UNIVERSITY OF TORONTO

Faculty of Arts and Science

December 2015 Final Examination

CSC411H1 F

Duration - 2 hours

No Aids Allowed

Please check that your exam has 10 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:	25
Problem 4:	35
Problem 5:	15
Total:	**115**

Name:

Student Number:

1. Ensemble Methods [20 points].

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

   - (5 pts) Bias-Variance Tradeoff: Justify ensemble methods from the point of view of (a) variance reduction and (b) bias reduction.

   - (5 pts) 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?

   - (5 pts) What is the key idea behing Bagging? Write down the steps needed to bag a regression model $y(x)$. How do you obtain the final prediction for regression after performing Bagging?

2. Mixture Models [20 points].

- (5pts) What is the key difference between K-means and soft K-means algorithm?

- (10 pts) 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 $\theta_k = (\pi_k, \mu_k, \sigma_k)$. What does the random variable z represent? Write down the update equation for z in the EM algorithm.

- (5 pts) Suppose that we have a dataset of observations: $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N)$, representing N observations of a multivariate variable $\mathbf{x}$. Assume that the observations are $i.i.d$ and are drawn from a K-component Gaussian mixture model with parameters $\{\pi_k, \mu_k, \Sigma_k\}$, $k = 1, ..., K$. Write down the log-likelihood function for this Gaussian mixture model.

3. Unsupervised Learning [25 points].

- (5 pts) Explain 3 reasons why one would want to do unsupervised learning

- (5 pts) Explain two ways to interpret the objective function of PCA. Write down the objective it is minimizing/maximazing, and argue why it is a good idea.

- (5 pts) Draw an auto encoder for the case that the inputs are $x \in \Re^7$ and we want to perform dimensionality reduction to have $z \in \Re^4$. What is the input? and the output? How many hidden units do you have? How many weights in total?

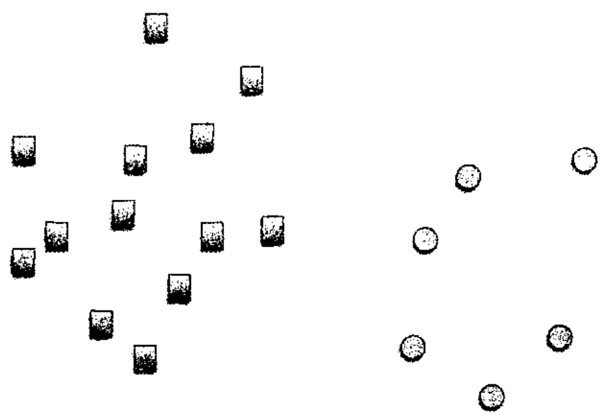
- (5 pts) When is an auto-encoder equivalent to PCA?

- (5 pts) Explain 3 ways to do regularization in auto-encoders.

4. SVMs [35 points].

- (5 pts) What is the primal optimization problem of an SVM? Write the exact equations and explain what they mean

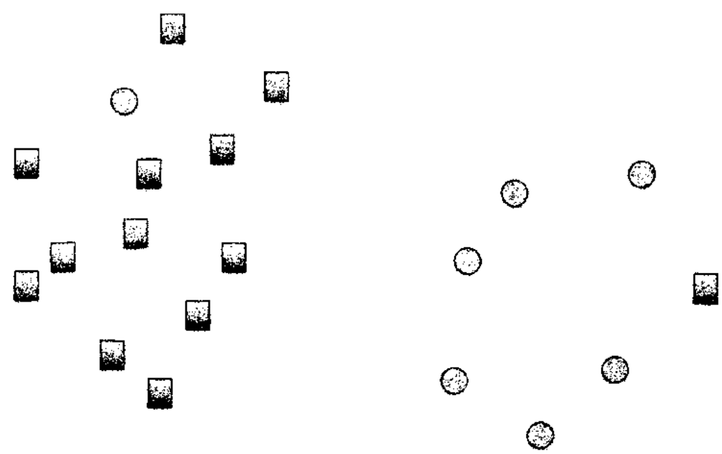
- (10 pts) Geometric interpretation of an SVM: draw w, the margin, the support vectors and the decision boundary on the following figure.



Explain also in words what w, the margin and the support vectors are, as well as why the shape of the decision boundary is what you draw.

- (5 pts) How can we make the decision boundary of an SVM non-linear?

- (5 pts) 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

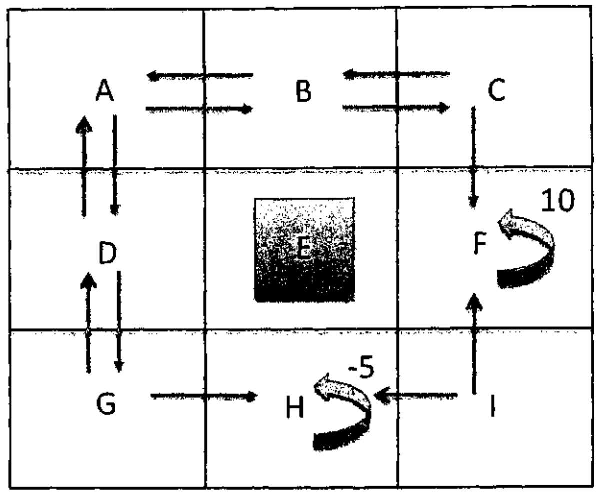


Write down the primal optimization problem in the non-linearly separable case.

- (5 pts) When is it preferable to use the primal vs the dual formulation of SVM. **Hint**: Remember that the dual formulation comes out of employing the method of Lagrange multipliers applied to the primal problem.

- (5 pts) What is the kernel trick? What's the advantage of using this trick? For which kernels can I use the kernel trick?

5. Reinforcement Learning [15 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 F and −5 for moving into H; these are both absorbing states. The reward for moving into every other state is 0.

- (5 pts) 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:

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

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: DABCF, DGH, ADGDABCF

- (5 pts) What is the exploration-exploitation dilemma? How can we take it into account when we do Q-learning?

- (5 pts) Why is reinforcement learning more difficult than supervised learning?