Final Examination

CS540-2: Introduction to Artificial Intelligence

May 9, 2018

LAST NAME: SOLUTIONS

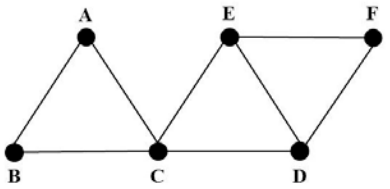
FIRST NAME: _____

Directions

1. This exam contains 33 questions worth a total of 100 points
2. Fill in your **name** and **student ID** number carefully on the answer sheet
3. Fill in each oval that you choose *completely*; do *not* use a check mark, an "X" or put a box or circle around the oval
4. Fill in the ovals with **pencil**, not pen
5. If you change an answer, be sure to completely erase the old filled-in oval
6. Fill in *only one oval* for each question
7. When you answer a question, be sure to check and make sure that the question number on the answer sheet matches the question number that you are answering on the exam
8. For True / False questions, fill in A for True and B for False
9. There is no penalty for guessing

Constraint Satisfaction

Consider the following graph representing 6 countries on a map that needs to be colored using three different colors, 1, 2 and 3, such that no adjacent countries have the same color. Adjacencies are represented by edges in the graph. We can represent this problem as a Constraint Satisfaction Problem where the variables, A – F, are the countries, and the values are the colors.



1. [4] The table below shows the current domains of all countries after country B has been assigned value 1 and country F has been assigned value 2. What are the reduced domains of the *two* countries C and D after applying the **Forward Checking** algorithm to this state? (The domains of countries A and E may change too but we're not interested in them.)

A	B	C	D	E	F
{1, 2, 3}	{1}	{1, 2, 3}	{1, 2, 3}	{1, 2, 3}	{2}

A. C = {1, 2, 3} and D = {1, 2, 3}
B. C = {2, 3} and D = {1, 2, 3}
C. C = {2, 3} and D = {1, 3}
D. C = {2} and D = {1, 3}
E. C = {2} and D = {3}

Answer: C. C = {2, 3}, D = {1, 3}

2. [4] The table below shows the current domains of all countries after country A has been assigned value 2 and country D has been assigned value 3. What are the reduced domains of the *two* countries C and F after applying the **Arc Consistency** algorithm (AC-3) to this state? (The domains of countries B and E may change too but we're not interested in them.)

A	B	C	D	E	F
{2}	{1, 2, 3}	{1, 2, 3}	{3}	{1, 2, 3}	{1, 2, 3}

A. C = {1} and F = {1}
B. C = {1} and F = {2}
C. C = {1, 3} and F = {1, 2)
D. C = {1, 3} and F = {1}
E. C = {1, 2} and F = {1, 2}

Answer: A. C = {1} and F = {1}

Neural Networks

3. [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.

   A. True
   B. False

   Answer: False. It can only learn linearly-separable functions.

4. [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.

   A. True
   B. False

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

5. [2] True or False: Training neural networks has the potential problem of overfitting the training data.

   A. True
   B. False

   Answer: True.

6. [2] True or False: The back-propagation algorithm, when run until a minimum is achieved, always finds the same solution (i.e., weights) no matter what the initial set of weights are.

   A. True
   B. False

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

7. [4] Consider a Perceptron that has two input units and one output unit, which uses an LTU activation function, plus a bias input of +1 and a bias weight $w_3 = 1$. If both inputs associated with an example are 0 and both weights, w_1 and w_2, connecting the input units to the output unit have value 1, and the desired (teacher) output value is 0, how will the weights change after applying the Perceptron Learning rule with learning rate parameter $\alpha = 1$?

   A. w_1, w_2 and w_3 will all decrease
   B. w_1 and w_2 will increase, and w_3 will decrease
   C. w_1 and w_2 will increase, and w_3 will not change
   D. w_1 and w_2 will decrease, and w_3 will not change
   E. w_1 and w_2 will not change and w_3 decreases

   Answer: E. w_1 and w_2 will not change and w_3 decreases

Convolutional Neural Networks

Consider a Convolutional Neural Network (CNN) that has an Input layer containing a 13 x 13 image that is connected to a Convolution layer using a 4 x 4 filter and a stride of 1 (i.e., the filter is shifted horizontally and vertically by 1 pixel, and only filters that are entirely inside the input array are connected to a unit in the Convolution layer). There is *no* activation function associated with the units in the Convolution layer. The Convolution layer is connected to a max Pooling layer using a 2 x 2 filter and a stride of 2. (Only filters that are entirely inside the array in the Convolution layer are connected to a unit in the Pooling layer.) The Output layer contains 4 units that each use an ReLU activation function and these units are fully-connected to the units in the Pooling layer.

8. [4] How many *units* are in the Convolution layer?

   A. 16
   B. 81
   C. 100
   D. 144
   E. 169

   Answer: C. because 10 x 10 = 100 units

9. [4] How many distinct *weights* must be learned for the connections to the Convolution layer?

   A. 16
   B. 169
   C. 2,704
   D. 16,900
   E. None of the above

   Answer: A because 4 x 4 = 16 weights because weights are shared

10. [4] How many *units* are in the Pooling layer?

    A. 4
    B. 25
    C. 81
    D. 100
    E. 169

    Answer: B because 5 x 5 = 25 units

11. [4] How many distinct *weights* must be learned for the connections to the Output layer?

    A. 4
    B. 5
    C. 16
    D. 100
    E. 104

    Answer: E because (25 x 4) + 4 = 104 (including the bias weights)

The next two questions do ***NOT*** refer to the specific CNN described above, but rather are about CNNs in general.

12. [2] True or False: CNNs can learn to recognize an object in an image no matter how the object is *translated* (i.e., shifted horizontally and/or vertically) even if the training set only includes that object in one position.

    A. True
    B. False

    Answer: True because of shared weights

13. [2] True or False: CNNs can learn to recognize an object in an image no matter how the object is *rotated* (in the image plane) even if the training set only includes that object in one orientation.

    A. True
    B. False

    Answer: False

Support Vector Machines

14. [2] True or False: Given a *linearly-separable* dataset for a 2-class classification problem, a Linear SVM is *better* to use than a Perceptron because the SVM will often be able to achieve a *better* classification accuracy on the *training set*.

    A. True
    B. False

    Answer: False because both with have 100% accuracy on the linearly-separable training set.

15. [2] True or False: Given a *linearly-separable* dataset for a 2-class classification problem, a Linear SVM is *better* to use than a Perceptron because the SVM will often be able to achieve *better* classification accuracy on the *testing set*.

    A. True
    B. False

    Answer: True because by maximizing the margin the decision boundary learned by the SVM will often be farther away from more of the training examples, leading to better performance on the testing examples that are close to the decision boundary.

16. [2] True or False: With a *non-linearly-separable* dataset that contains some extra "noise" data points, using an SVM with slack variables to create a soft margin classifier, and a *small* value for the penalty parameter, *C*, that controls how much to penalize misclassified points, will often *reduce overfitting* the training data.

    A. True
    B. False

    Answer: True because small C means the penalty for mis-classifying a few points will be small and therefore we are more likely to maximize the margin between most of the points while mis-classifying a few points including the noise points.

17. [2] True or False: An SVM using the quadratic kernel $K(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^T \cdot \mathbf{y})^2$, can *correctly classify* the dataset {((1,1), +), ((1,-1), –), ((-1,1), –), ((-1,-1), +)}.

    A. True
    B. False

    Answer: True because using this quadratic kernel will effectively map each 2D point, (p, q), into a 3D point, $(p^2, \sqrt{2}pq, q^2)$, meaning points (1,1) and (-1,-1) map to $(1, \sqrt{2}, 1)$, and points (-1,1) and (1,-1) map to $(1, -\sqrt{2}, 1)$. So, these four points are now linearly separable in this 3D space.

Probabilities

18. [3] Which *one* of the following is 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*?

    A. $P(A \mid B)\ P(B \mid C)\ P(C \mid A)$
    B. $P(C \mid A, B)\ P(A)\ P(B)$
    C. $P(A, B \mid C)\ P(C)$
    D. $P(A \mid B, C)\ P(B \mid A, C)\ P(C \mid A, B)$
    E. $P(A \mid B)\ P(B \mid C)\ P(C)$

    Answer: C

19. [3] Which *one* of the following expressions is guaranteed to equal 1, given (*not* necessarily Boolean) random variables *A* and *B*, and *no independence or conditional independence assumptions between them*? Sums are over all possible values of the associated random variable.

    A. $\sum_x P(A = x \mid B = b)$
    B. $\sum_y P(A = a \mid B = y)$
    C. $\sum_x P(A = x) + \sum_y P(B = y)$
    D. $\sum_x \sum_y P(A = x \mid B = y)$
    E. None of the above

    Answer: A

20. [4] Given two Boolean random variables, *A* and *B*, where *P*(*A*) = ½, *P*(*B*) = 1/3, and P(*A* | ¬*B*) = ¼, what is P(*A* | *B*)?

    A. 1/6
    B. ¼
    C. ¾
    D. 1
    E. Impossible to determine from the given information

    Answer: D because P(A | B) = (P(B | A) P(A)) / P(B) = 3/2 P(B | A). P(B | A) = 1 – P(~B | A), and P(~B | A) = (P(A | ~B) P(~B)) / P(A) = 1/3, so P(B | A) = 2/3 and therefore P(A | B) = (2/3)(3/2) = 1

21. [4] A 6-sided die is rolled 15 times and the results are: side 1 comes up 0 times; side 2: 1 time; side 3: 2 times; side 4: 3 times; side 5: 4 times; side 6: 5 times. Based on these results, what is the probability of side 3 coming up when using **Add-1 Smoothing**?

    A. 2/15
    B. 1/7
    C. 3/16
    D. 1/5
    E. None of the above

    Answer: B because (2+1)/(1+2+3+4+5+6) = 1/7

Probabilistic Reasoning

Say the incidence of a disease D is about 5 cases per 100 people (i.e., $P(D) = 0.05$). Let Boolean random variable *D* mean a patient "has disease D" and let Boolean random variable *TP* stand for "tests positive." Tests for disease D 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.

22. [4] Compute $P(TP)$, the prior probability of testing positive.

A. 0.0285
B. 0.0495
C. 0.051
D. 0.078
E. None of the above

Answer: D because $P(D) = 0.05$, $P(TP \mid D) = 0.99$, $P(\neg TP \mid \neg D) = 0.97$, and $P(TP) = P(TP \mid D)\, P(D) + P(TP \mid \neg D)\, P(\neg D) = (.99)(.05) + (1 - .97)(1 - .05) = 0.078$

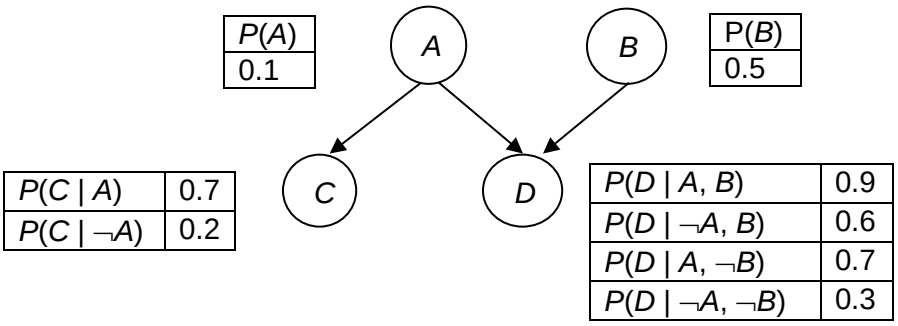
23. [4] Compute $P(D \mid TP)$, the posterior probability that you have disease D when the test is positive.

A. 0.0495
B. 0.078
C. 0.635
D. 0.97
E. None of the above

Answer: C because $P(D \mid TP) = (P(TP \mid D)\, P(D)) / P(TP) = ((.99)(.05)) / .078 = 0.635$

Bayesian Networks

Consider the following Bayesian Network containing four Boolean random variables.



24. [4] Compute $P(\neg A, B, \neg C, D)$

A. 0.216
B. 0.054
C. 0.024
D. 0.006
E. None of the above

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

25. [4] Compute $P(A \mid B, C, D)$

A. 0.0315
B. 0.0855
C. 0.368
D. 0.583
E. None of the above

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

The next two questions do ***NOT*** refer to the specific Bayesian Network shown above, but rather are about other Bayesian Networks.

26. [2] True or False: The Bayesian Network associated with the following computation of a joint probability: $P(A)\ P(B)\ P(C \mid A, B)\ P(D \mid C)\ P(E \mid B, C)$ has arcs from node *A* to *C*, from *B* to *C*, from *B* to *E*, from *C* to *D*, from *C* to *E*, and no other arcs.

    A. True
    B. False

    Answer: True

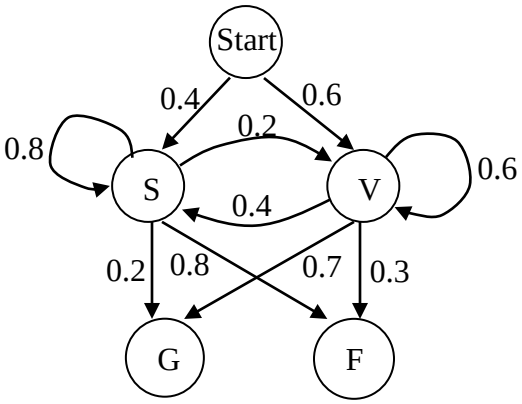
27. [2] True or False: The following product of factors corresponds to a valid Bayesian Network over the variables *A*, *B*, *C* and *D*: $P(A \mid B)\ P(B \mid C)\ P(C \mid D)\ P(D \mid A)$.

    A. True
    B. False

    Answer: False because this corresponds to a network that is not a DAG.

Hidden Markov Models

Consider the following HMM that models a student's activities every hour, where during each hour 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).



28. [4] If in the *second* hour the student is Studying, what's the probability that in the *fourth* hour the student is playing Video games? That is, compute $P(q_4 = V \mid q_2 = S)$.

A. 0.08
B. 0.12
C. 0.16
D. 0.2
E. 0.28

Answer: E because

$$P(q_4=V \mid q_2=S) = P(q_4=V, q_3=V \mid q_2=S) + P(q_4=V, q_3=S \mid q_2=S)$$
$$= P(q_4=V \mid q_3=V, q_2=S)\, P(q_3=V \mid q_2=S) + P(q_4=V \mid q_3=S, q_2=S)\, P(q_3=S \mid q_2=S)$$
$$= P(q_4=V \mid q_3=V)\, P(q_3=V \mid q_2=S) + P(q_4=V \mid q_3=S)\, P(q_3=S \mid q_2=S)$$
$$= (.6)(.2) + (.2)(.8) = 0.28$$

29. [4] What is the probability that in the *first* hour the student is Grinning? That is, compute $P(o_1 = G)$.

.

A. 0.08
B. 0.25
C. 0.42
D. 0.5
E. 1.0

Answer: D

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

AdaBoost

30. [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 where the data can be separated by a linear combination of the weak classifiers' decision boundaries.

A. True
B. False

Answer: True

31. [2] True or False: In AdaBoost, the error computed for a weak classifier is calculated by the ratio of the number of misclassified examples to the total number of examples in the training set.

    A. True
    B. False

    Answer: False

32. [2] True or False: For a 2-class classification problem where AdaBoost has selected five weak classifiers, the classification of a test example is determined by the majority class of the classes predicted by the five weak classifiers.

    A. True
    B. False

    Answer: False

33. [4] The AdaBoost algorithm creates an ensemble of weak classifiers by doing which *one* of the following before determining the next weak classifier:

    A. Chooses a new random subset of the training examples to use
    B. Decreases the weights of the training examples that were misclassified by the previous weak classifier
    C. Increases the weights of the training examples that were misclassified by the previous weak classifier
    D. Removes the training examples that were classified correctly by the previous weak classifier
    E. None of the above

    Answer: C