UNIVERSITY OF TORONTO

Faculty of Arts and Science

December 2013 Final Examination

CSC411H1 F

Duration - 2 hours

No Aids Allowed

Please check that your exam has 11 pages, including this one.

Use the back of the page if you need more space on a question.

Point Distribution

Problem 1:	20
Problem 2:	20
Problem 3:	20
Problem 4:	20
Problem 5:	10
Problem 6:	10
Total:	**100**

Name:

Student Number:

1. Ensemble Methods [20 points].

(A). What is the main difference between bagging and boosting?

(B). What are the two criteria when selecting individual classifiers for boosting? Explain them in one sentence each.

(C). How do the functions that control how the ensemble components are combined differ between the mixture-of-experts algorithm and boosting? How does this affect the resulting model?

(D). In a binary classification problem, we have 5 classifiers with accuracies (0.4, 0.5, 0.6, 0.7, 0.8). Assume that the errors of these individual classifiers are perfectly uncorrelated. Write down a good decision function that involves combining the binary (+1,-1) outputs of each component classifier, and justify it.

2. Mixture Models [20 points].

   (A). Describe 3 differences between the standard K-means algorithm and the EM algorithm for fitting a mixture of Gaussians.

   Consider a simple form of mixture model, in which each mixture component is a spherical Gaussian density of dimension d:

$$\begin{aligned} p(\mathbf{x}|\theta) &= \sum_{k=1}^{K} P(z=k|\theta) p(\mathbf{x}|z=k,\theta_k) \\ p(\mathbf{x}|z=k,\theta_k) &= \frac{1}{(2\pi\sigma_k^2)^{d/2}} \exp\left(-\frac{|\mathbf{x}-\mu_\mathbf{k}|^2}{2\sigma_k^2}\right) \end{aligned}$$

   (B). When learning this model using the EM algorithm, which random variable's expected value is computed in the E-step?

(C). In the M step, what objective function is maximized, and with respect to which variables?

(D). How will increasing the number of components impact the likelihood of the training data? For example, how do you expect the training data likelihood with K components compares to that with $K + 1$ components, for a general K?

(E). How will this increase in K impact the likelihood on the test data?

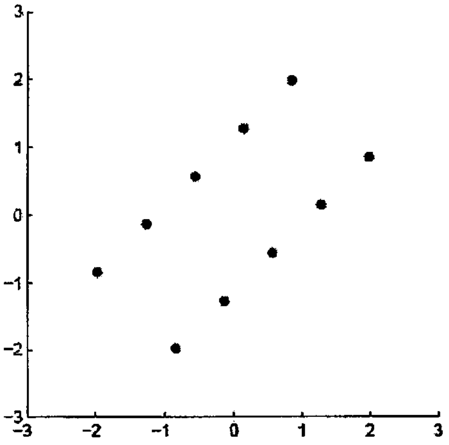
3. Unsupervised Learning [20 points].

PCA should be used with caution for classification problems, because it does not take information about classes into account. Here you will show that, depending on the dataset, the results may be very different. Suppose that the classification algorithm is 1-nearest-neighbor (1-NN), the data is 2-dimensional and PCA is used to reduce the dimensionality of the data to 1 dimension. There are 2 classes (+ and -). The datapoints (without class labels) are plotted below (the two plots are identical).

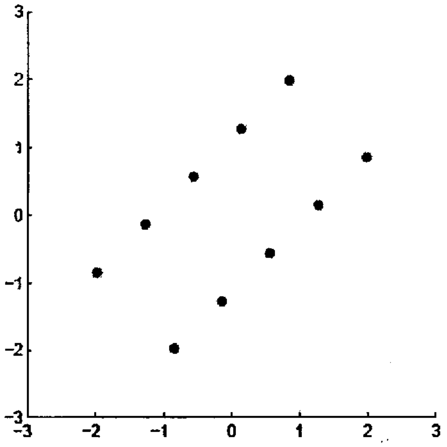
(A). On one of the plots draw the line that PCA will project the datapoints to.

(B). For each of the plots, assign the datapoints to the two classes so that 1-NN will have the following leave-one-out cross-validation error:

2D data: 100% error
1D data from PCA: 0% error



2D data 0% error
1D data from PCA: 100% error

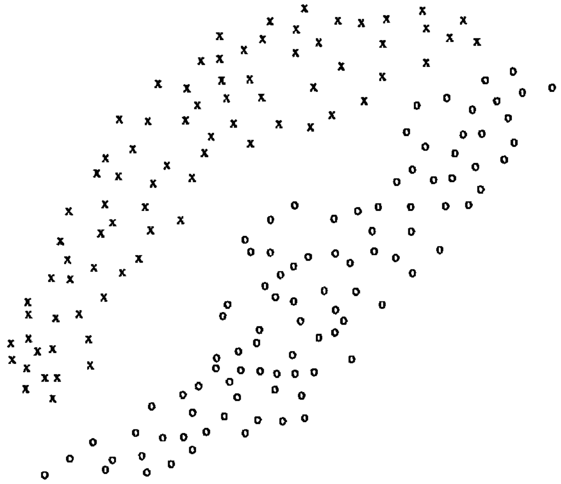


(C). Principal Components Analysis can be formulated as minimizing two very different objective functions – what are they?

(D). X is a dataset describing N points each with D features (dimension $D > 3$). We want to project X into a two-dimensional space using PCA. How is the best plane of projection defined (describe in relation to the data X)?

(E) .How do we judge whether our projection on only 2 dimensions is a good representation of the original data?

(F). If the projections on two dimensions are as shown below, with each datapoint in one of two classes as denoted by different symbols, do you expect a single run of K-means to work well on this projection? Justify your answer with at least two reaons.



4. Support Vector Machines [20 points].

(A). Consider fitting a linear SVM to the following data set:

Data Case	x_1	x_2	Class
1	3	0	+
2	-1	-2	-
3	1	2	+
4	-1	0	-

Plot the points below with x_1 along the x-axis and x_2 along the y-axis. Write down the decision rules the linear SVM will find for class (+) and class (-) in terms of x_1 and x_2. Draw the decision boundaries. What is/are the support vector(s) for this data set?

(B). What is the kernel trick, and what computational purpose does it serve?

(C). How does a non-linear SVM differ from a linear SVM?

(D). Describe two important differences between a non-linear SVM and a neural network.

5. Bayesian Learning [10 points].

Recall the equation for predictive probability in a Bayesian learning system:

$$P(y_{new}|\mathcal{D}) = \int P(y_{new}|\theta)P(\theta|\mathcal{D})d\theta$$

(A). Briefly define the two terms in the integral, and how they can be computed.

(B). Name two advantages and two disadvantages of this approach relative to standard learning.

6. Graphical Models [10 points].

   (A). How does a Restricted Boltzmann Machine differ from a standard neural network, with respect to both inference and learning?

   (B). How does conditioning on some visible variables affect inferring the posterior distribution in an RBM?

END OF EXAM