UNIVERSITY OF TORONTO

Faculty of Arts and Science

December 2016 Final Examination

CSC411H1 F

Duration - 2 hours

No Aids Allowed

Please check that your exam has 13 pages, including this one.

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

Point Distribution

Problem 1:	14
Problem 2:	16
Problem 3:	16
Problem 4:	20
Problem 5:	16
Problem 6:	18
Total:	**100**

Name:

Student Number:

1. Ensemble Methods [14 points].

(A). Describe the following concepts: bias, variance, and irreducible error.

(B). Explain why the following ensemble models work — bagging and boosting — from the bias-variance trade-off perspective.

(C). 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. Clustering and Mixture Models [16 points].

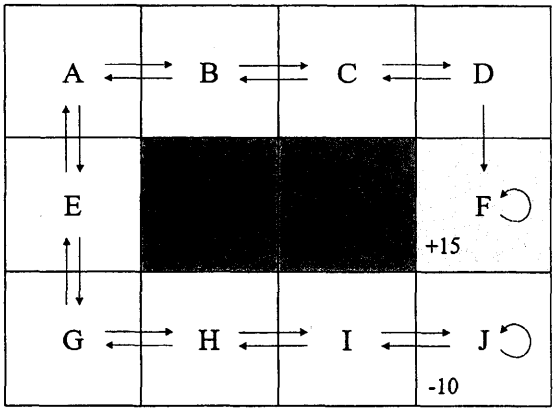
(A). Write down the objective function of the k-means algorithm, and explain your notation.

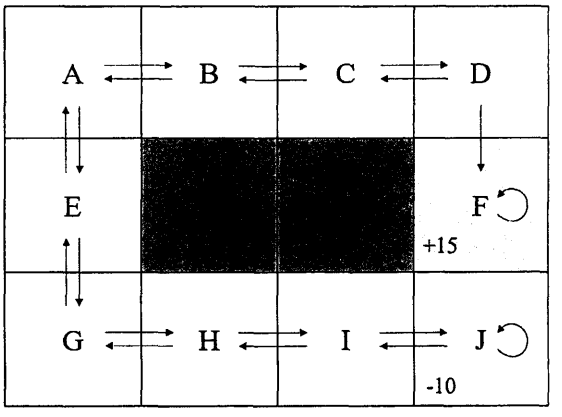
(B). Write down the pseudo code of the k-means algorithm, based on the objective function. Hint: each iteration should include two steps.

(C). Will the k-means algorithm converge or not? Explain why.

(D). Does the EM algorithm applied to a mixture of Gaussians optimize the same objective as K-means? Why or why not?

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 two central blocks are obstacles, so the robot cannot move into that state. The rewards are $+15$ for moving into state F and -10 for moving into J; 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 figure, after the robot has executed the following state sequences: ABCDF, GHIJ, GHGEABCDF.

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

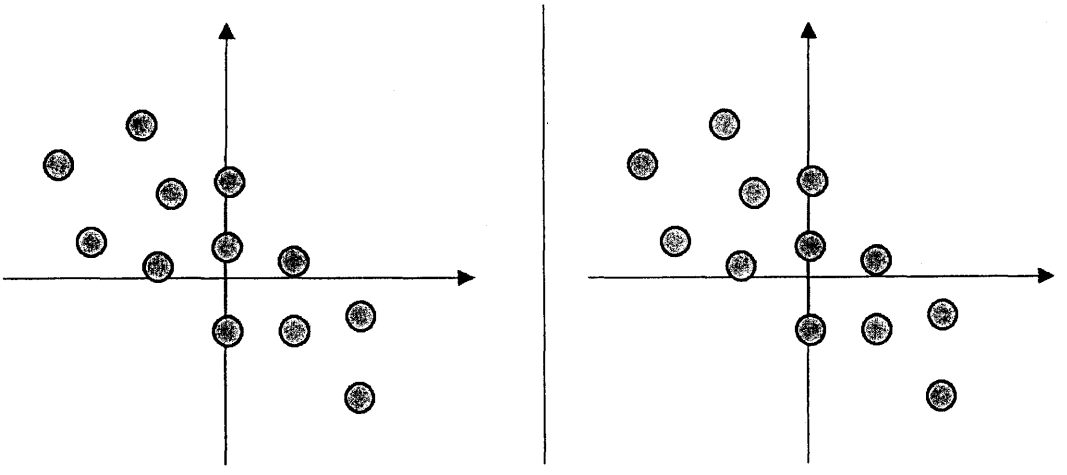
(C). What is the exploration-exploitation dilemma? How does it relate to the Q-learning algorithm?

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

4. PCA and Autoencoders [20 points].

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

(B) If we want to conduct a 1-D PCA over the 2D data shown in the following figure, interpret the objective function of PCA on the left figure. Interpret the principal component and the projection of the points on the right figure.

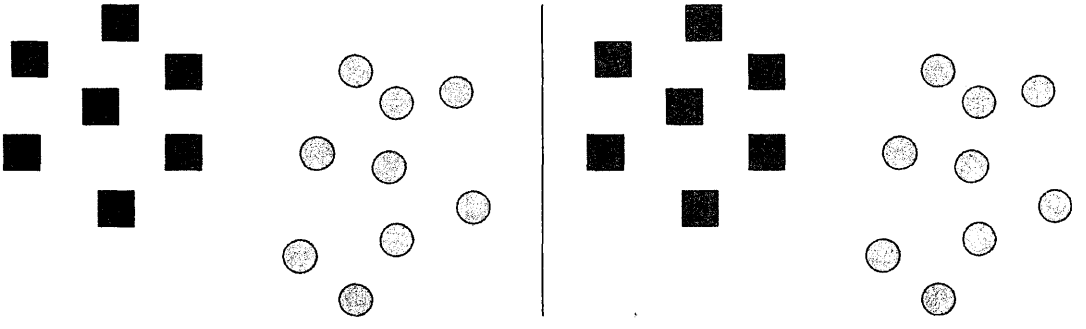


(C) What's an autoencoder? Describe the objective function of an autoencoder (in words or an equation). Explain (in words) when it is equivalent to PCA.

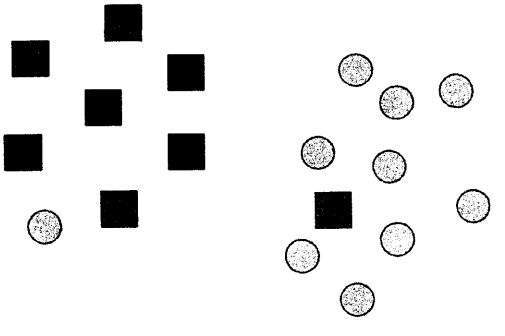
(D) Draw an autoencoder for the case that the inputs are $x \in \Re^5$ and we want to perform dimensionality reduction to have $z \in \Re^2$. What is the input? and the output? How many hidden units do you have? How many weights (including bias) in total?

5. SVMs [16 points].

(A) Draw the decision boundary of an SVM in the figure on the left. Explain its geometric interpretation utilizing the figure on the right, i.e., show what are **w**, margin, support vectors.

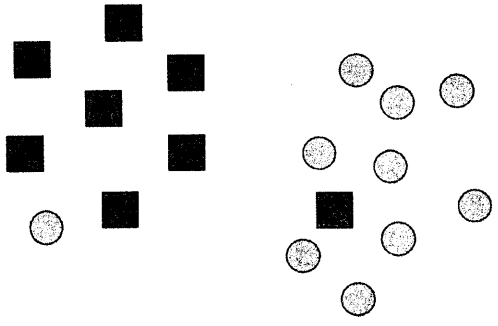


(B) How can the objective function of a linear SVM be modified when the problem is not linearly separable? Utilize the figure below to explain the changes in the objective.



(C) What is the kernel trick? What's the condition on a kernel such that one can use the kernel trick?

(D) Utilize the figure below to explain what the decision boundary looks like, if we use the kernel function $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{\|\mathbf{x}^{(i)}-\mathbf{x}^{(j)}\|}{2\sigma^2})$.



6. Neural Network [18 points].

(A) What is Soft-Max? Write the equation of $\sigma(z_j)$ as a function of $z_1, z_2, ..., z_K$. Provide an example where Soft-Max is used in neural network and explain why it is good to use Soft-Max.

(B) Compute the gradient of $\frac{\partial \sigma(z_j)}{\partial z_k}$. Notice the numerator is on z_j and denominator is on z_k. You can use the notation of $\sigma(z_j)$ and $\sigma(z_k)$ in the results.

(C) Given target distribution q and predicted distribution p, what is the cross entropy (write down equations)? In a classification problem of K classes, given the output values $z_1, z_2, ..., z_K$ and the ground truth label t, what is the cross entropy loss?

(D) What is the relation between Soft-Max and cross entropy loss?