1_part2.m with parameters 1,2,4, and 8 respectively."). Please use this course as an opportunity to learn how to write better figure captions. They should tell the reader how to interpret the figures, and the answer to obvious questions the reader might have which are not readily available from the figure.

For simplicity, problems are generally all worth the same, except ones marked by "+" that are expected to substantively more time consuming, and are worth double. Two "+" means triple value, etc.

Questions marked by * are required for grad students only. They count as challenge problems for undergraduates.

Questions marked by ** are challenge problems for both grads and undergraduates.

Any non-challenge problem can be replaced by challenge problems with collective value is at least that of the problem being replace (e.g., an undergraduate might replace a non starred "+" problem with two "*" problems without "+"). Please make it clear that this is what you are doing (e.g., for a required problem you could answer "see optional problem #3"). The point here is to enable students to avoid problems that they feel are not instructive.

For a complete assignment, undergraduates need to hand in problems that have total value of at least seven. Grad students need five more.

Extra problems (please indicated in your answer when you are doing an extra problem) are eligible for modest extra credit. The maximum score for an assignment will be capped at 120%. The maximum score for all assignments taken together is capped at 65/60.

Hints or answers to many of these problems can be looked up. If you are stuck and make use of a resource, simply make a note of it. For example, you might say that you had a glance at the solution to the same or similar problem solution in a particular source, and then attempted to recreate for yourself. This is better than being completely stuck, or copying the answer blindly.

Mathematical content. This is a mathematical subject and there is a wide variance in backgrounds of students who take this course. For example, there may be problems in the assignments which seem more difficult than they really are simply because you are not used to the kind of problem. In general, I am very willing to give hints, consider other work in exchange, and grade holistically, focused on effort and progress from whatever level you are at. However, this works best if you start the assignment early, and work through it steadily over time, rather than do the last minute thing.

1. (Regression explained using three problems).

(a)

$$\text{Let } y(x) = 1 + 2x + 3x^2$$

Express $y(2)$ as the dot product of two vectors of length 3.

(b)

$$\text{Let } y(x) = 4 + 3x + 2x^2 + 1x^4$$

$$\text{Express: } y_1 = y(0), y_2 = y\left(\frac{1}{2}\right), y_3 = y(1), y_4 = y(2), y_5 = y(3)$$

as a 5×4 matrix times a vectors of length 4.

(c)

Now suppose you have observed values that are assumed to come from a model like that one in (b) at the same x values, specifically $(0, \frac{1}{2}, 1, 2, 3)$, in a vector \mathbf{y} . Let the 5×4 matrix be \mathbf{A} , and the vector of length 4 be \mathbf{w} . Express the sum of the squared error between the estimate and the data using these matrices and vectors for a generic \mathbf{y} . (To connect this to the, next problem, notice that your answer does not depend on the particular polynomial model or data).

As a concrete example, provide the value for the \mathbf{A} from (b) and observed $\mathbf{y} = (3, 6, 12, 32, 120)^T$.

2. Now consider the general case of a polynomial with coefficient vector, \mathbf{w} , where there is no error in the x values, but observed y values are distributed normally around the values predicted by the model (the mean) with some known variance. Assume that the length of \mathbf{w} is given, but that its values are not known. Show that the MLE for \mathbf{w} is the minimum of the sum of squared error (as from the above problem).

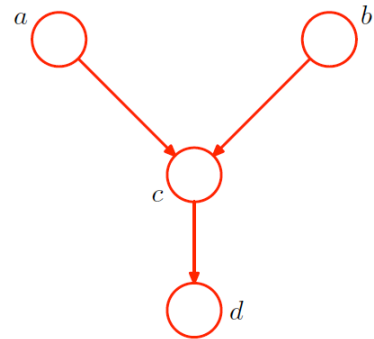
3. (*) Show that the solution to (2) is $\mathbf{w} = \mathbf{A}^\dagger \mathbf{y}$ where $\mathbf{A}^\dagger = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T$ (pseudoinverse).

4. Consider the function $f(x) = \cos(x)$ from 0 to 2π . In Matlab, generate three sets of 10 random data points from this known model by generating 10 values of x (uniformly spread is OK), computing the value of $y = f(x)$, and adding Gaussian noise with three different variances: 0.001, 0.01, and 0.1. Next your program should find \mathbf{w} for lengths 1 to 10 using the expression in 3 (even if you did not derive it). Plot the RMS error (the square root of the average of the squared error) as a function of the length of \mathbf{w} for the three variances. Based on lowest error, what is the best value for the length of \mathbf{w} ?

5. Repeat (4) but instead of the error, plot (a) the log of the likelihood, (b) the AIC value, and (c) the BIC value. Notice that to do this, you will need an estimate for the variance which you can compute from the data and the model fit. (Begin by computing the deviations of the data from the estimates). Which value of \mathbf{w} is suggested in each case?
6. Repeat (4) for lengths 1 through 9, holding out each point in turn. Plot the RMS error as before, but now plot averages of the RMS error over the 9 runs for the training data (the data used to fit the curve) and the held out data. (Note that since your held out data sets have only one point, the RMS is the absolute value). What is a good value for the length of \mathbf{w} for each variance based on the average of the RMS errors?
7. (*) Now let's put a prior on the coefficients of \mathbf{w} . Let's use a simple multivariate normal distribution with $\mathbf{0}$ mean and diagonal covariance (equal for each coefficient). So the precision matrix is simply a parameter, alpha, times the identity matrix. Derive an expression for the posterior distribution. It should depend on both the original precision (or variance) and the precision (or variance) for the prior.
8. (**) Minimize the expression developed in the previous problem to derive an expression for the MAP estimate.
9. (*) Derive the decision boundary between two classes with univariate Gaussian posteriors for given means and variances. In other words, you are laying out the algorithm for deciding between two Gaussians each with their own mean and variance. These four numbers would be the input to your program (if you were to write it).
10. For each of the factorizations below, **first** reorder the factors so that all variables first appear as not being conditioned on. In other words, they appear to the left of the "|" first. This is one canonical way to write the chain rule, i.e., $P(A,B,C)=P(A)P(B|A)P(C|A,B)$. (The answer is not necessarily unique). **Second**, identify any differences between this way of writing the factorization and the canonical form of the chain rule. This identifies where the factorization restricts the distribution from the general case. **Third**, draw a graph representing the factorization.
 - (a) $p(A,B,C,D,E)=p(A|D,E)p(B|D)p(C)p(D|C)p(E)$
 - (b) $p(A,B,C,D,E)=p(A|C)p(E|B)p(C)p(B)p(D|A,E)$
 - (c) $p(A,B,C,D,E)=p(C|A,B,D,E)p(A|B,D)p(D|B)p(B)p(E|A,B,D)$
11. For **each** of your graphs in the previous question, use d-separation to argue whether or not i) C and B are conditionally independent given D, and ii) E and C are conditionally independent given B.

12. (*, From Bishop)

Figure 8.54 Example of a graphical model used to explore the conditional independence properties of the head-to-head path $a-c-b$ when a descendant of c , namely the node d , is observed.



8.10 (*) Consider the directed graph shown in Figure 8.54 in which none of the variables is observed. Show that $a \perp\!\!\!\perp b \mid \emptyset$. Suppose we now observe the variable d . Show that in general $a \not\perp\!\!\!\perp b \mid d$.

13. (*) Draw the graphical model for a research problem that you are working on or thinking about, or would like to work on or think about. You will need to explain what the variables are, and what is generally going on as well.

14. (**) Any exercise in Kollar and Friedman chapter 3 that you find interesting.

15. (**) Same as (13) excluding the one that you have already done.

16. (**) Same as (14) excluding ones that you have already done.

17. (**)

8.11 (**) Consider the example of the car fuel system shown in Figure 8.21, and suppose that instead of observing the state of the fuel gauge G directly, the gauge is seen by the driver D who reports to us the reading on the gauge. This report is either that the gauge shows full $D = 1$ or that it shows empty $D = 0$. Our driver is a bit unreliable, as expressed through the following probabilities

$$p(D = 1|G = 1) = 0.9 \quad (8.105)$$

$$p(D = 0|G = 0) = 0.9. \quad (8.106)$$

Suppose that the driver tells us that the fuel gauge shows empty, in other words that we observe $D = 0$. Evaluate the probability that the tank is empty given only this observation. Similarly, evaluate the corresponding probability given also the observation that the battery is flat, and note that this second probability is lower. Discuss the intuition behind this result, and relate the result to Figure 8.54.