UNIVERSITY OF TORONTO AT SCARBOROUGH
Department of Computer and Mathematical Sciences

DECEMBER 2014 EXAMINATIONS

CSCC11H: Machine Learning and Data Mining

Duration: 3 hours

Aids allowed: None

There are 16 pages total (including this page)

Please answer questions on the exam pages in the space provided. You may use the back of the pages for scratch work but final answers should be in the provided spaces. Read questions carefully and answer as neatly, clearly and concisely as possible; illegible or incomprehensible answers will get no marks. Should a question be unclear or ambiguous, make a *reasonable* interpretation and state what you have assumed before answering. Partial credit will be given for clear formulations of how to solve the problems.

Family name: _____

Given names: _____

Student number: _____

Signature : _____

By signing above, I certify that the work contained within is my work and my work alone and understand that copying another students work or allowing another student to copy my work is a serious academic offence.

Question	Marks
1	_____/ 23
2	_____/ 15
3	_____/ 12
4	_____/ 20
5	_____/ 21
6	_____/ 9
Bonus	_____
Total	_____/ 100

1 Short Answer [23 marks]

(a) **[2 marks]** Ensemble methods combine together many simple, poorly performing classifiers in order to produce a single, high quality classifier. Name two ensemble methods.

(b) **[2 marks]** What is the principal assumption in the Naive Bayes' model, and when is this assumption useful?

(c) **[4 marks]** Describe the K-fold cross-validation algorithm for model selection.

Question 1 continued.

(d) **[1 marks]** Gradient Descent and Stochastic Gradient Descent are very similar in that they both iteratively update the parameter vector based on the gradient of a function. Describe either in words how they different.

(e) **[2 marks]** Stochastic Gradient Descent is motivated by the idea that there exists (in the universe) an infinite amount of data which is distributed according to $p(x, y)$. Then, given some parameters θ and a loss function $L(x, y|\theta)$ we wish to minimize the *expected loss*. Define mathematically the expected loss.

Question 1 continued.

(f) **[6 marks]** In class we discussed three different kinds of Unsupervised Learning problems. List the three types of problems and for each name a method which addresses that problem.

(g) **[3 marks]** What is the *Product Rule* of probability? Define it and use it to derive *Bayes' Rule*.

(h) **[3 marks]** Use *Bayes' Rule* to write the probability of the parameters θ given the data $\mathcal{D}$. Specify which elements correspond to the *posterior*, the *likelihood*, the *prior* and the *evidence*.

2 Basis Function Regression [15 marks]

Suppose you are given a dataset of N training pairs $\{(x_i, y_i)\}_{i=1}^N$, such that $x_i \in \mathbb{R}, y_i \in \mathbb{R}$. We want to fit the following model to the data:

$$y = f(x) + n, \quad \text{where } f(x) = \sum_{j=1}^{K} w_j \sin(cx + d_j) , \quad \text{and } n \sim \mathcal{N}(0, 1/\alpha) . \tag{1}$$

Model parameters include the weights $\mathbf{w} = [w_1, ...w_K]^T$, the frequency c, the phase shifts $\mathbf{d} = [d_j ... d_K]$, and the variance $1/\alpha$.

(a) **[2 marks]** Provide an analytic expression for $p(y|x, c, \mathbf{w}, \alpha, \mathbf{d})$.

(b) **[2 marks]** Formulate the likelihood of the training data given this model, $p(y|x, c, \mathbf{w}, \alpha, \mathbf{d})$. Simplify where possible and include all terms and constants.

(c) **[3 marks]** Write the negative log-likelihood, and simplify as much as possible. Include all terms and constants. To simplify the remainder of the question, it is recommended that you write the summation in terms of a dot product (*i.e.*, $\mathbf{w}^T\mathbf{b}$).

Question 2 continued.

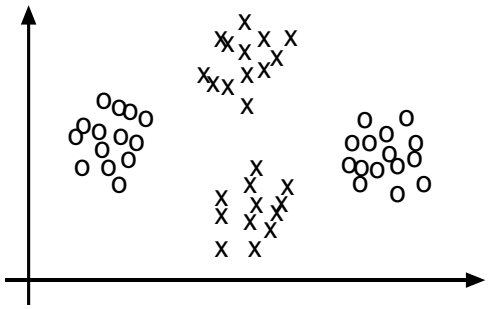
(d) **[4 marks]** Suppose all parameters but $\mathbf{w}$ are known. Give a closed-form expression for the optimal weight vector $\mathbf{w}$ that maximizes the data likelihood.

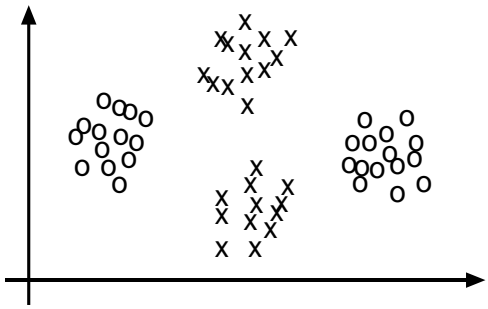
Question 2 continued.

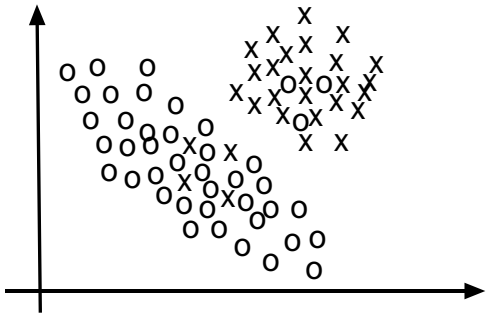
(e) **[4 marks]** Suppose we assume a Gaussian prior on the weights $\mathbf{w} \sim \mathcal{N}(0, \beta\mathbf{I})$. Derive the posterior distribution over the weights $\mathbf{w}$, given c, α, and $\mathbf{d}$. What type of distribution is this posterior?

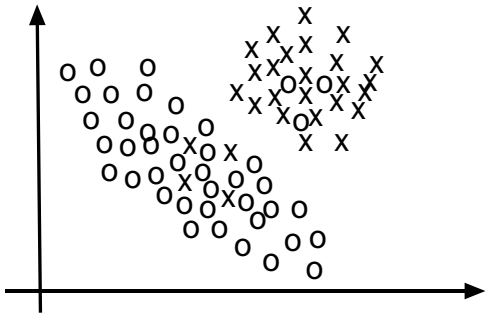
3 Classification [12 marks]

Consider the following classification algorithms: Gaussian Class Conditional, K-Nearest Neighbours (K=1), K-Nearest Neighbours (K=5), Logistic Regression and AdaBoost. In this question for each dataset below name the classification algorithm which would work best and draw its decision boundary on the left plot. Similarly, for each dataset, name the classifiction algorithm which work worst and draw its decision boundary on the right plot. For both algorithms describe briefly why it is the best/worst choice for the dataset. Note: there may be more than one right answer for these!









4 PCA [20 marks]

(a) **[4 marks]** PCA assumes a specific relationship between the unobserved latent coordinates $\mathbf{x}$ and the observed data points $\mathbf{y}$. Express this relationship as an equation. Clearly identify and name the parameters which are learned.

(b) **[2 marks]** What objective function is minimized to learn the parameters of PCA?

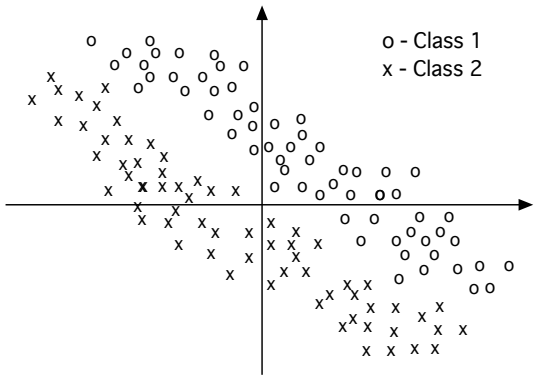
(c) **[2 marks]** Probabilistic PCA is very similar to regular PCA but differs in two key regards: the assumptions that it makes and what it produces. Precisely describe these differences.

Question 4 continued.

(d) **[4 marks]** Specify the steps of the PCA algorithm.

Question 4 continued.

(e) **[4 marks]** Suppose, you are given 2D feature vectors for a classification task which are distributed according to the figure below. You apply PCA to the data from each class seperately, keeping only a single eigenvector in each case. On the figure, for each class draw the location of the mean and the direction of the eigenvector which corresponds to the largest eigenvalue.



(f) **[2 marks]** Explain briefly how you might use the results of PCA on each class seperately to perform classification.

(g) **[2 marks]** A more principled, probabilistic approach to classification could be found by using Probabilistic PCA and a generative classification algorithm we learned in class. Name the classification algorithm. Would it be better, worse or about the same? Why?

5 Clustering [21 marks]

(a) **[2 marks]** What is the name of the algorithm used when fitting a Gaussian Mixture Model?

(b) **[2 marks]** The algorithm for fitting Gaussian Mixture Models indirectly optimizes the likelihood of the data by instead optimizing the *free energy*. What is the relationship between the free energy and the likelihood of the data?

(c) **[2 marks]** For the K-Means algorithm: What are the inputs? Which parameters are usually specified by the user, and what does the algorithm estimate?

(d) **[3 marks]** What objective function does the K-Means algorithm minimize?

(e) **[5 marks]** Specify the steps of the K-Means algorithm.

(f) **[2 marks]** Explain how you know that this algorithm will always converge.

Question 5 continued.

(g) **[2 marks]** The K-Means algorithm will in general converge to a local optima rather than a global one. Given this, how would you adapt the algorithm to increase the chances of finding a good solution?

(h) **[3 marks]** Sketch a dataset on which K-Means would work poorly but a Gaussian Mixture Model with the same number of clusters would do well. Describe why K-Means wouldn't work well.

6 Vignettes [9 marks]

Below three different scenarios are described where machine learning might be applied. For each scenario, indicate what machine learning algorithm you would apply and why.

(a) **Scene Identification** Most cameras have the ability to change their settings (ISO, shutter speed, white balance, etc) to improve the quality of the image based on the content (urban, landscape, night, indoors, etc). Unfortunately, this has to be done manually but you would like to create software which will automatically identify the type of scene and allow the camera to automatically adjust these settings. To begin with you'll only consider trying to determine if an image is indoors or not. Someone has already labelled a large number of images as indoors or outdoors as well as extracted a compact vector representation. You need to build a classifier which will use these vector representations and classify their corresponding images as indoors or outdoors.

(b) **Document Classification** A company has a large number of documents that need to be sorted into one of three categories: Research & Development, Finance or Marketing. They have been able to identify a number of phrases which are commonly used in these documents and may help disambiguate them but there are a large number (thousands) of these phrases and each one only appears in a small number of documents. Thankfully someone has labeled a few hundred documents for you but you need to build a method to automatically label the rest.

(c) **High-Frequency Trading** In high-frequency trading computers are used to make predictions about the value of a stock, currency or other financial instrument in the very near future (sometimes less than a second) based on the recent past. Data comes very quickly, with there being potentially thousands or more trades every second and predictions need to be constantly made. Making such predictions can be thought of as a regression problem where the input is the value of the instrument at several points in recent history and the output is the value in the future.

Bonus Question

Write and answer your own exam question based on the material in the course. This question can be on anything we covered in class or was in the course notes. This is your chance to show what you learned that may not have been otherwise tested in the exam. Up to four bonus marks will be given based on difficulty, correctness and originality of the question. No marks will be given for reproducing a question from this exam.

END OF EXAM