Final Examination

CS540-2: Introduction to Artificial Intelligence

December 22, 2009

For Section 2, Prof. Zhu's class

20 questions, 5 points each

LAST NAME: SOLUTION

FIRST NAME: ______________________________

EMAIL : ______________________________

1. Consider a Support Vector Machine (SVM) with decision boundary w'x+b=0 for a three-dimensional feature space, where ' standards for vector transpose. The weight vector is w=(1 2 3)', and b=4.
   a. Which of the following points will be classified incorrectly by this SVM?
      i. x_1=(0 0 0)', y_1= -1 (minus one)
      ii. x_2=(-1 -1 -1)', y_2= 1 (plus one)
      iii. x_3=(-3.99 -3 2)', y_3= 1 (plus one)
   b. What would have been the slack variables ξ_i for the three points above, if they were in the training set? Assume w and b do not change. Hint: the constraints are $y_i(w' xi + b) \geq 1 - \xi_i$

answer: compare sign(w'xi+b) and yi. The first two are classified wrong. To satisfy the constraints, the slacks are 5, 3, and 0.99.

2. Consider a kernel $k(x, y) = e^{x+y} + \sqrt{xy} + 3$, where both x and y are positive real numbers. What is the feature vector $\phi(x)$ induced by this kernel?

Answer: $\theta(x) = (e^x, \sqrt{x}, \sqrt{3})$

3. Consider all possible kernels K and all possible values of input feature vector x. What is $\min_K \min_x K(x, x)$?

Answer: 0, because K(x,x)=inner product of the same vector, thus cannot be smaller than zero.

4. Fill in the missing values in the following joint probability table so that A and B are independent:

	A=T	A=F
B=T	3/12	6/12
B=F		

Answer: Computer the marginal p(B=F)=1-(3/12+6/12). Since A and B are independent, this probability mass has to be split with the same ratio as the first row. Thus 1/12 and 2/12.

5. Order each triple from large to small (allow equality), or explain why they can't be ordered.
    a. Triple 1: P(A), P(A, B), P(A, B, C)
    b. Triple 2: P(A), P(A|B), P(A|B,C)
    c. Triple 3: P(A|D), P(A,B|D), P(A,B,C|D)

Answer: triple 1 and 3 are in correct order. There is no definite order in triple 2.

6. Consider two Boolean random variables A and B. If P(A=F,B=T)/P(A=T,B=T)=2, what is P(A=T|B=T)?

Answer: 1/3

7. Consider a Naïve Bayes network B←A→C where the variables are Boolean. Complete the following CPTs so that P(A=T,B=T,C=T)=1/18, and P(A=F,B=F,C=F)=1/24:

    P(A=T)=?

    P(B=T|A=T)=?

    P(B=T|A=F)=1/2

    P(C=T|A=T)=1/4

    P(C=T|A=F)=3/4

Answer: solving a system of 2 equations, P(A=T)=2/3, P(B=T|A=T)=1/3.

8. Language X has 10,000 different words in its vocabulary. In theory, how many entries need to be stored if you want to create a 5-gram language model?

Answer: the history is 4-word long. There are $10{,}000^4$ histories. Each history needs 9,999 entries for its conditional probability. Therefore, $9{,}999*10{,}000^4$.

9. Grandma cannot hear very well. You invent a device which blows a tiny puff of air to her face as soon as the device detects that someone is starting to say a word. Which one of the following words do you expect Grandma may hear better with your device? “nod, pod, rod”. Briefly explain why.

Answer: this is an example of tactile McGurk’s effect. “pod” is the most likely one as ‘p’ is a plosive which is associated with air puff in face to face communication.

10. Consider a D-dimensional feature space where each feature takes value 0 or 1. What is the largest Euclidean distance between two points in this feature space? What is the smallest Euclidean distance between two non-overlapping points in this feature space?

Answer: largest: all-0 to all-1, sqrt(D). smallest: 1.

11. Why is it a bad idea to tune k for k-nearest-neighbor classifier on the training set? Your answer must be more specific than “overfitting.”

Answer: tuning on training set will result in k=1 with zero training error.

12. Consider three sequences: ♣♦♥♠, ♣♥♠♦, and ♠♦♥♣. An “adjacent swap” takes two elements next to each other in a sequence and switches their order. For example, an adjacent swap applied to the first two elements in the first sequence would result in ♦♣♥♠. Define the distance between two sequences as the minimum number of adjacent swaps needed to change one sequence into the other. Use this distance and Hierarchical Agglomerative Clustering to cluster the three sequences into two clusters. (Hint: one cluster has two sequences; the other cluster has one sequence.)

Answer: we only need to find out the two sequences that’s closest. (♣♦♥♠, ♣♥♠♦), (♠♦♥♣).

13. Welcome to the Terrible-Three-Day-Tour™! We will visit New York on Day 1. The rules for Day 2 and Day 3 are:
    a. If we were at New York the day before, flip a fair coin to decide either stay in New York (head), or go to Baltimore (tail).
    b. If we were at Baltimore the day before, flip a fair coin to decide either stay in Baltimore (head), or go to Washington D.C. (tail).

    On average, before you start the tour, what is your chance to visit Washington D.C.? What is your chance to visit Baltimore?

Answer: this is a Markov chain with 3 states. DC is only visited in the sequence NBW, which has probability ¼. B is always visited except the sequence NNN, thus has probability ¾.

14. Which of the following First-Order Logic sentences are correct translations of "No two adjacent countries have the same color?"

a. $\forall x, y\ \neg Country(x) \vee \neg Country(y) \vee \neg Adjacent(x, y) \vee \neg(Color(x) = Color(y))$

Yes

b. $\forall x, y\ (Country(x) \wedge Country(y) \wedge Adjacent(x, y)) \Rightarrow \neg(Color(x) = Color(y))$

Yes

c. $\forall x, y\ Country(x) \wedge Country(y) \wedge Adjacent(x, y) \wedge \neg(Color(x) = Color(y))$

No

d. $\forall x, y\ (Country(x) \wedge Country(y) \wedge Adjacent(x, y)) \Rightarrow Color(x \neq y)$

No

15. Let C1 be the clause **¬Republican(Mother(*x*)) ∨ Republican(*x*)** and let C2 be the clause **¬Republican(*y*) ∨ likes(*y*, Sarah) ∨ ¬Resident(*y*, Alaska)**. What is the result of applying the *resolution rule of inference* to C1 and C2?

¬Republican(Mother(x)) ∨ likes(x, Sarah) ∨ ¬Resident(x, Alaska)

16. Are the following two expressions *unifiable*? If so, what is the most general unifier? If not, why not?

P(*x*, *g*(*y*, A, *h*(*y*, B))) and *P*(*h*(A,B), *g*(A, *y*, *x*))

MGU = {x/h(A, B), y/A}

17. Can a Perceptron learn the SAME function of three binary inputs, defined to be 1 if all inputs are the same value and 0 otherwise? Either argue/show that this is impossible or construct a Perceptron that correctly represents this function.

No. SAME is the complement of XOR. It is not linearly separable and therefore cannot be represented by a Perceptron. Proof is by a figure showing the 3D space.

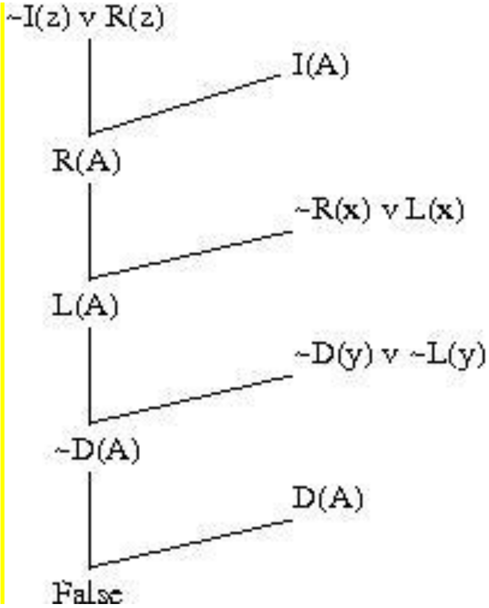
18. Given the following FOL sentences in CNF:

1. $\neg R(x) \vee L(x)$
2. $\neg D(y) \vee \neg L(y)$
3. D(A)
4. I(A)

Use the resolution algorithm to prove the query: $\exists z\, (I(z) \wedge \neg R(z))$ by constructing a proof tree.

First, negate the query and convert it to CNF: $\neg I(z) \vee R(z)$

Then construct the refutation proof tree:



19. True or False:

    a. A Perceptron can learn to correctly classify the following data, where each consists of three binary input values and a binary classification value: (111,1), (110,1), (011,1), (010,0), (000,0).

    b. The Perceptron Learning Rule is a sound and complete method for a Perceptron to learn to correctly classify any two-class problem.

    c. Training neural networks has the potential problem of over-fitting the training data.

```
a. True.  Output is 1 if at least 2 of the 3 inputs are 1.
   Therefore a Perceptron with all three weights equal to 0.5 and
   a threshold value of 0.8 will work.
b. False.  It can only learn linearly-separable functions.
c. True
```

20. Back-Propagation Learning in Neural Networks

    a. What is the search space and what is the search method used by the back-propagation algorithm for training neural networks?

```
The search space is the n-dimensional weight space defined by the
n weights (including biases) in a given network.  The search
method is gradient descent.
```

    b. Back-propagation minimizes what quantity?

```
Minimizes the sum-squared error between the output of the network
and the desired output.
```

    c. Does the back-propagation algorithm, when run until a minimum is achieved, always find the same solution no matter what the initial set of weights are? Briefly explain why or why not.

```
No.  It will iterate until a local minimum in the squared error
is reached.
```