University of Toronto
Faculty of Arts & Science
December 2017 Examinations

CSC411H1 F 2017 Final Test
Duration — 3 Hours
Aids allowed: none

Student Number: __________

Last Name: ____________________ First Name: ____________________

Do not turn this page until you have received the signal to start.
(Please fill out the identification section above and read the instructions below.)
Good Luck!

This final consists of 7 questions on 15 pages (including this one). When you receive the signal to start, please make sure that your copy is complete.

1. Do not turn the page until told to do so.
2. If a question asks you to do some calculations, you must show your work to receive full credit.
3. You can use either pen or pencil for the exam. But please be aware that you are not allowed to dispute any credit after the exam is returned if you use a pencil.
4. Use the back of the page if you need more space on a question. If you require additional paper you may request some from an invigilator.
5. If you use any space for rough work, indicate clearly what you do not want marked.
6. Note: A minimum of 30% will need to be scored on the final exam to ensure a passing grade for this course.
7. Lastly, enjoy the problems!

# 1:	____/ 15
# 2:	____/ 30
# 3:	____/ 15
# 4:	____/ 30
# 5:	____/ 35
# 6:	____/ 20
# 7:	____/ 20
TOTAL:	____/165

Question 1. [15 marks]

Mark whether the following statements are true or false by placing a tick in the corresponding column for each row.

Statement	True	False
A Neural Network with no hidden layers and logistic activation function in the output layer produces a linear decision boundary		
A Gaussian Naive Bayes classifier assumes a diagonal covariance matrix for the input features given the class labels		
Gaussian Discriminant Analysis has a quadratic decision boundary when we use a full covariance matrix		
Boosting improves performance by reducing variance		
Nearest neighbors scales well to high dimensions since it does not need to learn any parameters		

Question 2. [30 marks]

Explain each of the following terms. Be concise. Mentioning irrelevant information will **result** in losing marks. Try to use no more than 5 lines.

- Bagging

- Kernel trick

- Convolution layer

- Exploration vs exploitation

- Naive Bayes

- One-VS-all classifier

Question 3. [15 marks]

Part (a) [5 marks]

How can you tell if the model you are training is overfitting?

Part (b) [10 marks]

Describe two methods to reduce overfitting

Question 4. [30 marks]

For the following you do not need to write vectorized code - for loops are fine. Your code should be detailed enough to be implemented by somebody without machine learning expertise using a package like numpy.

Part (a) [20 marks]

Write pseudo-code implementing K-means clustering given inputs $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)} \in \mathbb{R}^d$ and number of clusters k. Your algorithm should use Euclidean distance. Clearly state your stopping condition.

Part (b) [10 marks]

Consider a differentiable loss function $\ell(\mathbf{w}, \mathbf{x}, y)$ and a dataset $D = (\mathbf{x}^{(1)}, y^{(1)}), ...(\mathbf{x}^{(n)}, y^{(N)})$. We aim to optimize the average loss $\frac{1}{N} \sum_{i=1}^{N} \ell(\mathbf{w}, \mathbf{x}^{(i)}, y^{(i)})$ with respect to $\mathbf{w}$.

Write pseudo-code implementing mini-batch SGD optimization with the following inputs:

1. Differentiable loss $\ell(\mathbf{w}, \mathbf{x}, y)$ with gradient $\nabla_{\mathbf{w}} \ell(\mathbf{w}, \mathbf{x}, y)$
2. Dataset D
3. Batch size m
4. Initial point $\mathbf{w}_0$
5. Number of steps T and learning rate policy $\alpha_1, ..., \alpha_T$.

Question 5. [35 marks]

In this question we will derive the EM algorithm for a mixture model with Bernoulli Naive Bayes components.

Consider a dataset consisting of inputs $\mathbf{x}^{(1)}, ..., \mathbf{x}^{(N)}$ which are binary $\{0, 1\}$ vectors of dimension d. We will model these points as being distributed according to a mixture of K Bernoulli Naive Bayes components.

Take $p(z = k|\pi) = \pi_k$ and the vector of parameters of the jth Bernoulli Naive Bayes component as $\boldsymbol{\mu}_j$. We write $\Theta = \{\pi, \boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_K\}$ to represent the collection of all model parameters. Then we have,

$$p(\mathbf{x}|z = k, \Theta) = \prod_{j=1}^{d} \mu_{kj}^{x_j}(1 - \mu_{kj})^{(1-x_j)}$$

Part (a) [5 marks]

Derive the explicit formula for the log-likelihood $\log(p(\mathbf{x}^{(1)}, ..., \mathbf{x}^{(N)}; \Theta))$.

Part (b) [5 marks]

Define $\gamma_{ik} = P(z = k|\mathbf{x}^{(i)}; \Theta^{old})$ where Θ^{old} are some fixed values of the parameters. Derive an expression for γ_{ik} in terms of these fixed parameter values and the data.

Part (c) [15 marks]

Derive the closed form solution for $\Theta^{new} = \arg\max_\Theta \sum_{i=1}^{N} \mathbb{E}_{P(z^{(i)}|\mathbf{x}^{(i)};\Theta^{old})}[\log(p(\mathbf{x}^{(i)}, z^{(i)}; \Theta))]$ in terms of γ_{ik} and the data. You only need to optimize $\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_K$.

This page is left intentionally blank.

Part (d) [10 marks]

Write pseudo-code implementing the EM algorithm for optimizing a mixture of Bernoulli Naive Bayes components. Fix $\pi_k = 1/K$.

Question 6. [20 marks]

To show that a function, $k(\mathbf{x}, \mathbf{y})$, is a kernel it is sufficient to show that its symmetric Gram matrix is positive semi-definite. That is, the matrix $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ satisfies $\mathbf{x}^T K \mathbf{x} \geq 0$ for all $\mathbf{x}$. Equivalently we could show that we can write $k(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle$ for some mapping ϕ.

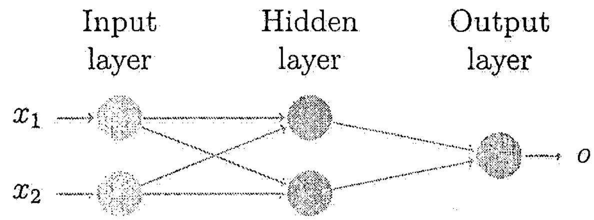
Prove the following properties of kernels:

1. The function $k(\mathbf{x}, \mathbf{y}) = \alpha$ is a kernel for $\alpha > 0$.
2. $k(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}) \cdot f(\mathbf{y})$ is a kernel for all $f : \mathbb{R}^d \to \mathbb{R}$.
3. If $k_1(\mathbf{x}, \mathbf{y})$ and $k_2(\mathbf{x}, \mathbf{y})$ are kernels then $k(\mathbf{x}, \mathbf{y}) = a \cdot k_1(\mathbf{x}, \mathbf{y}) + b \cdot k_2(\mathbf{x}, \mathbf{y})$ for $a, b > 0$ is a kernel.
4. If $k_1(\mathbf{x}, \mathbf{y})$ is a kernel then $k(\mathbf{x}, \mathbf{y}) = \frac{k_1(\mathbf{x}, \mathbf{y})}{\sqrt{k_1(\mathbf{x}, \mathbf{x})}\sqrt{k_1(\mathbf{y}, \mathbf{y})}}$ is a kernel (hint: use the features ϕ such that $k_1(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle$).

This page is left intentionally blank.

Question 7. [20 marks]

Consider a 2-layer neural network. The network has two input units, two hidden units and a single output unit. For this question we do not include a bias term in any layer of the network.



The hidden layer uses a Sigmoid activation function: $f(x) = \dfrac{1}{1+e^{-x}}$.

Part (a) [2 marks]

How many total parameters does the network contain? Do not count hyperparameters.

Part (b) [5 marks]

Denote the network parameters in layer j by the matrix $\mathbf{W}^j$ (for $j = 1, 2$). Write an expression for the neural network output, o, using the inputs $\mathbf{x} = (x_1, x_2)$ and the network parameters.

Part (c) [13 marks]

Consider training this network to minimize the squared error: $\ell(o, t) = (o - t)^2$. Derive the **gradient** of each layer's parameters using the backpropagation algorithm.

Print your name in this box.

End of Examination