UNIVERSITY OF TORONTO

Faculty of Arts and Science

April 2018 EXAMINATIONS

STA414/2104H1S L5101

Duration - 3 Hours

Aids Allowed: Non-programmable calculator

First Name:_____

Last Name:_____

Student Number:_____

This exam question booklet contains 14 pages. Answer all questions in the provided answers booklet (separate from this booklet), or on the provided scantron sheet. Use pen for long answers, and pencil for scantron answers. This yellow booklet must be handed in, and nothing written in it will be marked.

Questions:

Question	Marks Achieved	Total Possible
MC		60
1		34
2		22
3		16
4		18
5		22
Total		172

PLEASE HAND IN. DO NOT WRITE ANY ANSWERS ON THIS YELLOW QUESTION PAPER.

Each of the following multiple choice questions is worth 2 marks, for a total of 60 marks. Fill in the correct answer on the provided scantron sheet. Answers written in the booklet will not be marked.

1. Optimization procedures, assuming they do converge to a minimum of some kind:
   (a) Always find global minima
   (b) Are never guaranteed to find global minima for arbitrary objective functions
   (c) May be guaranteed to find global minima, depending on the procedure
   (d) Always find global minima if there are no constraints
2. Subject to certain technical constraints, Automatic Differentiation:
   (a) Computes approximate derivatives of arbitrary computer code
   (b) Computes numeric derivatives of arbitrary computer code
   (c) Computes exact derivatives of arbitrary computer code
3. Which of the following modifications to gradient descent produces stochastic gradient descent?
   (a) At each iteration, randomly subsample the training set and compute the gradient only for this subsample
   (b) Before the procedure starts, randomly subsample the training set and compute the gradient only for this subsample at each iteration
   (c) At each iteration, randomly perturb the direction of the gradient
   (d) At each iteration, randomly select a learning rate from a $Unif(0, 1)$ distribution
4. Linear Basis Function Models produce predictions that...
   (a) Are always linear in the parameters
   (b) Depend only on the first-order effects of the features, i.e. features X_i and X_j will always affect the model independently
   (c) Are the same as linear regression if the data is normally distributed
   (d) Are the same as linear regression if the data follows an exponential family distribution
5. Loss can be decomposed into which of the following?
   (a) Bias and Variance
   (b) Variance and Noise
   (c) Bias and Noise
   (d) Bias, Variance and Noise
6. Which of the following statements is true for Bayesian inference?
   (a) The posterior has a closed form answer only if we use conjugate priors
   (b) The posterior always has a closed-form answer if the data are IID
   (c) We can choose the prior to be of different forms, leading to different posteriors
   (d) The normalizing constant will be the same for any models built using a given dataset, and can be ignored

7. Linear smoothers of the form $\hat{y} = f(\mathbf{x}) = \sum_{i=1}^{n} k(\mathbf{x}, \mathbf{x}_n) y_n$:
   (a) Always produce models that are locally sensitive to changes in $\mathbf{x}$
   (b) Always produce models that are globally sensitive to changes in $\mathbf{x}$
   (c) Can produce models that are either globally or locally sensitive, depending on the choice of $k(\cdot, \cdot)$
8. Which is not a problem when using linear least squares for classification as opposed to logistic regression?
   (a) Sensitivity to outliers
   (b) Violation of independence assumptions
   (c) Model predictions not bounded between $(0, 1)$
   (d) When there are ≥ 3 classes, Linear Least Squares can cause some classes to be masked
9. Logistic regression:
   (a) Has a unique closed-form solution for any dataset
   (b) Is based on an underlying distribution for the targets that is a member of the exponential family
   (c) Is more sensitive to outliers than linear least squares
   (d) Is equivalent to a single-layer fully connected neural network using a linear activation function
10. The K-Means procedure
    (a) Is equivalent to K-Nearest Neighbours if the data is normally distributed
    (b) Is the Bayesian version of K-Nearest Neighbours
    (c) Is the generalization of a gaussian mixture model which gives soft classifications
    (d) Is a special case of a gaussian mixture model which gives hard classifications
11. In Gaussian mixture models:
    (a) Maximum likelihood is not possible because the likelihood is too complicated
    (b) There is no closed-form solution for the maximum likelihood estimators
    (c) A closed form solution for the maximum likelihood estimators only exists when the component covariance matrices are all constrained to be equal
    (d) Gradient descent can't be used to estimate the parameters because no suitable loss function exists
12. An EM Algorithm is
    (a) A procedure for finding maximum likelihood estimates in the presence of discrete latent variables only
    (b) A procedure for finding maximum likelihood estimates when the data are not IID
    (c) A procedure for finding maximum likelihood estimates in the presence of missing data
    (d) Equivalent to Newton's method with the Hessian approximated by a rank-1 projection

13. Regarding the tractability of the E and M steps,
    (a) Both the E step and the M step might need to be done numerically
    (b) The E step might need to be done numerically, and the M step is always tractable
    (c) Both steps are always tractable, which is why the EM algorithm is so popular
    (d) The E step is tractable when the marginal (observed) data log-likelihood is a linear function of the missing data
14. Which of the following is not a challenge when extending the linear model to have a general, unstructured covariance matrix for the errors, Σ?
    (a) The data is independent, so trying to estimate covariance parameters will lead to a degenerate solution
    (b) We have n datapoints, but Σ has $n(n+1)/2$ parameters
    (c) Estimating Σ will affect the precision with which we can estimate other parameters in the model
    (d) Σ would need to be constrained to be symmetric and positive definite.
15. How could we introduce dependency into our linear regression model in a tractable and interpretable way?
    (a) Use our knowledge about the form of the dependency in the data to come up with a latent variable formulation that gives us the covariance structure we want
    (b) Put a prior distribution on $\mathbf{w}$, the parameters of the model
    (c) Extend our model to incorporate latent variables in the form of a single layer fully connected neural network
    (d) Perform a principal components analysis on the data matrix first
16. How could we introduce dependency into our linear regression model in a way that is readily extendible to further types and depths of dependency?
    (a) Add a continuous latent variable directly into the linear model
    (b) Add a continuous latent variable into the covariance matrix Σ
    (c) Use a neural network: we can add more layers to capture deeper dependencies between the observations
    (d) Performing a principal components analysis as suggested above also leads to a readily extensible model
17. Principal Components Analysis tries to factor the data matrix $\mathbf{X}$ as $\mathbf{X} = \mathbf{UZ}$ in a way that preserves:
    (a) The most of all types of structure in $\mathbf{X}$
    (b) The maximum variability in $\mathbf{X}$
    (c) The maximum correlation between columns in $\mathbf{X}$
    (d) The maximum correlation between rows in $\mathbf{X}$

18. Neural Networks...
    (a) Capture dependencies between training observations through their complex layer structure
    (b) Capture interactions between features through their complex layer structure
    (c) Model a nested linear relationship between features and targets
    (d) Model a nonlinear relationship between features and targets, but do not capture interactions between features
19. Neural networks are typically trained...
    (a) Using gradient descent, which is called "backpropagation"
    (b) Using backpropagation, which is a computationally efficient implementation of gradient descent for neural networks
    (c) Using backpropagation, which is a computationally efficient implementation of maximum likelihood for neural networks
    (d) By fitting a sequence of logistic regression models to the hidden layers
20. Which of the following is not a practical consideration when training neural networks?
    (a) Overparametrization
    (b) Zero weights
    (c) Choosing an activation function
    (d) Features with zero variance
21. Which of the following is not a form of regularization in neural networks?
    (a) Dropout
    (b) Adding a penalty to the loss function
    (c) Early stopping
    (d) Subsampling the training data
22. The AIC and BIC are typically used to
    (a) Compare the fit of nested models to the training set
    (b) Compare the fit of any models to the training set
    (c) Compare the fit of nested models to the test set
    (d) Compare the fit of any models to the test set
23. The ROC/AUC and KS statistics are measures of
    (a) How well the model fits the data
    (b) How well the model predicts new targets
    (c) The importance of the features in the the model
    (d) The deviation of the targets from a normal distribution

24. More specifically, the Area Under the ROC Curve is best described as a measure of
    (a) The probability that a new point will be correctly classified
    (b) Classification error
    (c) The probability that a training set point is correctly classified
    (d) The probability that a randomly selected pair of positive and negative points will be correctly rank-ordered by the model
25. A saturated model...
    (a) Fits the training data perfectly (zero loss)
    (b) Fits the test data perfectly (zero loss)
    (c) Is nested in every other possible model
    (d) Predicts a single value for all targets
26. A null model...
    (a) Fits the training data perfectly (zero loss)
    (b) Fits the test data perfectly (zero loss)
    (c) Is nested in every other possible model
    (d) Predicts a single value for all targets
27. A correct model refers to
    (a) The model that actually generated the data
    (b) The model that is most closely nested in the model that actually generated the data
    (c) Any model that is nested in the model that actually generated the data
    (d) Any model that is nested in a sataurated model
28. Why do decision trees overfit more than regression, usually?
    (a) Decision trees are locally optimal, where as regression models are globally optimal
    (b) Decision trees use only a subset of the training data
    (c) Decision trees consider only a random subset of the features
    (d) Regression models are optimized specifically to not overfit
29. In a decision tree, what is the meaning of a surrogate variable?
    (a) A variable that splits almost as well as the splitting variable, and so will be used in subsequent splits in the tree
    (b) A variable that splits the data well, conditional on the current splitting variable
    (c) A variable that would split the data the best at a given node, if it were not partially missing
    (d) A variable that splits the data almost as well as the splitting variable, and will be used in cases where the splitting variable is missing
30. Boosting...
    (a) Is a linear basis function model where the basis functions are chosen to minimize the distance between the model and a globally optimal decision tree
    (b) Is a linear basis function model where the basis functions are individual classifiers/regressors fit in a sequence
    (c) Is a procedure for reducing the variance of a decision tree model
    (d) Is a procedure for linearizing a decision tree model

1. (34 marks) Suppose we have two coins, with different probabilities of heads. We are flipping these coins to see whether they come up heads. We do this by first randomly choosing coin 1 or coin 2 with a fixed, unknown probability, and then flipping it and recording whether it was heads. To denote this, we define

$$Z = \begin{cases} 1 \text{ if we flip coin 1} \\ 0 \text{ if we flip coin 2} \end{cases}$$

and

$$X = \begin{cases} 1 \text{ if the coin is heads} \\ 0 \text{ else} \end{cases}$$

We have the following unknown parameters:

$$P(Z = 1) = \alpha \in (0, 1)$$
$$P(X = 1|Z = 1) = \phi_1 \in (0, 1)$$
$$P(X = 1|Z = 0) = \phi_2 \in (0, 1)$$

We flip the coins n times. The data is then

$$\mathbf{x} = (x_1, \ldots, x_n)$$
$$\mathbf{z} = (z_1, \ldots, z_n)$$

where $x_i \in \{0, 1\}$ represents the result of the i^{th} flip (1 = heads), and $z_i \in \{0, 1\}$ is coded as 1 if we flipped coin 1 and 0 if we flipped coin 2 on the i^{th} flip. You may recall that if a random variable takes on values $\{0, 1\}$ with fixed probabilities α and $1 - \alpha$, then it follows a Bernoulli distribution with

$$P(X = x) = \alpha^x (1 - \alpha)^{1-x}$$

Using this, we can write the conditional distribution of $x_i|z_i$ as

$$P(X_i = x_i|z_i) = \left(\phi_1^{x_i}(1 - \phi_1)^{1-x_i}\right)^{z_i} \times \left(\phi_2^{x_i}(1 - \phi_2)^{1-x_i}\right)^{1-z_i}$$

The problem is: we lost the information on which coin was flipped. Can we still estimate α, θ_1, and θ_2?

(a) (3 marks) Briefly explain what a latent variable is, and state which quantity above is a latent variable in this problem.

(b) (2 marks) Write down $p(z_i) = P(Z_i = z_i)$, the marginal distribution of each z_i, using the information given in the question.

(c) (4 marks) Find the joint distribution $p(x_i, z_i) = P(X_i = x_i, Z_i = z_i)$.

(d) (4 marks) Write down the complete-data log-likelihood $\ell(\theta|\mathbf{x}, \mathbf{z})$, where the parameter vector $\theta = (\alpha, \phi_1, \phi_2)$. You may assume each (x_i, z_i) are independent of (x_j, z_j), $j \neq i = 1 \dots n$.

(e) (4 marks) Find the marginal distribution $p(x_i)$. Briefly describe what this quantity represents.

(f) (2 marks) Write down the marginal log-likelihood $\ell(\theta|\mathbf{x})$. Briefly explain why performing maximum likelihood on this function directly is difficult.

(g) (3 marks) If we wanted to perform maximum likelihood via an EM algorithm, state what would be the complete data, observed data, and missing data.

(h) (4 marks) Evaluate $P(Z_i = 1|x_i)$ and $E_{\mathbf{z}|\mathbf{x},\theta^t}(Z_i)$, where θ^t is some fixed value of θ.

(i) (8 marks) State an EM algorithm for finding maximum likelihood estimates of θ, as follows:

- State the E-step in complete detail
- Describe the M-Step, but do not perform the maximization. State which quantity is to be maximized, and how the result is used to update the parameter vector.
- An answer worth full marks will be very explicit as to which θ's are free parameters, and which are fixed at their current iteration estimates, for both the E- and M-steps.

2. (22 marks) Consider the plot of a sample from a bivariate probability distribution $F(X_1, X_2)$, entitled "Question 2, Plot 1".

   (a) (4 marks) Draw arrows pointing in the direction of the principal components that would be obtained by doing a Principal Components Analysis on these data. Just draw a rough sketch of the axes in your exam booklet, and draw arrows pointing in the appropriate direction. Clearly label which one is PC1 and which is PC2.

   (b) (4 marks) Which of the following three eigendecompositions most reasonably could have been obtained from the covariance matrix of the above data? Write (i), (ii) or (iii) in your exam booklet.

   (i)
```
## eigen() decomposition
## $values
## [1] 4.6477295 0.1733598
##
## $vectors
##              [,1]         [,2]
## [1,] -0.999997429 -0.002267785
## [2,]  0.002267785 -0.999997429
```

   (ii)
```
## eigen() decomposition
## $values
## [1] 6.169855 0.011874
##
## $vectors
##            [,1]       [,2]
## [1,] -0.9138437  0.4060661
## [2,] -0.4060661 -0.9138437
```

   (iii)
```
## eigen() decomposition
## $values
## [1] 1.0873477 0.9298689
##
## $vectors
##            [,1]       [,2]
## [1,] -0.7302111 -0.6832216
## [2,]  0.6832216 -0.7302111
```

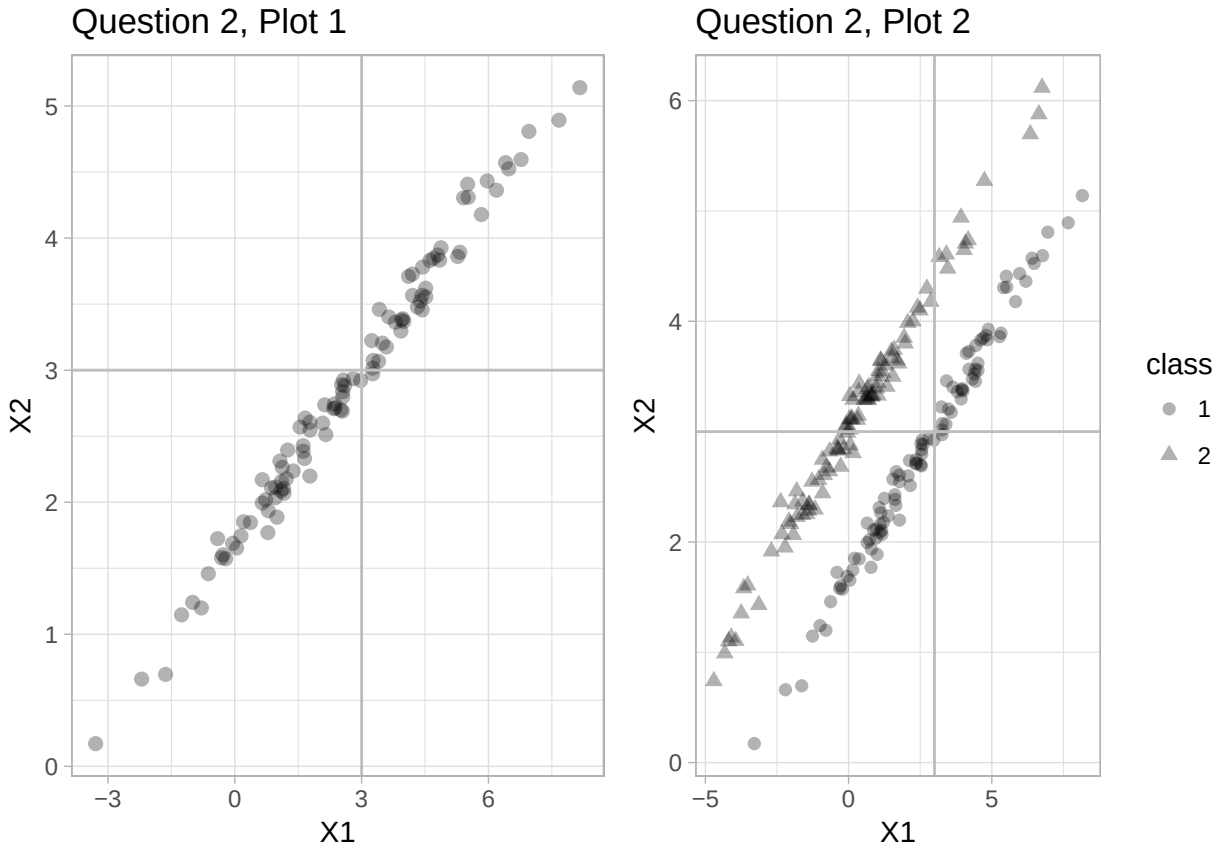
   (c) (4 marks) Based on your answer, what is the total variance present in the data, and the proportion of variance explained by the first PC?

   (d) (2 marks) How many PCs would you keep, if your sole goal were dimension reduction? Give an

informal argument for your answer, do not bother with any arbitrary rules that may or may not have been discussed in lecture.

(e) (4 marks) Now consider the plot entitled "Question 2, Plot 2". In your exam booklet, draw another rough plot axis, and on it, draw an arrow pointing in the direction of the first PC. Again, you don't have to be too exact here.

(f) (4 marks) If your goal were to build a classification model for the data shown in Question 2 Plot 2, would you keep 1 PC or 2? Explain your answer.



3. (16 marks) Suppose we have a dataset of features $\mathbf{x}_i \in \mathbb{R}^p$ and real-valued targets, $y_i, i = 1 \dots n$. We wish to build a decision tree for these data.

   (a) (8 marks) State the decision tree algorithm, as an algorithm (list of steps). Use the loss function $L(\mathbf{y}, \hat{\mathbf{y}}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$, where $\hat{y}_i$ represents the tree's prediction for y_i.

   (b) (6 marks) Denoting the regions found by the above procedure as R_m, write a decision tree as a linear basis function model in the $\mathbf{x}_i$,

$$\hat{y_i} = \sum_{m=1}^{M} \alpha_m h_m(\mathbf{x}_i)$$

   Explicitly define the functions $h(\mathbf{x}_i)$, and the constants α_m.

   (c) (2 marks) Recall the K-Nearest Neighbours (KNN) prediction function for a new point $\mathbf{x}^*$: $\hat{y}$ gets assigned to the average of the K nearest points in the training set to $\mathbf{x}^*$. What decision tree nodes give a decision tree that is equivalent to a KNN regressor of this form?

4. (18 marks) Suppose we have a dataset of features $\mathbf{x}_i \in \mathbb{R}^p$ and real-valued targets, $y_i, i = 1 \dots n$. We build a model that provides predictions $\hat{y} = f(\mathbf{x})$.

   (a) (8 marks) State the algorithm for bagging this model. Denote the bagged prediction function as $f_{bag}(\mathbf{x})$

   (b) (2 marks) It is known that bagging does not provide a reduction in variance for linear prediction functions, of the form $f(\mathbf{x}) = \beta'\mathbf{x}$ for fixed $\beta \in \mathbb{R}^p$. Briefly explain why this means that, even though in general $f_{bag}(\mathbf{x}) \neq f(\mathbf{x})$, bagging is useless in this case.

   (c) (2 marks) In the previous question we said that a decision tree is a linear basis function model. Give a brief explanation as to why (b) does not contradict what was discussed in lecture: that bagging provides variance reduction for trees.

   (d) (2 marks) Why can't we reduce the variance of our prediction function as close to zero by increasing the number of bootstrapped resamples to an arbitrarily high number? You can give either an intuitive or a mathematical argument here.

   (e) (4 marks) State the modification to the bagging procedure that produces Random Forests. Why does this help mitigate the effect of the above problem?

5. (22 marks) The following fictional dataset represents credit card customers at a bank. We wish to build a model to classify whether these customers will default on their credit card debt in some specified time period. Denote $y = 0$ as a non-default and $y = 1$ as a default. y appears in the dataset as the variable **target**. Here is a basic view of the data:

```
## # A tibble: 10,000 x 4
##    limit   age num_products target
##    <dbl> <dbl>        <dbl>  <int>
##  1 10500    32            0      0
##  2 12900    19            3      0
##  3 37800    18            7      0
##  4 45900    18            3      0
##  5 21700    53            4      0
##  6 43300    41            7      1
##  7 10000    18            3      0
##  8 27000    31            0      0
##  9 26400    20            4      0
## 10 11700    59            4      0
## # ... with 9,990 more rows
```

```
## Table of target:
```

```
##
##    0    1
## 9009  991
```

(a) (4 marks) What would be the classification accuracy on a model that predicted all customers as being non-defaulters ($y = 0$)? Briefly explain why classification accuracy is not the only metric we should consider when building a model for these data.

(b) (2 marks) A logistic regression model was built, giving a prediction function:

$$\hat{y}(\mathbf{x}) = \frac{1}{1 + \exp(\mathbf{x}'\hat{\mathbf{w}})}$$

where the features are (limit,age,num_products). However, I am only allowed to pick 2 features. Argue why it is acceptable (not optimal, just acceptable) to use the AIC or BIC to choose features in this situation.

(c) (4 marks) State the algorithm for performing forwards stepwise selection using either AIC or BIC.

(d) (4 marks) Give one argument in favour of using each of the AIC and BIC to choose features.

(e) (6 marks) Consider the ROC curve shown in the plot "Question 5, ROC Curve". Give a detailed explanation of how this curve was generated from the model described in (b), which produces soft classifications.

(f) (2 marks) If we pull a randomly selected pair of points from the training set and observe $y_1 = 1$ and $y_2 = 0$, then using the prediction function from the fitted model, evaluate $P(\hat{y}_1 > \hat{y}_2)$.