Graduate Preliminary exam
Spring 2014 AI Sub-Area #3
Time Limit: 3 Hours

Name (Print): _____
Date: 2014-05-03
Student ID: _____

This exam contains 7 pages (including this cover page) and 20 problems. Check to see if any pages are missing. Enter all requested information on the top of this page, and put your initials on the top of every page, in case the pages become separated.

You may *not* use your books, notes, or any calculator on this exam.

You are required to show your work on each problem on this exam. The following rules apply:

- **If you use a "theorem" you must indicate this** and explain why the theorem may be applied.
- **Mysterious or unsupported answers will not receive full credit.** An incorrect answer supported by substantially correct calculations and explanations might receive partial credit.
- Write your answer in the exam answer book.
- Write everything in English.

Do not write in the table to the right.

Problem	Points	Score
1	6	
2	10	
3	8	
4	8	
5	8	
6	5	
7	5	
8	2	
9	3	
10	5	
11	3	
12	1	
13	3	
14	4	
15	4	
16	5	
17	3	
18	5	
19	2	
20	10	
Total:	100	

Part I: Theory

1. (6 points) Longest palindrome subsequence.

   A palindrome is a nonempty string over some alphabet that reads the same forward and backward. For example, aaaabaaaa, 00000, abcddcba are all palindrome. Give an efficient algorithm to find the longest palindrome that is a subsequence of a given input string. Prove the correctness of your algorithm. You will get full points if your algorithm runs in $O(n^2)$ time. Slower but still correct algorithm may get partial credits.

2. (10 points) Monge Matrix.

   An $m \times n$ array A of real number is a *Monge matrix* if

   $$A[i,j] + A[k,l] \leq A[i,l] + A[k,j], \forall i < k \text{ and } j < l.$$

   Answer the following questions. Let $f(i)$ the column index of the leftmost minimum element of row i. For example:

   $$\begin{pmatrix} 10 & 9 & 12 & 10 \\ 11 & 10 & 13 & 10 \\ 9 & 8 & 11 & 7 \\ 11 & 10 & 11 & 8 \end{pmatrix}$$

   is a Monge Matrix. $f(1) = 2, f(2) = 2, f(3) = 4, f(4) = 4$.

   1. (5 points) Show that $f(1) \leq f(2) \leq \ldots \leq f(m)$.
   2. (5 points) Given an $m \times n$ Monge matrix, Design an algorithm that computes $f(1), \ldots, f(m)$ in $O(m + n \log m)$ time.

3. (8 points) An interval graph is an undirected graph $G = (V; E)$ whose vertices correspond to intervals on the real line, where each interval is specified by a leftmost value v_1 and a rightmost value v_2. Two vertices in G are connected iff the corresponding intervals overlap. Let G be an interval graph whose corresponding intervals are provided. Give a polynomial time algorithm to find a maximum independent set in G. You need to prove the correctness of your algorithm. (An independent set in a graph is a subset of vertices such that no two vertices in the subset are joined by an edge.)

4. (8 points) Show that finding a min-cost matching with exactly k edges in a bipartite graph $G(U, V; E)$ (with positive edge costs) can be solved in polynomial time. Note that $|U|$ may not be the same as $|V|$. You can assume that the min-cost *perfect matching* problem in a bipartite graph can be solved in polynomial time.

5. (8 points) Finding the maximum area polygon.

   We are given a unit circle and n points on the circle. Design a polynomial time algorithm that, given a number $m < n$, finds m points (out of n points) such that the area of the polygon formed by the m points is maximized. (You can assume that you can compute the area of a given triangle in constant time.)

Part II: Artificial Intelligence

6. Search.

When are the following search algorithms optimal (i.e, find a solution that satisfies the goal conditions while incurs minimum cost)? Please be precise about the conditions you define.

(a) (1 point) Breath-first search;

(b) (1 point) Depth-first search;

(c) (3 points) A*.

7. Probability theory and probabilistic reasoning.

(a) (2 points) Suppose that you are a developer that wants to know more about the market before developing your app for a new social network platform called TalkNow (T).

Market researchers have partitioned the population into three groups: Young people (Y), Adults (A), and senior citizens (S). The research indicates that among all young people, 80% use TalkNow. Among adults, 40% use TalkNow. Among the senior citizens, just 20% use TalkNow. There are as many adults as young people and senior citizens combined.

What is the probability that a person is a young person, given that we know this person uses TalkNow?

(b) (1 point) A Bayesian Network is a Directed Acyclic Graph (DAG) to represent the conditional dependencies for a set of random variables. A colleague has sent you an email explaining that she has a Bayesian Network with 12 nodes but she has not provided further details about it. Given n random variables, how many possible Bayesian Networks could be constructed to represent their conditional dependencies?

(c) (1 point) Later, another colleague sends you a new Bayesian Network. In this network, there are only five nodes, for the random variables A, B, C, D, and E. All variables are Boolean. In the Bayesian Network, there is a directed edge from A to B, from A to C, from B to C, from C to D, and from C to E.

There are no other edges in this network. In particular, there is no edge from A to D or from D to A. Does this network structure imply that D is independent from A? Explain why or why not.

The colleague sends you all the conditional probability tables (CPTs) for this Bayesian Network. How would you use them to compute the probability of E given A? Provide the formula.

(d) (1 point) Are smaller Markov blankets better when performing inference using Gibbs sampling? Briefly explain in one or two sentences why or why not.

8. General machine learning.

   (a) (1 point) In the curve fitting problem on a set of data points, what is the disadvantage of Nearest Neighbors approach compared to Least Squares method?

   (b) (1 point) Answer True/False for the following statement. Explain your reasoning in 1-2 sentences.

   "When training a linear regression estimator, 10-fold cross-validation has smaller bias than 5-fold cross-validation".

9. Maximum Likelihood (ML).

   Suppose you are given two coins. Coin 1 has the probability θ of heads, while Coin 2 has the probability 3θ of heads. Suppose after flipping these two coins, you observe the following results: $Coin1(Head), Coin2(Head), Coin2(Tail), Coin2(Tail)$.

   (a) (1 point) Write down the likelihood function of the observation given θ.

   (b) (2 points) Compute the maximum likelihood estimate of parameter θ.

10. (5 points) Probabilistic model learning.

    Given training data $x_1, \cdots, x_n, \cdots, x_N$, single densities $i \in 1, \cdots, I$ where I is the total number of densities and unknown model parameter λ, estimate the the model density $p(i|\lambda)$ in mixture density $p(x_n|\lambda) = \sum_i p(i|\lambda) \cdot p(x_n|i, \theta_i)$ using Expectation Maximization, according to the maximum likelihood criterion.

Part III: Machine Learning/Computational Biology

11. (3 points) Maximum a Posteriori (MAP).

    Earlier, we considered Maximum Likelihood estimates (9). Suppose you are again considering the same two coins. Coin 1 has the probability θ of heads, while Coin 2 has the probability 3θ of heads. You are again studying the following observed results after flipping these two coins: $Coin1(Head), Coin2(Head), Coin2(Tail), Coin2(Tail)$.

    This time, however, you assume that θ can only take two possible values from $\{0.1, 0.3\}$. Additionally assume that we have the following prior for the parameter θ: $P[p = 0.1] = 0.3$ and $P[p = 0.3] = 0.7$. Compute the maximum a posteriori (MAP) estimate of θ over the above observation, using this prior.

12. (1 point) Cross-Validation.

    You are planning to use k-fold cross-validation for deciding on the values of the hyperparameters of your learning algorithm. However, for this, you first need to decide what k to use. What is the advantage of using a larger value, e.g. $k = 10$, rather than e.g. $k = 2$? Is there any advantage of using $k = 2$ rather than $k = 10$?

13. (3 points) Clustering.

You are given a very large dataset of search engine queries and want to find meaningful groups of semantically related queries. For this, you are planning to use hierarchical agglomerative clustering, which requires a measure of dissimilarity between clusters.

a) Which of the two choices Single Linkage or Complete Linkage clustering will typically result in clusters with a smaller diameter? Why?

b) The algorithm's output can be visualized using a dendrogram. Explain in one or two sentences if there are any disadvantages of dendrograms on this dataset.

c) Explain in one or two sentences if there would be any advantage of instead using the k-means algorithm on this dataset.

14. (4 points) SVMs.

You would like to classify connections on a social network as close friends ($y = 1$) or not so close friends ($y = -1$). The data is not linearly separable, so you decide to use a linear SVM with soft margins, where instead of enforcing a distance of at least 1 from the decision surface, there are additional slack variables ξ that allow for relaxing this hard constraint. You are thus attempting to learn a linear SVM model with soft margins by solving the dual Lagrangian. The training data and the corresponding Lagrange multipliers α_i that you have obtained using this approach are shown in the following table.

x_{i1}	x_{i2}	y_i	α_i
1.0	4.0	1	2.0 ✓
1.5	7.0	1	0.0
2.0	6.0	1	0.0
4.0	6.0	1	1.5 ✓
4.5	3.0	1	2.0 ✓
1.5	5.0	1	2.0 ✓
2.0	2.0	1	1.1 ✓
3.0	1.0	1	0.0
4.0	1.0	1	0.0
5.0	3.0	1	0.4 ✓
5.0	4.5	1	2.0 ✓

$y = w^T x + b$

$\arg\min_w \frac{1}{2}\|w\|^2$ ✓

$w^T x + b \geq 1 - \xi$

a) What are the support vectors for this training data?

b) Assume that the decision surface bias is $b = -1.8$. Compute the weight vector $\vec{w}$ using the data in the table given above.

c) Consider the points $(1, 4)$, $(4.5, 3)$, $(2, 2)$, and $(3, 1)$ from the table above. How are these points classified by the model? Compute the slack variables x_i for these points and for each of them determine whether the point is 1) correctly classified, outside the margin, 2) correctly classified, at the margin hyperplane, 3) correctly classified, inside the margin, or 4) misclassified.

For these computations, use b and $\vec{w}$ from above. If you were not able to determine the real $\vec{w}$, you may proceed assuming $(-0.5, 0.5)$ for $\vec{w}$ to get partial credit for your

solution.

d) Use the SVM model to compute a classification for a test data point $(4, 4)$.

15. (4 points) Perceptron and Kernel Methods.

a) Does the Perceptron algorithm have any advantages over the regular SVM algorithm?

b) Under what circumstances can a machine learning algorithm be kernelized (i.e., adapted to work with kernels)? Can you find a way to do this for the Perceptron algorithm? Explain the changes you need to make to it and why these changes are acceptable. Also explain whether this algorithm has any disadvantages.

Hint: Instead of storing the weight vector $\vec{w}$ explicitly, you may use a dual formulation, with $\vec{w} = \sum_{i=1}^{n} \alpha_i y_i x_i$ for n training examples (x_i, y_i), where $y_i \in \{-1, 1\}$, $\alpha_i \in \mathbb{R}^n$.

16. Graphical Models.

In an undirected graphical model with three random variables x, y and z, suppose any two nodes are connected to each other (that is, x, y and z form a complete graph). Assume that x, y and z can only take values from $\{-1, 1\}$. In this MRF model, the associated potential functions are defined by

$$\phi(x, y) = 2xy + x^3,$$
$$\phi(x, z) = xz + z,$$
$$\phi(y, z) = yz^3.$$

$P(x,y) = \frac{1}{Z} e^{-\phi(x,y)}$

$P(x|y) = \frac{P(x,y)}{P(x=1)}$

(a) (3 points) Compute the posterior marginal probability $P(y \mid x = 1)$.

(b) (2 points) In the following three directed graphical models, assume variable y has been observed:

$$x \longrightarrow y \longrightarrow z,$$
$$x \longrightarrow y \longleftarrow z,$$
$$x \longleftarrow y \longrightarrow z.$$

In the above three directed models, argue whether x and z are conditionally independent, given y. Also provide the brief proofs for your statements.

17. (3 points) Ensemble learning

Bootstrap aggregation, often abbreviated as bagging, involves having each model in the ensemble vote with equal weight. Recall that bagging trains each model in the ensemble using an independently randomly drawn subset of the training set. Given a standard training set D of size n, bagging generates m new training sets (bootstrap sample), each of size n, by sampling from D uniformly and with replacement. These m models are trained using the above m bootstrap samples respectively and combined by voting for classification. Write a rigorous derivation showing that bagging decreases the mean squared error rate of a regression task, compared to the single regressor trained on D.

$P(S_1) P(X_1|S_1) P(S_2|S_1) P(X_2|S_2)$

18. Hidden Markov model

Consider a given time sequence of vectors $x_1, ..., x_T$ as observations and the time sequence of states $s_1, ..., s_T$ as hidden states. We can model $p(x_1^T) = \sum_{s_1^T} \prod_{t=1}^{T} p(x_t, s_t | x_1^{t-1}, s_1^{t-1})$.

(a) (2 points) Write down the model for $p(x_1^T)$ using dependencies only on the previous state (first-order Markov).

(b) (2 points) Write a generalized Hidden Markov Model that employs the forward algorithm (i.e. the standard dynamic programming algorithm) for scoring.

(c) (1 point) The direct calculation of the sum in (1) is time-consuming, while (2) is an efficient one. What is the time complexity of the algorithm in direct calculation of (1) and forward algorithm of (2).

19. (2 points) EM

Explain and prove the convergence of EM algorithm.

20. (10 points) Computational Biology.

Answer True/False for the following statement.

(1) Most of the DNA sequences in the human genome code for proteins.

(2) There are 64 codons for translating RNAs into proteins.

(3) In the central dogma of molecular biology, DNAs are transcribed into RNAs, while RNAs are translated into proteins.

(4) In genetic code, each RNA nucleotide codes for each amino acid.

(5) For protein backbone structure, each residue only has one degree of freedom.

(6)All proteins have unique backbone conformations.

(7) The protein design problem is an NP-hard problem.

(8) Both DNAs and proteins have fixed 3D conformations.

(9) RNA strands can double back and form loops.

(10) Each drug can only interact with a unique target protein.

$P(A, B, C, D, E, F, G, H, I) = P(A) P(B) P(C|A,B) P(D|C) P(E|B) P(I|E)$

$P(G|D,F) P(H|F) P(I|H)$

$2+4+2+2+2+4+2+2$