UNIVERSITY OF TORONTO

Faculty of Arts and Science

December 2014 Final Examination

CSC411H1 F

Duration - 2 hours

No Aids Allowed

Please check that your exam has 12 pages, including this one.

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

Point Distribution

Problem 1:	16
Problem 2:	20
Problem 3:	16
Problem 4:	16
Problem 5:	20
Problem 6:	12
Total:	**100**

Name:

Student Number:

1. Ensemble Methods [16 points].

(A). What is the key underlying idea that may allow an ensemble method to achieve lower error rate than a single model?

(B). For each of these ensemble methods—bagging, boosting, and mixture-of-experts—explain how they are designed to achieve this.

(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). The K-means algorithm can be seen as a special case of the EM algorithm. Describe the steps in K-means that correspond to the E and M steps, respectively, and what happens in each step.

(B). Which algorithm would you expect to get stuck more often in less favorable solutions, hard or soft K-means? Explain your answer.

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

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

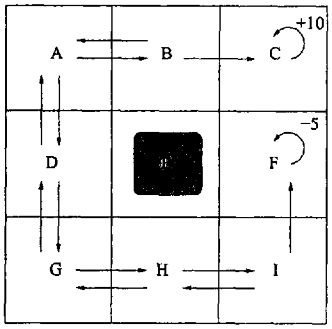
where $\theta_k = (\pi_k, \mu_k, \sigma_k)$.

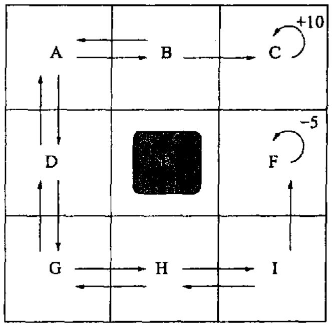
(C). What does the random variable z represent, and how is it updated in the EM algorithm?

(D). What is the objective that is optimized by hard K-means?

(E). Does the EM algorithm applied to a mixture of Gaussians optimize the same objective? Explain your answer.

3. Reinforcement Learning [16 points].





Consider the robot navigation task shown on the left above. The possible actions in each state are depicted by the arrows. The central state is an obstacle, so the robot cannot move into that state. The rewards are $+10$ for moving into state C and -5 for moving into F; these are both absorbing states. The reward for moving into every other state is 0.

(A). Assume that the state transitions are deterministic. Recall that under the simple Q-learning algorithm, the estimated Q values are updated using the following rule:

$$\hat{Q}(s,a) = r + \gamma \max_{a'} \hat{Q}(s',a')$$

Consider applying this algorithm when all the $\hat{Q}$ values are initialized to zero, and $\gamma = 0.8$. Write all the Q estimates on the left-hand figure, after the robot has executed the following state sequences: GDABC, GHIF, GHGDABC.

(B). Indicate the optimal Q values on the right-hand figure.

(C). What is the exploration-exploitation dilemma? How does it relate to the Q-learning algorithm?

(D). Describe one strategy to trade off these two approaches.

4. PCA and Autoencoders [16 points].

(A) Provide three reasons why one may want to use dimensionality reduction.

(B) Explain two ways to interpret the objective function of PCA (what objective is it minimizing/maximizing)? Why are these good?

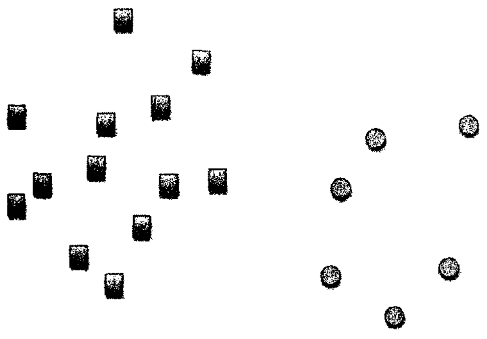
(C) Is PCA robust to outliers? Why?

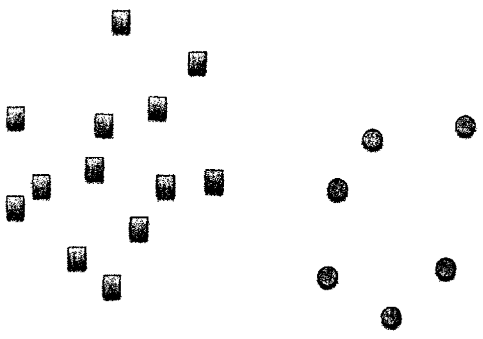
(D) What's an autoencoder? When is it equivalent to PCA?

(E) Draw an autoencoder for the case that the inputs are $x \in \Re^{10}$ and we want to perform dimensionality reduction to have $z \in \Re^{3}$. What is the input? and the output? How many hidden units do you have? How many weights in total?

5. SVMs [20 points].

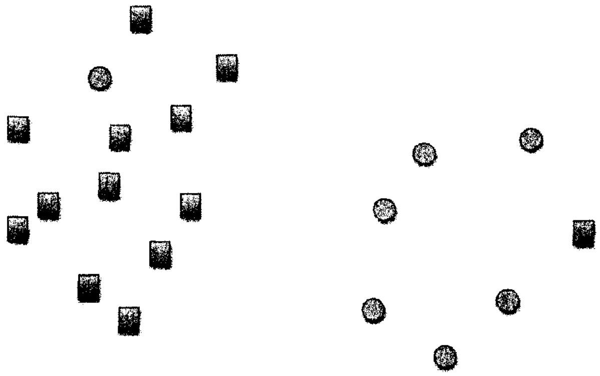
(A) Draw the decision boundary of an SVM in the figure on the left.





(B) What is the objective function of an SVM? Explain its geometric interpretation utilizing the figure on the right, i.e., **w**, margin, support vectors.

(C) What is the objective function of an SVM when the problem is not linearly separable? Utilize the figure below to explain the changes in the objective.



(D) When is it preferable to use the primal vs. the dual formulation of SVM. Remember that the dual is the Lagrangian formulation derived from the objective function in (B).

(E) What is the kernel trick? Why is it beneficial? What's the condition on a kernel such that I can use the kernel trick?

6. Bayesian methods [12 points].

(A) Explain Bayesian Linear Regression. What is the model? What is the prior, and the noise distribution? What about the likelihood? How can you compute the posterior?

(B) How is learning performed in this model?

(C) How is prediction formulated?