Qualifier: CS 6375
Machine Learning
Fall 2015

This exam contains 13 pages (including this cover page) and 6 problems. Check to see if any pages are missing.

You may **NOT** use books, notes, or any electronic devices on this exam. Examinees found to be using any materials other than a pen or pencil will receive a zero on the exam and face possible disciplinary action.

The following rules apply:

- **Organize your work**, in a reasonably neat and coherent way, in the space provided. Work scattered all over the page without a clear ordering will receive very little credit. Ask for additional paper if needed.
- **To ensure maximum credit** on short answer / algorithmic questions, be sure to **EXPLAIN** your solution.
- **Problems/subproblems** are not ordered by difficulty.
- **Do not** write in the table to the right.

Problem	Points	Score
1	15	
2	15	
3	15	
4	15	
5	14	
6	26	
Total:	100	

1. **Support Vector Machines:** Consider the following training data for a binary classification problem.

x_1	x_2	y
1	1	+
1	-1	-
-1	-1	+
2	2	+
-1	1	-

   (a) (3 points) Argue that the above data set is not linearly separable in $\mathbb{R}^2$.

   (b) (3 points) Is the above data set is linearly separable under the kernel $K(x^{(1)}, x^{(2)}) = (x^{(1)} \cdot x^{(2)})^2$? If so, provide a separator, if not, show why one cannot exist.

(SVMs continued)

(c) (9 points) For a collection of training data $(x^{(1)}, y_1), \ldots, (x^{(n)}, y_n)$, consider, as an alternative to the SVM algorithm, minimizing the following loss function.

$$\sum_i \left(\max\{0, -y_i f(x^{(i)})\}\right)^2$$

Can this loss function be minimized with gradient methods? If yes, derive the appropriate update rules, if no, provide an alternative method to minimize the loss.

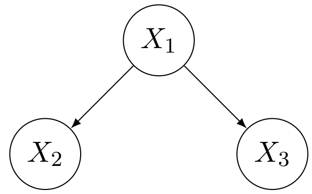
2. **VC Dimension:** Consider data points restricted to lie in a ball of radius r about the origin in $\mathbb{R}^2$. Suppose that each of the data points receives a label from the set $\{+, -\}$. Consider the hypothesis space of all linear classifiers such that the distance of any point in the data set to each classifier is at least d, for some nonnegative $d \in \mathbb{R}$.

   (a) (10 points) The VC dimension of this hypothesis space depends on both the data and d. What are all of the possible values of the VC dimension? Explain.

   (b) (5 points) If $d = \frac{3}{4}r$, what is the VC dimension of the hypothesis space? Prove it. (Hint: $\sin(60°) = \sqrt{3}/2$ and $\sin(30°) = 1/2$)

3. **Bayesian Networks:**

   (a) (10 points) Consider the following Bayesian network over three binary valued random variables with $p(X_1 = 1) = .2$, $p(X_2 = 1|X_1 = 0) = .4$, $p(X_3 = 1|X_1 = 0) = .8$, $p(X_2 = 0|X_1 = 1) = .5$, $p(X_3 = 0|X_1 = 1) = .3$.



1. What is $p(X_2 = 1, X_3 = 1)$?

2. What is $p(X_3 = 0)$?

3. What is $p(X_1 = 1|X_2 = 0)$?

4. How many parameters are required to estimate this Bayesian network?

5. Suppose now that the conditional probability tables are unknown. If we learn the parameters of the above model using the EM algorithm on the following data, what will $p(X_2 = 1|X_1 = 0)$ be at convergence?

x_1	x_2	x_3
0	?	1
0	1	?
0	1	?
1	?	?
0	1	1

(b) (5 points) Consider the following data set.

x_1	x_2	x_3
1	0	1
0	0	1
1	0	1
1	1	0
1	1	1

Which tree-structured Bayesian network maximizes the likelihood of the data?

4. **Maximum Likelihood Estimation:** Consider the following collection of univariate probability distributions over the integers, $\mathbb{Z}$, parameterized by $\theta \geq 0$.

$$p(x|\theta) = \frac{1}{2\theta + 1} I_{x \in [-\theta, \theta] \cap \mathbb{Z}}$$

where $I_{x \in [-\theta, \theta]}$ is the $0-1$ function indicator whether (1) or not (0) x is in the interval $[-\theta, \theta]$.

(a) (3 points) Given n data points $x^{(1)}, \ldots, x^{(n)} \in \mathbb{Z}$, what is the maximum likelihood estimate of θ?

(b) (2 points) If $\widehat{\theta}$ is the maximum likelihood estimator, what is the probability of a new data point $x^{(n+1)}$ under the model given by $\widehat{\theta}$.

(c) (4 points) Consider a prior distribution $p(\theta) \propto \exp(-(\theta - 2)^2/2) \cdot I_{\theta \geq 0}$. How does the MAP estimator (using this prior) compare to the maximum likelihood estimator for the data set $\{1\}$?

(MLE continued)

(d) (6 points) Suppose that the data points are actually drawn from $p(x|\theta)$ for $\theta = 10$. How many training samples do you need to guarantee that the maximum likelihood estimator is within one of the correct θ with probability at least 95%? You do not need to compute the precise number for full credit.

5. **Decision Trees:** Suppose that we want to consider the hypothesis space consisting only of "decision lists." That is, decision trees such that each node in the tree can have at most one non-leaf child. For this question, consider labeled data points in $\mathbb{R}^m$?

   (a) (4 points) Describe an algorithm to learn decision lists.

   (b) (3 points) What is the maximum depth of a decision list?

(c) (3 points) What is the size of the decision list hypothesis space?

(d) (4 points) What is the VC dimension of the hypothesis space of decision lists?

6. **Short Answer:**

   (a) (4 points) Explain the difference between bagging and boosting. Give one example of when you would use each of them.

   (b) (4 points) Explain why the backpropagation algorithm, without modificaiton, may perform poorly on large neural networks in practice.

(c) (10 points) If the best element in the hypothesis space overfits the training data, explain whether or not each of the following can be used to counteract overfitting.

1. Bagging
2. Boosting
3. Enlarging the training set
4. Changing the kernel space
5. Adding regularization

(d) (5 points) Consider a binary classification task over a data set with n data points and m features. Suppose that logisitic regression yields a perfect classifier for this data set. If the decision tree algorithm is also run on this data set, will it yield a perfect classifier? If yes, what is the maximum depth of the resulting tree. If no, why does the algorithm fail?

(e) (3 points) If complete DAGs are guaranteed to maximize the likelihood, why bother with the Chow-Liu algorithm?