STA414 / STA2104 Midterm Test

Dept of Statistical Sciences, University of Toronto
13 February 2017

Name:

Student Number:

Section *(circle one)*: L0101 = Mon, L5101 = Tues

1	/ 10
2	/ 15
3	/ 15
4	/ 10
5	/ 13
Total	/ 63

Instructions:

- Time allowed: 90 minutes
- Answer all questions. Page 8 has space for overflow
- Any questions completed in pencil rather than pen may not be eligible to be remarked even if there was a marking error
- Aids allowed: You are allowed to bring in one $8.5'' \times 11''$ sheet with handwriting on one side, and a non-programmable calculator

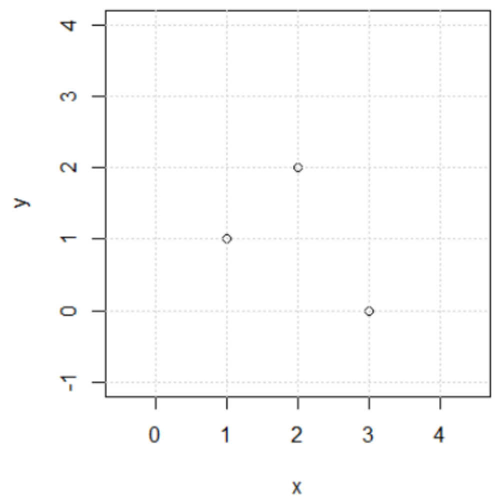
This test should have 8 pages including this page

1. **Cross-validation for Regression [10 Points]**

   A given dataset has three observations of (x_i, y_i) pairs: (1,1), (2,2), (3, 0).

   Suppose you plan to model the data with $y = mx + b$, fitting m and b by ridge regression with a quadratic penalty of the form λm^2. You consider only two values: $\lambda \to 0$ or $\lambda \to \infty$.

   For both possible values of λ, run three-fold cross-validation (i.e., with each validation set having only one case) and report the total squared error for the validation sets. Which choice of λ gives the lower squared error?



2. **Bayesian Inference**. **[15 Points]**

Given a sequence of independent coin flips, each with probability of success μ, the "negative binomial distribution" (not described in lecture) is a distribution over the positive integers that counts the number of successes x before there are r failures. For example, if we set $r = 2$, then draws from the negative binomial distribution might look like:

- $T - H - H - T$ $(x = 2)$
- $H - T - H - H - H - T$ $(x = 4)$
- $H - H - T - T$ $(x = 2)$

Note that the final flip is always tails because we end on the r^{th} failure. The probability mass function (pmf) for the negative binomial distribution is given by

$$p(x) = \binom{x + r - 1}{x} \mu^x (1 - \mu)^r.$$

In this problem, we will let r be a fixed parameter and focus on Bayesian inference on the probability μ.

(a) **[5 Points]** In words, describe how the three factors in the pmf — $\binom{x+r-1}{x}$, μ^x, and $(1 - \mu)^r$ — correspond to the description of the generative process.

(b) **[10 Points]** The conjugate prior for the negative binomial distribution is the β distribution:

$$p(\mu \mid \alpha_0, \beta_0) = \frac{\Gamma(\alpha_0 + \beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} \mu^{\alpha_0 - 1} (1 - \mu)^{\beta_0 - 1}.$$

Given draws $\{x_n\}_{n=1}^N$ where $x_n \in \{1, 2, 3, \cdots\}$, the posterior has the form

$$p(\mu \mid \alpha_N, \beta_N) = \frac{\Gamma(\alpha_N + \beta_N)}{\Gamma(\alpha_N)\Gamma(\beta_N)} \mu^{\alpha_N - 1} (1 - \mu)^{\beta_N - 1}.$$

Write the expressions for the posterior parameters α_N and β_N. Explain the expressions in words.

This page is for your answer to question 2(b).

3. **The Bernoulli distribution [15 Points]**

   Let $p(x_1, x_2)$ be a distribution over two Bernoulli variables $x_1 \in \{0, 1\}$ and $x_2 \in \{0, 1\}$. Suppose you are seeking an approximation to $p(x_1, x_2)$ which we'll call $q(x_1, x_2)$. One way to find $q(x_1, x_2)$ is to minimize the "Kullback-Leibler divergence" (not covered in lecture). This divergence is defined as:

$$\mathrm{KL}(p\,||\,q) = \sum_{x_1, x_2} p(x_1, x_2) \ln \frac{p(x_1, x_2)}{q(x_1, x_2)}.$$

   In particular, suppose you want to approximate $p(x_1, x_2)$ with a factored distribution

$$q(x_1, x_2) = q_1(x_1) q_2(x_2)$$

   where each $q_1(x_1)$ and $q_2(x_2)$ are Bernoulli distributions with means μ_1 and μ_2, respectively. Show that the KL divergence above is minimized by setting these parameters to the expectations $E_p[x_1]$ and $E_p[x_2]$ respectively.

4. **Manipulating Gaussians [10 Points]**

   In this question we have a joint probability distribution, $p(x_0, x_1, x_2)$, in which:

$$
\begin{aligned}
x_0 &\sim \mathcal{N}(0, \sigma^2) \\
x_1 &\sim \mathcal{N}(ax_0, \sigma^2) \\
x_2 &\sim \mathcal{N}(bx_0, \sigma^2)
\end{aligned}
$$

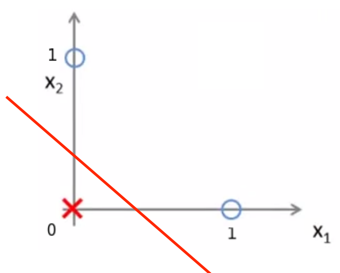
   Compute the marginal distribution $p(x_1, x_2)$. You may use properties of means, variances, expectations, and Gaussian distributions.

5. **Linear Binary Classification Models [13 Points]**

Consider the problem of building a binary classification model $p(c|\boldsymbol{x})$, given input-class pairs $(\boldsymbol{x}_1, t_1), (\boldsymbol{x}_2, t_2), \ldots (\boldsymbol{x}_3, t_3)$, where $t \in \{0, 1\}$, and $\boldsymbol{x} \in \mathbb{R}^2$.

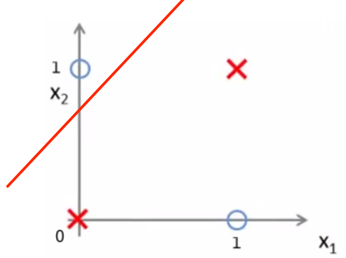
(a) **[3 Points]** Write down a parametric classification model $p(c|\boldsymbol{x}, \mathbf{w})$ with parameters $\mathbf{w}$, whose decision boundaries (lines along which $p(c|\boldsymbol{x})$ is constant) are linear in $\boldsymbol{x}$.

(b) **[3 Points]** Given the following three datapoints, what is the maximum likelihood that can be assigned to this dataset using a non-featurized logistic regression model, maximizing over $\mathbf{w}$? i.e. what is: $\max_{\mathbf{w}} \prod_{i=1}^{3} p(t_i|\boldsymbol{x}_i, \mathbf{w})$? (You don't need to state $\mathbf{w}$.)



likelihood = 1*1*1 = 1

(c) **[3 Points]** Given the following four datapoints, what is the maximum likelihood that can be assigned to this dataset using a non-featurized logistic regression model by maximizing over $\mathbf{w}$? i.e. what is: $\max_{\mathbf{w}} \prod_{i=1}^{4} p(t_i|\boldsymbol{x}_i, \mathbf{w})$? (You don't need to state $\mathbf{w}$.)



(d) **[4 Points]** Write down a set of features $\{\boldsymbol{\phi}(\boldsymbol{x})\}$ that would allow a linear model to correctly classify all the datapoints in part (c).

This page is for rough work.
If you include a solution here, you must indicate so near the question itself.