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.

Please create a PDF document with your answers and/or the results of any programs that you write. You should also hand in your programs.

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 a challenge problem; 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.

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.

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 (or two no "+" problems). Additional "+"s scales linearly.

Note: 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.

Undergraduates need to do 6 problems reasonably correctly for full credit, graduate students need to do 10 problems reasonable correctly for full credit.

1. (Regression explained using three problems).

(a)

Let $y(x) = 1 + 2x + 3x^2$

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

(b)

Let $y(x) = 4 + 3x + 2x^2 + 1x^4$

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

as a 5x4 matrix times a vectors of length 4.

(c)

Now suppose you have observed values that might come the model in (b) at the same x values, specifically $(0, \frac{1}{2}, 1, 2, 3)$, in a vector **y**.

Let the 5x4 matrix be **A**, and the vector of length 4 be **w**.

Express the sum of the squared error between the estimate and the data using these matrices and vectors.

(To connect this to the next problem, notice that your answer does not depend on the particular polynomial model or data).

2. Now consider the general case of a polynomial with coefficient vector, **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 know variance. Assume that the length of **w** is given, but that its values are not known. Show that the MLE for **w** is the minimum of the expression from the above problem.

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

4. Consider the function f(x)=cos(x) from 0 to 2 pi. 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 1) Gaussian noise with three different variances: 0.001, 0.01, and 0.1. Next your program should find **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 **w** for the three variances. Based on lowest error, what is the best value for the length of **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 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 **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 **w** for each variance based on the average of the RMS errors?

7.(*) Now let's put a prior on the coefficients of w. Let's use a simple multivariate normal distribution with **0** mean and diagonal covariance (equal for each coefficient). So the precision matrix is simply a parameter, alpha, times the identify 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 in (7) 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. (From Bishop)

Table 8.2 The joint distribution over three binary variables.

a	b	c	$p(a, b, c) \times 1000$
0	0	0	192
0	0	1	144
0	1	0	48
0	1	1	216
1	0	0	192
1	0	1	64
1	1	0	48
1	1	1	96

8.3 ($\star\star$) Consider three binary variables $a, b, c \in \{0, 1\}$ having the joint distribution given in Table 8.2. Show by direct evaluation that this distribution has the property that a and b are marginally dependent, so that $p(a, b) \neq p(a)p(b)$, but that they become independent when conditioned on c, so that $p(a, b|c) = p(a|c)p(b|c)$ for both $c = 0$ and $c = 1$.

11. (*, From Bishop)

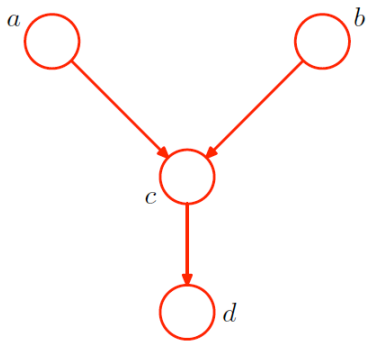

Figure 8.54 Example of a graphical model used to explore the conditional independence properties of the head-to-head path a–c–b when a descendant of c, namely the node d, is observed.

8.10 ($\star$) 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 $a \not\perp\!\!\!\perp b \mid d$.

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

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

14. (**) Same as (13) excluding the same one.

15. (**) Same as (14) excluding one that you have already done.

16. (**)

8.11 ($\star\star$) 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

$$p(D = 1|G = 1) = 0.9 \tag{8.105}$$
$$p(D = 0|G = 0) = 0.9. \tag{8.106}$$

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.