STA2104 L5101: Midterm

GRADUATE STUDENTS

February 13, 2018

First Name:____

Last Name:____

Student Number:____

- This midterm consists of 15 multiple choice questions and 3 written answer questions, for a total of 70 marks.
- This midterm contains 14 pages.
- Your TCard must be displayed on your desk at all times.
- **Circle your final answer to each problem. Questions that do not have a circled answer will receive zero marks**.
- All final answers must be written on the **front** of the page. Nothing written on the backs of pages will be marked. You should do your rough work on the back of the page first, then write and circle your final answer on the front of the page.
- Your answer for a question must appear on the same page as the question.
- Because of this requirement, note that the amount of space given to answer a question might be *much* more than is required.
- Do not write at the top of the page, above the question number. This will mess with the scanning of the QR code.
- Non-programmable calculators may be used. No other aids are permitted. No other papers are allowed on your desk
- If you need extra paper for writing, raise your hand. You must hand in all extra sheets of paper used. Nothing written on extra sheets of paper will be marked.
- Use pen. Questions answered in pencil will not be eligible for remark requests.

Each multiple choice question is worth 2 points, for a total of 30 points. Circle the answer corresponding to the correct statement.

1. Which of the following models is a *linear* model, for feature x and parameters $\mathbf{w}$?
   (a) $y(x, \mathbf{w}) = x^{w_1} + x^{w_2}$
   (b) $y(x, \mathbf{w}) = w_1 x + w_2/x$
   (c) $y(x, \mathbf{w}) = w_1 x + w_2 x^2 + \ldots + w_d x^d$
   (d) a and b
   (e) a and c
   (f) b and c
   (g) All of the above

2. Suppose we have a loss function $L(\mathbf{w}, \mathbf{x})$ that we wish to minimize with respect to $\mathbf{w}$. Suppose that we want to add a regularization penalty to this loss function. Which of the following penalties would be a valid regularization penalty?
   (a) $\frac{\lambda}{2}\mathbf{w}'\mathbf{w}$
   (b) $\frac{\lambda}{2}\mathbf{w}'\mathbf{x}$
   (c) $\frac{\lambda}{2}\mathbf{x}'\mathbf{x}$
   (d) All of the above

3. Suppose we have a dataset with n observations and p features. Which of the following conditions are necessary for a *ridge regression* model to have a unique solution, assuming $\lambda > 0$?
   (a) $n > p$
   (b) $n < p$
   (c) The columns of the design matrix $\mathbf{X}$ are linearly independent
   (d) None of the above

4. The *likelihood* function is
   (a) The joint distribution of the parameters
   (b) The joint distribution of the sample, treated as a function of the parameters
   (c) A function we choose to represent our prior belief about the parameter values
   (d) The function that most likely generated the sample we observed

5. Consider a 2-dimensional Gaussian distribution with mean $\mathbf{0}$ and covariance matrix $\Sigma = \sigma^2\mathbf{I}$, where $\mathbf{I}$ is the identity matrix. The contours of constant density for this distribution form which geometric shape?
   (a) Axis-aligned ellipse centred at $(0, 0)$ with unequal length axes
   (b) Circle centred at $(0, 0)$
   (c) Circle, centered at an unknown point
   (d) Impossible to say with the information given
6. Suppose we have a linear discriminant function $y(\mathbf{x}) = \mathbf{w}'\mathbf{x}$, and define a classification rule such that we classify a point $\mathbf{x}$ to be in class A if $y(\mathbf{x}) > 0$ and class B if $y(\mathbf{x}) < 0$. This separates the input space using a decision surface that is a linear function of $\mathbf{x}$. Which of the following *activation functions* also results in a decision surface which is a linear function of $\mathbf{x}$?
   (a) $y(\mathbf{x}) = \frac{\exp(\mathbf{w}'\mathbf{x})}{1+\exp(\mathbf{w}'\mathbf{x})}$
   (b) $y(\mathbf{x}) = \log\left(\mathbf{w}'\mathbf{x}\right)$
   (c) $y(\mathbf{x}) = \tanh\left(\mathbf{w}'\mathbf{x}\right)$
   (d) All of the above
7. We wish to build a model to predict class C_k given input $\mathbf{x}$. We choose to model $P(C_k|\mathbf{x})$ directly. This is referred to as a
   (a) Generative model
   (b) Discriminative model
   (c) Probabilstic model
   (d) Linear least squares
8. We wish to build a model to predict class C_k given input $\mathbf{x}$. We choose to model $P(\mathbf{x}|C_k)$ and $P(C_k)$, and use these to calculate $P(C_k|\mathbf{x})$. This is referred to as a
   (a) Generative model
   (b) Discriminative model
   (c) Probabilstic model
   (d) Linear least squares

9. Fisher's Linear Discriminant Analysis finds the projection $y = \mathbf{w}'\mathbf{x}$ that
   (a) Maximizes the separation of the classes
   (b) Minimizes the separation of the classes
   (c) Maximizes the within-class variance of the classes
   (d) Minimizes the within-class variance of the classes
10. In its unmodified form, Linear Discriminant Analysis performs best when
    (a) Class covariance matrices are all differently shaped
    (b) Class covariance matrices are all differently shaped, but axis-aligned
    (c) Class covariance matrices are all spherical
    (d) Class covariance matrices all have determinant equal to 1
11. The following 5 questions consider the following situation. Suppose we have a model with one *hyperparameter* λ that we would like to choose via S-fold cross-validation. We have N training cases available, and our metric is squared error. As a first step, we
    (a) Split the training set into S folds, each with N/S observations, sampling without replacement
    (b) Split the training set into S folds, each with N/S observations, sampling with replacement
    (c) Split the training set into N folds, each with S observations, sampling with replacement
    (d) Split the training set into S folds, each with N observations, sampling with replacement
12. Next, for some single value of λ we
    (a) Fit our model on all of the folds and report the squared error on each
    (b) Fit our model on all of the folds and report the average squared error on those folds
    (c) Fit our model on all but one of the folds and report the average squared error on those folds
    (d) Fit our model on all but one of the folds and report the squared error on the remaining fold

13. We do this
    (a) For a pre-specified grid of values of λ that we choose arbitrarily
    (b) For a pre-specified grid of values of λ that are most likely according to a prior distribution on λ
    (c) For values of λ that we choose via cross-validation
    (d) For the value of λ that maximizes the log-likelihood
14. We choose the value for λ that
    (a) Maximizes the in-sample squared error
    (b) Maximizes the out-of-sample squared error
    (c) Minimizes the in-sample squared error
    (d) Minimizes the out-of-sample squared error
15. Suppose we do this for $\lambda = 0.1, 0.01, 0.001$, and we get a cross-validated squared error of $0.4, 0.1, 0.1$ respectively. What would you do?
    (a) Pick $\lambda = 0.01$ because you should always choose the highest value of λ in the event that multiple values give the same cross-validated squared error
    (b) Try more λ values
    (c) Conclude that the model doesn't fit the data well
    (d) Try a more complex model with more features

1. (14 marks) Suppose a single datapoint x follows an *Exponential* distribution, with density

$$f(x|\beta) = \beta e^{-x\beta}$$

We wish to estimate β in the Bayesian framework, so we put a $Gamma(a, b)$ prior distribution on β with density

$$p(\beta|a, b) = \frac{b^a}{\Gamma(a)}\beta^{a-1}e^{-\beta b}$$

The (a, b) are hyperparameters.

(a) (2 marks) Write down the formula for the *posterior* distribution $p(\beta|x, a, b)$, in terms of $f(x|\beta)$ and $p(\beta|a, b)$. Don't plug anything in- your answer contains the terms $f(x|\beta)$ and $p(\beta|a, b)$.

(b) (4 marks) The Gamma distribution is the *conjugate prior* for the Exponential distribution. Use this fact to evaluate the *normalization constant*, that is, the denominator of your answer to part a).

(c) (4 marks) Show that the posterior distribution is a $Gamma(a_1, b_1)$ distribution, and give explicit expressions for the parameters a_1, b_1. These expressions depend on a, b, and x.

(d) (4 marks) Evaluate the *predictive distribution* of a new datapoint x^* up to a constant. This means your answer does not need to be a properly normalized distribution.

2. (16 marks) The *Poisson* distribution is sometimes used as a model for count data. It has probability mass function

$$P(Y = y|\lambda) = \frac{\lambda^y e^{-\lambda}}{y!}$$

(a) (2 marks) Write the *likelihood function* for an independent, identically distributed random sample of size N, $y_1 \ldots y_N$, from this distribution.

(b) (4 marks) For natural parameter $\eta = \log \lambda$ and sufficient statistic $u(y) = y$, write the likelihood function in exponential family form,

$$L(\lambda) = \left(\prod_{n=1}^{N} h(y_n) \right) g(\eta)^N \exp \left(\eta \times \sum_{n=1}^{N} u(y_n) \right)$$

Identify the functions $h(y)$ and $g(\eta)$.

(c) (4 marks) Find the *maximum likelihood estimator* for λ.

(d) (6 marks) Now suppose for each y_n we observe a p-dimensional vector of features, $\mathbf{x}_n$, and we wish to use it to build a model for y_n. Define a p-dimensional parameter vector $\mathbf{w}$, and write

$$\eta_n = \mathbf{w}'\mathbf{x}_n, n = 1 \ldots N$$

(i) (3 marks) Write down the likelihood function for $\mathbf{w}$.

(ii) (3 marks) State the equation that the *maximum likelihood estimator* for $\mathbf{w}$ must satisfy. You won't be able to find a closed-form expression for $\hat{\mathbf{w}}$; just state the equation that must be solved.

3. (10 marks) Recall the k-Nearest-Neighbours algorithm discussed in lecture. For a training set of features and targets $(\mathbf{x}_n, t_n), n = 1 \dots N$ where t_n is a categorical target, we wish to predict the class of a new point $\mathbf{x}$.

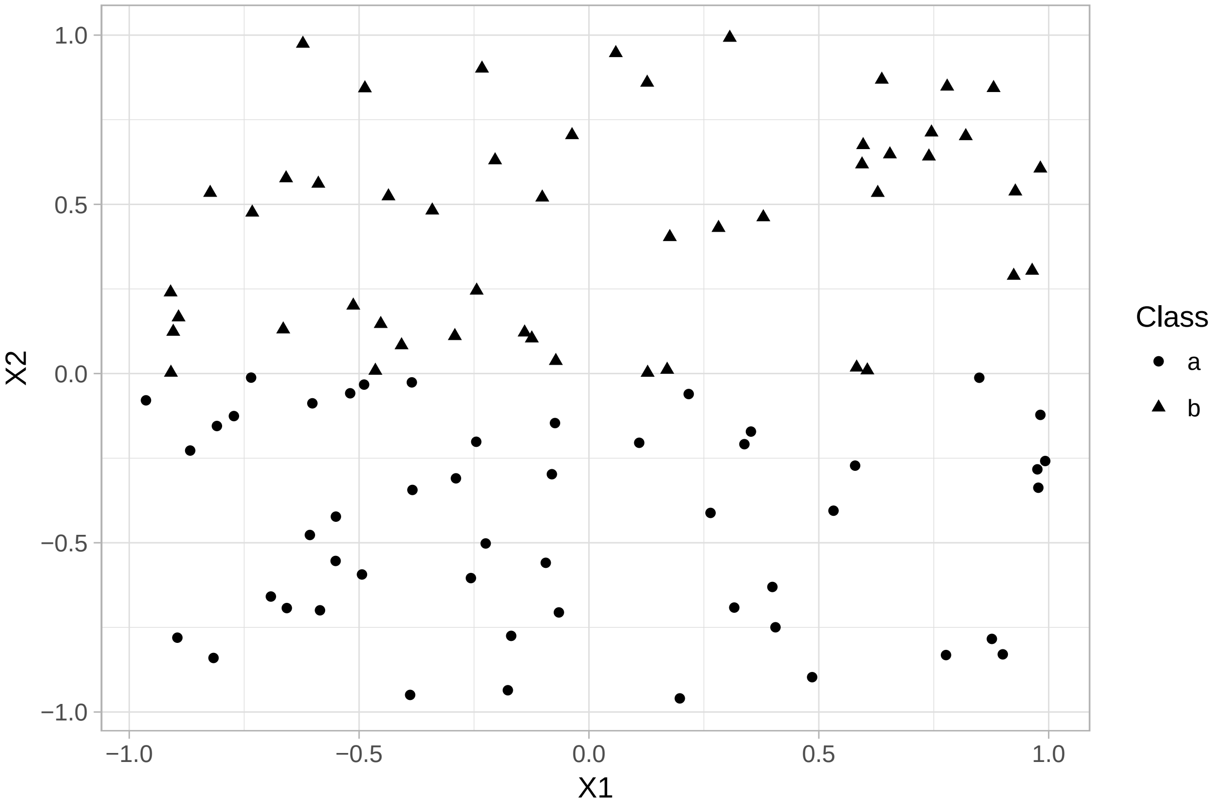
   (a) (4 marks) State the k-Nearest-Neighbours algorithm. Your distance metric should be Euclidean distance.

(b) (4 marks) We discussed the concept of a *linear smoother*, a prediction function $y(\mathbf{x})$ of the form

$$y(\mathbf{x}) = \sum_{n=1}^{N} k(\mathbf{x}, \mathbf{x}_n) t_n$$

a linear combination of the training set targets, where $k(\mathbf{x}, \mathbf{x}_n)$ depends on the distance between $\mathbf{x}$ and $\mathbf{x}_n$. Find the function $k(\mathbf{x}, \mathbf{x}_n)$ that gives a linear smoother that is equivalent to the k-Nearest-Neighbours algorithm you described in part a).

(c) (2 marks) For the dataset of size $N = 100$ shown in the plot, would you prefer a 1-NN or a 100-NN approach?


X2
X1
Class
● a
▲ b