Final Examination

CS540-2: Introduction to Artificial Intelligence

May 12, 2016

LAST NAME: SOLUTIONS

FIRST NAME: _____

Problem	Score	Max Score
1		18
2		6
3		9
4		8
5		12
6		11
7		14
8		5
9		17
Total		100

Question 1. [18] Neural Networks

(a) [2] True or False: The Perceptron Learning Rule is a sound and complete method for a Perceptron to learn to correctly classify any 2-class classification problem.

False. It can only learn linearly-separable functions.

(b) [2] True or False: A Perceptron can learn the Majority function, i.e., where each input unit is a binary value (0 or 1) and it outputs 1 if there are more 1s in the input than 0s. Assume there are *n* input units where *n* is odd.

True. Set all weights from the *n* input units to be 1 and set the bias to be $-n/2$.

(c) [2] True or False: Training neural networks has the potential problem of overfitting the training data.

True

(d) [2] True or False: The back-propagation algorithm, when run until a minimum is achieved, always finds the same solution no matter what the initial set of weights are.

False. It will iterate until a local minimum in the squared error is reached.

(e) [3] The back-propagation algorithm minimizes what quantity?

It minimizes the sum-squared error between the output of the network and the desired output, summed over all the training examples (or a subset of the examples if stochastic gradient descent is used).

(f) [3] The most popular activation function used in deep neural networks, because it typically learns much faster, is called ___Rectified Linear Unit (ReLU)___

(g) [4] Convolutional Neural Networks are multi-layer neural nets containing layers of different types (not counting the activation function). Name *two* (2) types of layers used.

Convolutional layer and Pooling layer

Question 2. [6] Support Vector Machines

Consider a Support Vector Machine with decision boundary $\mathbf{w}^T\mathbf{x} + b = 0$ for a three-dimensional feature space, where ^{T} standards for vector transpose, $\mathbf{w} = (1\ 2\ 3)^T$ and $b = 4$.

(a) [3] Will the example defined by $\mathbf{x} = (-2\ \ -3\ \ 4)^T$ and desired output $y = +1$ be classified *correctly* or *incorrectly*?

Compute sign($\mathbf{w}^T\mathbf{x} + b$) and compare with y. Here, $\mathbf{w}^T\mathbf{x} + b$ = (1)(-2) + (2)(-3) + (3)(4) + 4 = 8 > 0 and y = +1, so it is classified correctly.

(b) [3] Compute the **margin** associated with this decision boundary.

The margin $M = \frac{2}{\|w\|} = \frac{2}{\sqrt{w^T w}} = \frac{2}{\sqrt{(1)(1)+(2)(2)+(3)(3)}} = \frac{2}{\sqrt{14}} = 0.53$

Question 3. [9] Probabilities

(a) [5] Which (0 or more) of the following are equal to $P(A, B, C)$ given Boolean random variables A, B and C, and *no independence or conditional independence assumptions between any of them*?

(i) $P(A \mid B,C)P(B \mid C)P(C)$
(ii) $P(C \mid A,B)P(A)P(B)$
(iii) $P(A,B \mid C)P(C)$
(iv) $P(A \mid B,C)P(B \mid A,C)P(C \mid A,B)$
(v) $P(A \mid B)P(B \mid C)P(C)$
(vi) $P(C \mid A,B)P(A,B)$

(i), (iii) and (vi)

(b) [4] Which (0 or more) of the following expressions are guaranteed to equal 1, given (*not* necessarily Boolean) random variables A and B, and *no independence or conditional independence assumptions between them*?

(i) $\sum_a P(A = a \mid B)$
(ii) $\sum_b P(A \mid B = b)$
(iii) $\sum_a \sum_b P(A = a, B = b)$
(iv) $\sum_a \sum_b P(A = a \mid B = b)$

(i) and (iii)

Question 4. [8] Probabilistic Reasoning

Incidence of the disease Tuberculosis in the U.S. is about 5 cases per 10,000 people (i.e., $P(TB) = 0.0005$). Let Boolean random variable *TB* mean a patient "has Tuberculosis" and let Boolean random variable *TP* stand for "tested positive." TB tests are known to be very accurate in the sense that the probability of testing positive when you have the disease is 0.99, and the probability of testing negative when you do *not* have the disease is 0.97.

(a) [4] Compute $P(TP)$, the prior probability of testing positive.

Given: $P(TB) = 5/10000 = 0.0005$,
$P(TP \mid TB) = 0.99$, and
$P(\neg TP \mid \neg TB) = 0.97$

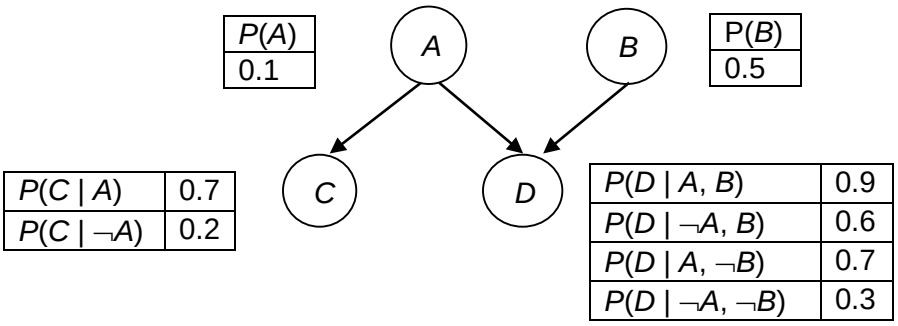
$P(TP) = P(TP \mid TB)P(TB) + P(TP \mid \neg TB)P(\neg TB)$
$= (0.99)(0.0005) + (1 - 0.97)(1 - 0.0005)$
$= 0.03048$

(b) [4] Compute $P(TB \mid TP)$, the posterior probability that you have TB when the test is positive.

$P(TB \mid TP) = (P(TP \mid TB)\ P(TB)) / P(TP)$ by Bayes's rule
$= (0.99)(0.0005)/(0.03048)$
$= 0.016$

Question 5. [12] Bayesian Networks

Consider the following Bayesian Network containing four Boolean random variables.



(a) [4] Compute $P(\neg A, B, \neg C, D)$

$$P(\neg A, B, \neg C, D) = P(\neg C \mid \neg A)\, P(D \mid \neg A, B)\, P(\neg A)\, P(B) =$$
$$= (1 - 0.2)(0.6)(1 - 0.1)(0.5)$$
$$= 0.216$$

(b) [4] Compute $P(\neg C \mid A, \neg B)$

$$P(\neg C \mid A, \neg B) = P(\neg C \mid A) = 1 - 0.7 = 0.3$$

(c) [4] Compute $P(A \mid B, C, D)$

$$P(A \mid B, C, D) = P(A, B, C, D) \,/\, P(B, C, D)$$
$$P(A, B, C, D) = (0.7)(0.9)(0.1)(0.5) = 0.0315$$
$$P(B, C, D) = P(A, B, C, D) + P(\neg A, B, C, D)$$
$$= 0.0315 + (0.2)(0.6)(0.9)(0.5) = 0.0855$$
$$\text{So, } P(A \mid B, C, D) = 0.0315 \,/\, 0.0855 = 0.368$$

Question 6. [11] Naïve Bayes Classifier

Consider the following dataset where there are three Boolean attributes, *W*, *X* and *Y*, and one Boolean class variable, *C*:

W	*X*	*Y*	*C*
T	T	T	T
T	F	T	F
T	F	F	F
F	T	T	F
F	F	F	T

(a) [3] What is *P*(*W*=T | *C*=F)?

There are 3 examples where *C*=F and *W*=T in two of them, so *P*(*W*=T | *C*=F) = 2/3 = 0.67

(b) [5] How would a Naïve Bayes Classifier classify the test example (*W*=F, *X*=T, *Y*=F)? No credit without showing how you got your answer. Do *not* compute logs, do *not* do smoothing, and do *not* compute the normalization term.

P(C | ¬W, X, ¬Y) = P(¬W, X, ¬Y | C)P(C) / P(¬W, X, ¬Y) by Bayes's rule
= P(C | ¬W, X, ¬Y) = P(¬W, X, ¬Y | C)P(C) after dropping the normalization term
= P(¬W | C)P(P(X | C) P(¬Y | C) P(C) by conditional independence assumptions
= (1/2)(1/2)(1/2)(2/5) = 0.05

P(¬C | ¬W, X, ¬Y) = P(¬W, X, ¬Y | ¬C)P(¬C) / P(¬W, X, ¬Y) by Bayes's rule
= P(¬C | ¬W, X, ¬Y) = P(¬W, X, ¬Y | ¬C)P(¬C) after dropping the normalization term
= P(¬W | ¬C)P(P(X | ¬C) P(¬Y | ¬C) P(¬C) by conditional independence assumptions
= (1/3)(1/3)(1/3)(3/5) = 0.022

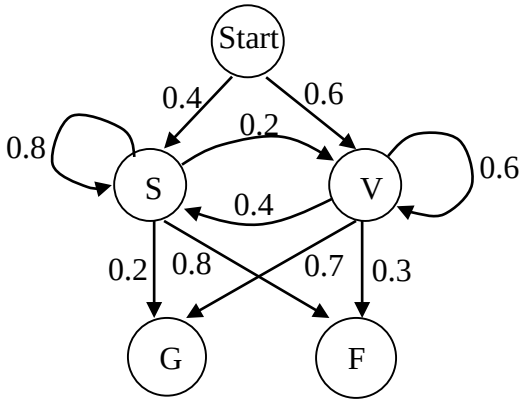
So, classify this test example as *C*=T because 0.05 > 0.022

(c) [3] Use "Add 1 Smoothing" to compute P(*W*=T | *C*=F)

"Add 1 Smoothing" counts every occurrence as having happened one more time than it did in the training data. Here there are two examples where *C*=F and *W*=T, and there is one example where *C*=F and *W*=F.
So, *P*(*W*=T | *C*=F) = (2+1) / [(2+1) + (1+1)] = 3/5 = 0.6

Question 7. [14] Hidden Markov Models

Consider the following HMM that models a student's activities every 5 minutes, where in each 5-minute period the student is in one of two possible hidden states: Studying (S) or playing Video games (V). While doing each activity the student will exhibit an observable facial expression of either Grinning (G) or Frowning (F).



(a) [4] Compute $P(q_1 = S, q_2 = S, q_3 = V, q_4 = V)$

$P(q_1 = S, q_2 = S, q_3 = V, q_4 = V)$
$= (0.4)(0.8)(0.2)(0.6)$
$= 0.0384$

(b) [5] Compute $P(o_1 = G)$

$P(o_1 = G) = P(q_1 = S, o_1 = G) + P(q_1 = V, o_1 = G)$
$P(q_1 = S, o_1 = G) = P(q_1 = S)\ P(o_1 = G \mid q_1 = S)$
$= (0.4)(0.2) = 0.08$
$P(q_1 = V, o_1 = G) = P(q_1 = V)\ P(o_1 = G \mid q_1 = V)$
$= (0.6)(0.7) = 0.42$
So, $P(o_1 = G) = 0.08 + 0.42 = 0.5$

(c) [5] Compute $P(q_1 = V \mid o_1 = F)$

$P(q_1 = V \mid o_1 = F) = P(o_1 = F \mid q_1 = V)P(q_1 = V) / P(o_1 = F)$
by Bayes's rule
$P(o_1 = F) = P(o_1 = F \mid q_1 = S)P(q_1 = S) + P(o_1 = F \mid q_1 = V)P(q_1 = V)$
$= (0.3)(0.6)/ [(0.8)(0.4) + (0.3)(0.6)]$
$= 0.36$

Question 8. [5] AdaBoost

(a) [2] True or False: If the number of weak classifiers is sufficiently large and each weak classifier is more accurate than chance, the AdaBoost algorithm's accuracy on the *training set* can always achieve 100% correct classification for any 2-class classification problem.

`True`

(b) [3] The AdaBoost algorithm creates an ensemble of weak classifiers by doing which *one* (1) of the following before determining the next weak classifier:

(i) Choose a new random subset of the training examples to use.
(ii) Decrease the weights of the training examples that were misclassified by the previous weak classifier.
(iii) Increase the weights of the training examples that were misclassified by the previous weak classifier.
(iv) Map the training examples into a higher dimensional space.
(v) None of the above.

`(iii)`

Question 9. [17] Propositional Logic

(a) [3] If |- denotes a *complete* inference algorithm and α |= β then α ∧ ¬β |- _____

False

(b) [3] True or False: $A \Leftrightarrow B \models A \wedge B$

False because when A=F and B=F, $A \Leftrightarrow B$ is true but $A \wedge B$ is false.

(c) [3] True or False: $(A \Leftrightarrow B) \wedge (\neg A \vee B)$ is *unsatisfiable*.

False because when A=T and B=T the sentence is true.

(d) [3] True or False: The sentence $(P \wedge Q) \Leftrightarrow R$ can be represented by an equivalent set of *Horn clauses*.

True because this is equivalent to
$(\neg P \vee \neg Q \vee R) \wedge (\neg R \vee P) \wedge (\neg R \vee Q)$

(e) [5] Prove that the query sentence *S* is true using the *Resolution Refutation algorithm* and the KB consists of the four sentences: ¬P ∨ ¬Q, (P ∨ Q) ∧ (P ∨ R), (S ∨ Q) ∧ (S ∨ ¬R), and ¬R. Show your answer as a proof tree.

1. Resolve ¬S and S ∨ Q to obtain Q
2. Resolve Q and ¬P ∨ ¬Q to obtain ¬P
3. Resolve ¬P and P ∨ R to obtain R
4. Resolve R and ¬R to obtain False