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 6. Grad students need 4 more for a total of 10.

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.

This assignment has 6 questions with no stars, 4 with a single star, and 2 with two stars. There are no extra length problems.

1. Recall that, for a particular form of the likelihood, a conjugate prior is one where the posterior distribution is the same kind of function, just with different parameters. Fill in a few steps to show that the Beta distribution is a conjugate prior of the Binomial distribution ($). Make sure you specify the new parameters of the distribution for the posterior ($).

2. (*) Remind yourself about the Poisson distribution. Show that a conjugate prior for it is the Gamma distribution ($).

3. (**) Most familiar distributions are in the exponential family, which means that they have the form:

$$p(\mathbf{x} \mid \boldsymbol{\eta}) = h(\mathbf{x}) g(\boldsymbol{\eta}) \exp\{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})\}$$

where x may be scalar or vector, discrete or continuous. The parameters η are called the natural parameters of the distribution. $g(\eta)$ is the the normalizing constant.

   (a) Show that the multinomial, Gaussian, gamma, and Poisson distributions are part of this family ($).

   (b) Suggest a generalized conjugate prior for this family, and show that it has the conjugacy property ($).

4. (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 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 **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 for a generic **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 **A** from (b) and observed **y**=$(3,6,12,32,120)^T$ ($).

5. 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 known 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 sum of squared error (i.e., exactly what you expresses in the previous problem) ($).

6. (*) 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). ($)

7. Consider the function f(x)=cos(2x) from 0 to 2 pi which we will use as a simple way to generate some points (there is no real significance to the choice of cos(2x), and if initial reports suggest that a different f(x) would be more interesting, we might need to switch it). In Matlab, generate three sets of 12 data points from this model by generating 12 values of x (uniformly spread is OK), computing the value of y=f(x), and adding Gaussian noise with three different variances: 0.002, 0.02, and 0.2. Call these synthetically generated data sets the "observed data". Next your program should find **w** for lengths 1 to 12 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 using the corresponding observed data ($). Based on lowest error, what is the best value for the length of **w?** ($) In addition, compute the RMS using the "truth" in f(x) ($). What is the best value for the length of **w** based on this test? ($)

8. Repeat the previous question (7), but instead of the error, plot (a) the log of the likelihood, (b) the AIC value, and (c) the BIC value for the observed data ($). Notice that to do this, you will need an estimate for the variance which you can compute from the data and the model fit from the deviations of the data from the estimates. Provide this estimate in your writeup ($). Which value for the length of **w** is suggested in each case? ($)

9. Repeat (7) for lengths 1 through 11, holding out each point in turn. Plot the RMS error as before, but now plot averages of the RMS error over the 11 runs for the training data (the data used to fit the curve) and the held out data ($). Note that since your held out data sets have only one point, the RMS is the absolute value. What is a good value for the length of **w** for each variance based on the average of the RMS errors? ($)

10. (*) 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 ($).

11. (**) Minimize the expression developed in the previous problem to derive an expression for the MAP estimate ($).

12. (*) Derive the decision boundary between two classes with univariate Gaussian posteriors for given means and variances ($). In other words, you are laying out the algorithm for deciding between two models consisting of Gaussians each with their own mean and variance. These means and variances (four numbers) would be initial input to a program that decided which class subsequent values belonged to (no need to write the program). For concreteness, you can think of distinguishing male/ female based on height. Finally, suppose that mistaking the first class for the second is N times more costly than the reverse. Provide a revised expression for the decision boundary ($).