UNIVERSITY OF TORONTO
Faculty of Arts and Science

APRIL 2018 EXAMINATIONS

DURATION: 3 hours

CSC 411 H1S — Machine Learning and Data Mining

No Aids Allowed
Reference/draft sheets are distributed with the exam
Examiner(s): M. Guerzhoy, L. Zhang

Student Number: |_|_|_|_|_|_|_|_|_|_|

Family Name(s): _____

Given Name(s): _____

---

*Do **not** turn this page until you have received the signal to start.*
In the meantime, please read the instructions below *carefully.*

This final examination paper consists of 8 questions on 32 pages (including this one), printed on both sides of the paper. *When you receive the signal to start, please make sure that your copy is complete, fill in the identification section above, and write your student number where indicated at the bottom of every odd-numbered page (except page 1).*

Answer each question directly on this paper, in the space provided, and use the reverse side of the previous page for rough work. If you need more space for one of your solutions, use the reverse side of a page or the pages at the end of the exam and *indicate clearly the part of your work that should be marked.*

Write up your solutions carefully! In particular, use notation and terminology correctly and explain what you are trying to do—part marks *may* be given for showing that you know some aspects of the answer, even if your solution is incomplete.

A mark of at least **40%** (after adjustment, if there is an adjustment) on this exam is required to obtain a passing grade in the course.

MARKING GUIDE

# 1:	_____/ 20
# 2:	_____/ 10
# 3:	_____/ 15
# 4:	_____/ 15
# 5:	_____/ 10
# 6:	_____/ 10
# 7:	_____/ 10
# 8:	_____/ 10
TOTAL:	_____/100

Good Luck!

OVER...

Use this page for rough work—clearly indicate any section(s) to be marked.

CONT'D...

Question 1. [20 MARKS]

Write down whether each of the statement is **True** or **False**. Each correct answer is worth 1 mark.

1. _____ The MAP and MLE estimates can only be equal when the number of training examples is very large.
2. _____ The k-Means algorithm is sometimes equivalent to k-Nearest Neighbours for small k.
3. _____ We can help avoid overfitting by obtaining more training data.
4. _____ Convolutional Neural Networks are only used for images.
5. _____ Dropout is equivalent to L1 regularization since they both encourage sparsity.
6. _____ An ensemble of models always has greater capacity than a single model.
7. _____ When training decision trees, we choose splits that minimize the error rate.
8. _____ Naive Bayes is less computationally expensive to train than Decision Trees.
9. _____ We can train the weights of a CNN using guided backpropagation to interpret the network.
10. _____ Non-parametric models do not have parameters.
11. _____ In linear SVMs with a soft margin, a large ϵ means the learning rate is small.
12. _____ For the k-Nearest Neighbours classifier, for larger values of k, the capacity of the classifier will tend to be smaller.
13. _____ To solve an unregularized linear regression problem, we solve for $\text{argmin}_\theta \sum_i (\theta^\top x^{(i)} - y^{(i)})^2$.
14. _____ A linear SVM always produces the same decision boundary as Logistic Regression.
15. _____ The Maximization stage of EM is analogous to the computation of cluster centers in k-Means
16. _____ The decision boundary of a neural network with a single hidden layer is linear.
17. _____ Bagging can be used to decrease the variance of a classifier
18. _____ Leave-one-out is always the same as N-fold cross validation where N is the size of the training set.
19. _____ Autoencoders are trained using an application of the Spectral Theorem.
20. _____ It is obvious for humans when an input image contains an adversarial attack.

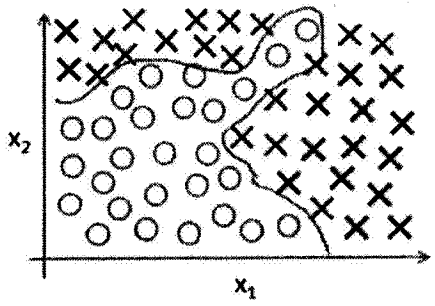
Use this page for rough work—clearly indicate any section(s) to be marked.

CONT'D...

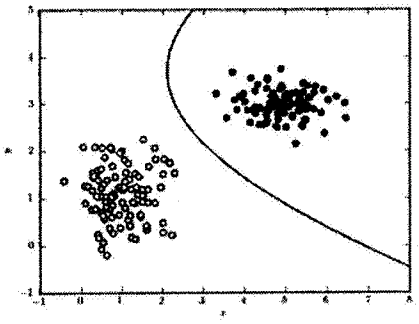
Question 2. [10 MARKS]

Part (a) [4 MARKS]

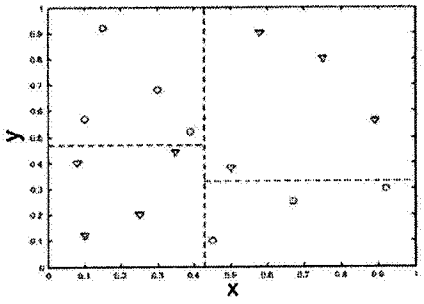
Match the decision boundaries with the models that produced them by drawing a line connecting the decision boundary image on the left and the name of the model on the right. Each decision boundary image should be matched with exactly one model.



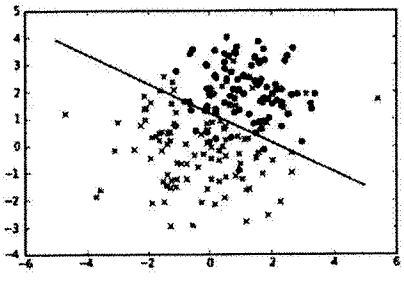
Gaussian Discriminant Analysis



Logistic Regression



Neural Network



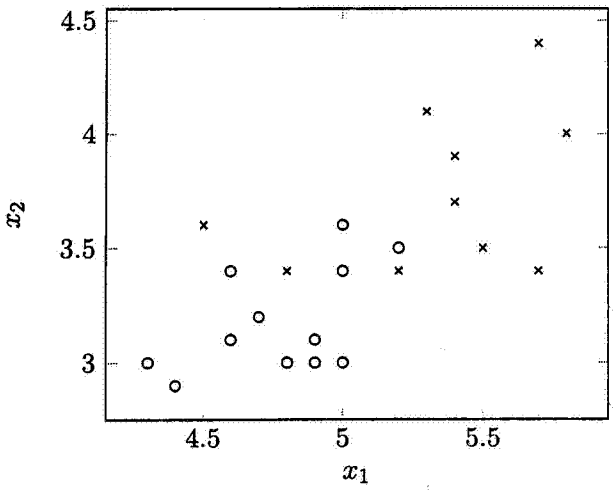
Decision Tree

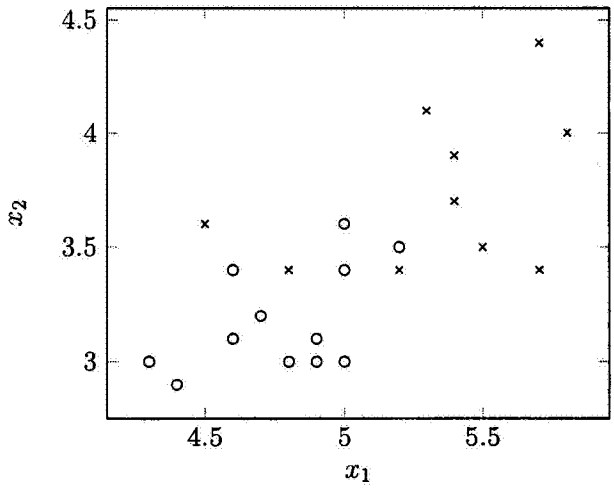
Use this page for rough work—clearly indicate any section(s) to be marked.

CONT'D...

Part (b) [6 MARKS]

Suppose Alice wants to perform binary classification on the dataset below, with the two input features x_1 and x_2. Alice decided to first run PCA (on the entire centered dataset, ignoring class labels) to obtain the two principal components u_1 and u_2. Alice then built a decision tree with maxdepth=2 on the new features she obtained by projecting the data onto u_1 and u_2. On the **left** figure, draw and label the two principal components u_1 and u_2. Assume the components are given in order of decreasing eigenvalues. On the **right** figure, draw a plausible decision boundary for the classifier Alice built.





Use this page for rough work—clearly indicate any section(s) to be marked.

CONT'D...

Question 3. [15 MARKS]

Suppose you have a dataset of news headlines, as in Project 3. Let $x^{(i)}$ represent the ith headline, with

$$x_j^{(i)} = \begin{cases} 1 \text{ if word } j \text{ is in headline } i, \\ 0 \text{ otherwise.} \end{cases}$$

Unlike in Project 3, we have no class labels. We would like to cluster the headlines into 10 clusters using the EM Algorithm.

We will make similar assumptions about $x_j^{(i)}$ to what we did in Project 3 when applying the Naive Bayes algorithm: that the x_js come from a Bernoulli distribution, and that the x_js are conditionally independent given the clusters.

Part (a) [2 MARKS]

List all the parameters of the model. Use i to index over different headlines, j to index over different words, and k to index over different clusters. Clearly define any other notation that you introduce.

Part (b) [3 MARKS]

What is the likelihood $P(X)$ for this model, where X is our entire training set? Write out your final answer below. Your answer should be expanded so that all the probability computations are explicit. Use the parameters you listed in Part (a), and clearly define any other notation that you introduce.

$P(X) =$

Use this page for rough work—clearly indicate any section(s) to be marked.

CONT'D...

Part (c) [5 MARKS]

Derive the E-step for this model. Your answer should include a mathematical justification. **Draw a rectangle around the formula that you derived.**

Part (d) [5 MARKS]

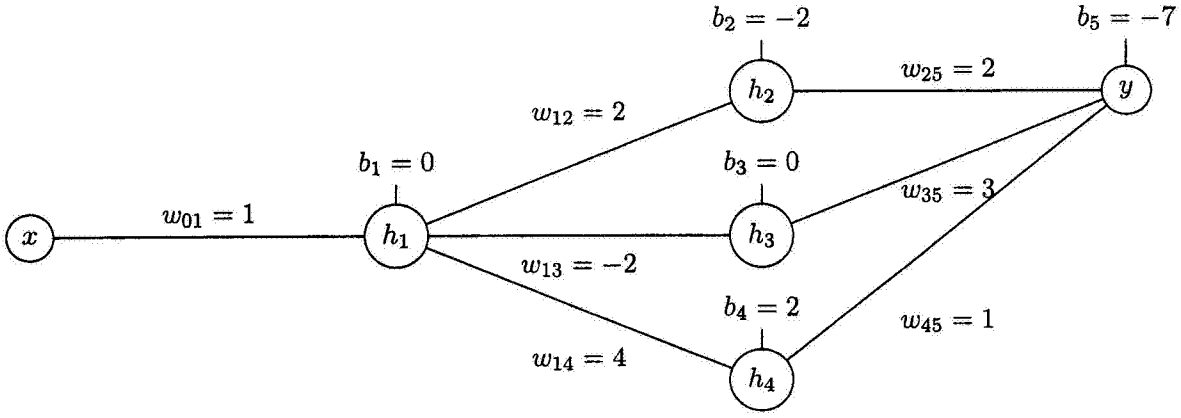
Derive the M-step for this model. Your answer should include a mathematical justification. **Draw a rectangle around the formula that you derived.**

Use this page for rough work—clearly indicate any section(s) to be marked.

CONT'D...

Question 4. [15 MARKS]

Suppose you are training the following neural network to perform regression. You decide to use the quadratic cost function $Cost(y, y^*) = (y - y^*)^2$ where y is the predicted value and y^* is the ground truth value. The hidden units h_1, h_2, h_3 and h_4 all use the ReLU activation function, and the output neuron y uses no activation function. The current weights of the network are shown below:



Part (a) [5 MARKS]

Compute the forward pass for $x = 2$. What are the post-activation values (i.e, the units' outputs) of h_1, h_2, h_3, h_4 and the prediction y?

$h_1 =$

$h_2 =$

$h_3 =$

$h_4 =$

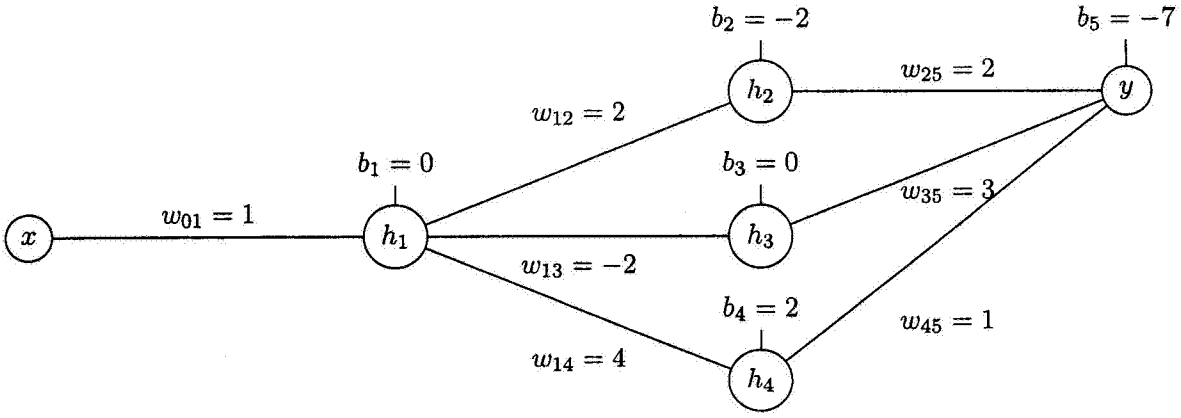
$y =$

Use this page for rough work—clearly indicate any section(s) to be marked.

CONT'D...

Part (b) [10 MARKS]

For the data point $x = 2$ in the training set, the ground truth output is $y^* = 8$. Use Backpropagation to compute the gradient $\frac{\partial Cost}{\partial w_{01}}$, where the cost is computed for the data point $(2, 8)$. Recall that you are using the quadratic cost function $Cost(y, y^*) = (y - y^*)^2$. Show each step in your computation, in order. The same network diagram is included again below for your convenience.



Use this page for rough work—clearly indicate any section(s) to be marked.

CONT'D...

Question 5. [10 MARKS]

Part (a) [5 MARKS]

Assume we know the principal component directions $u_1, u_2, \ldots, u_K$ for a training set of faces, as well as the mean face $\hat{\mu}$. State how to to compute the projection of face x on the space spanned by the first k principal components $u_1, u_2, \ldots, u_k$.

Part (b) [5 MARKS]

We discussed both the Kernel Trick and PCA in class. Kernel PCA is a combination of the two ideas: for training data $x^{(i)}$, we obtain the principal components of the transformed training set $\phi(x^{(i)})$. The key insight of the Kernel PCA algorithm is that a principal component of the transformed dataset can be represented as $v = \sum_{i=1}^{n} \alpha_i \phi(x^{(i)})$ where n is the total number of training examples, and α_i are scalar coefficients. You do not need to prove this.

Show mathematically that one can compute the projection of $\phi(x)$ on v without computing the features $\phi(x)$ using the Kernel Trick.

Use this page for rough work—clearly indicate any section(s) to be marked.

CONT'D...

Question 6. [10 MARKS]

Suppose you are in the process of training a Generative Adversarial Network (GAN). Your discriminator is D_{θ_d} and your generator is G_{θ_g}. You have just updated the discriminator parameters θ_d. Write the pseudocode to perform an update to the generator parameters θ_g. Assume you can compute any derivatives. For example, you can compute $\nabla_{\theta_g} G_{\theta_g}(z)$. Assume you have access to a function $normal()$ that generates samples from the standard normal distribution $N(0, 1)$.

Use this page for rough work—clearly indicate any section(s) to be marked.

CONT'D...

Question 7. [10 MARKS]

In Project 4, we used Policy Gradients to train a Tic-Tac-Toe policy. We kept track of rewards r_t provided by the environment at each time step, and then computed the discounted returns

$$G_t = \sum_{s=t}^{T} \gamma^{T-s} r_s,$$

where T is the episode length, and γ is the discount factor. We trained our policy by performing gradient ascent on

$$\sum_{t=1}^{T} G_t \log \pi_\theta(a_t|s_t).$$

Bob decides that he wants to train the policy parameters using rewards rather than returns, and perform gradient ascent on

$$\sum_{t=1}^{T} r_t \log \pi_\theta(a_t|s_t).$$

Part (a) [3 MARKS]

Why is this a bad idea for Tic-Tac-Toe?

Part (b) [2 MARKS]

Propose a Reinforcement Learning task for which Bob's idea could make sense, and briefly justify your answer.

Use this page for rough work—clearly indicate any section(s) to be marked.

CONT'D...

Part (c) [5 MARKS]

Consider a task with a discrete action space for which Bob's idea makes sense. Show that this task can be re-framed as a supervised learning problem. Make sure you include the definition of your policy, and the loss function for the supervised learning task.

Use this page for rough work—clearly indicate any section(s) to be marked.

CONT'D...

Question 8. [10 MARKS]

Suppose you have a dataset D that consists of data points and targets $(x^{(i)}, y^{(i)})$ with $i = 1 \ldots N$. You would like to train a regression model f to predict y given x. Propose a way to estimate the variance (in the sense of the Bias-Variance Decomposition) of f. You may assume that y is continuous. Use pseudocode.

Use this page for rough work—clearly indicate any section(s) to be marked.

CONT'D...

(This page is left intentionally blank)

Use this page for rough work—clearly indicate any section(s) to be marked.

CONT'D...

This page was intentionally left blank

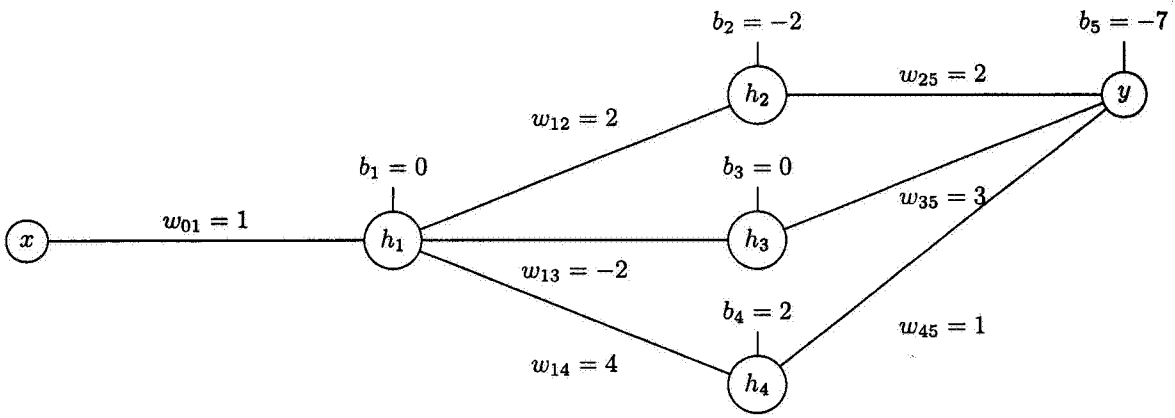
Use this page for rough work—clearly indicate any section(s) to be marked.

CONT'D...

This page was intentionally left blank

PLEASE WRITE NOTHING ON THIS PAGE

CSC411 Exam Draft Paper, APRIL 2018
For reference only – do not hand in



Do not hand in

Do not hand in

Do not hand in