CIS 419/519 Introduction to Machine Learning
Fall 2014 Final Exam

Instructions:

- Please turn off your cell phone.
- This examination is closed-book and closed-notes. Calculators and any other external resources are prohibited.
- Please use a pencil to fill in the bubble sheet.
- All answers must be recorded on the bubble sheet. Fill in the bubble of the correct answer for each question. Be certain to fill in the bubble completely.
- To correct a mistake, be certain to erase the bubble completely. (Do NOT cross out incorrect bubbles as with the midterm, please erase them.)
- For true/false questions, fill in (A) for TRUE and fill in (B) for FALSE.
- For multiple choice questions, fill in the bubble(s) of the correct answer. Some multiple choice questions may have multiple correct answers; in this case, you can fill in multiple bubbles for a single question.
- Only the bubble sheet will be counted as your official answers; any answers you write on the exam will not be considered.
- If you find yourself spending too long on a problem, skip it and move on to the next one. Scrap paper is available upon request. Be certain to write you name on all materials you turn in.
- If you do not understand a problem or require clarification, please ask the instructor.
- There are 100 points total. Students in CIS 419 will be graded out of 94 points; points earned beyond 94 will not count as extra credit. Students in CIS 519 will be graded out of 100 points.
  - There are some questions on the exam that focus on topics from the 519 homeworks (e.g., VC dimension). Students from 419 are welcome to attempt these questions, but 419 students are not expected to know this material. For this reason, CIS 419 students are graded out of less total points on the exam.
- I wish each of you the best of luck. It has been a pleasure teaching you in this course, and I hope you enjoyed it as well. Have a great winter break!

I Preliminaries

There are multiple versions of this exam. Fill in the version number of the exam on your answer sheet:

- If you are enrolled in CIS 419, fill in the bubble for Version 1
- If you are enrolled in CIS 519, fill in the bubble for Version 3

Write your name on the bubble sheet, the first page of the exam, and on any scrap paper you use.

II General Knowledge (26 pts total)

Choose the correct answer for each question below:

1.) (1 pts) Check that your exam has all 12 pages. Choose any answer to this question to let me know you did it: of the following choices, which is your favorite machine learning algorithm?

(A) *Support vector machines* are awesome!
(B) Nothing beats *neural networks*
(C) Except for *deep belief nets*
(D) Nah, give me good ol' *logistic regression*
(E) *Spectral clustering*, FTW!

Truth or LIES For each question below, answer whether it is a True statement (A) or a bloody LIE (B)!

2.) (1 pts) If a reinforcement learning policy π is **greedy** with respect to **the optimal value function** $\mathbf{V}^*$, then it is an optimal policy for solving the problem.

3.) (1 pts) One disadvantage of Q-learning is that it can only be used when the learner has prior knowledge of how its actions affect its environment.

4.) (1 pts) PCA finds a **higher** dimensional embedding of a data set.

5.) (1 pts) Both PCA and linear regression can be thought of a minimizing a sum of squared errors.

6.) (1 pts) PCA and spectral clustering (such as Andrew Ng's algorithm) perform eigendecomposition on two different matrices. However, the size of these two matrices are the same.

7.) (1 pts) Since classification is a special case of regression, logistic regression is a special case of linear regression.

8.) (1 pts) The lowest level features learned by a deep belief network correspond closely to the features extracted by the first few layers of the human visual processing system.

9.) (1 pts) If a learning algorithm prompted a physician to input the diagnosis for a specific patient with a previously unknown disease, this would be an example of *active learning.*

VC Dimension

10.) (2 pts) True (A) or False (B): There is at least one set of 4 points in $\mathbb{R}^3$ that can be shattered by the hypothesis set of all 2D planes in $\mathbb{R}^3$.

11.) (4 pts) What is the VC-dimension of the k-Nearest Neighbor classifier when $k = 1$?
(A) 2 (B) 3 (C) 4 (D) 5 (E) ∞

Properties of Learning Algorithms

For each machine learning algorithm listed below, choose the correct properties of that algorithm.

Does the learning algorithm yield the **globally** optimal solution or a **locally** optimal solution?

12.) (1 pts) ID3 Decision Tree: (A) global optimum (B) local optimum

13.) (1 pts) Logistic regression: (A) global optimum (B) local optimum

14.) (1 pts) Neural network with 1 hidden layer: (A) global optimum (B) local optimum

15.) (1 pts) SVM: (A) global optimum (B) local optimum

16.) (1 pts) K-means: (A) global optimum (B) local optimum

17.) (1 pts) Q-learning: (A) global optimum (B) local optimum

What loss function does each algorithm use? Choose the best option from the following choices:
(A) zero-one loss (B) sum of squared error (C) log loss (D) hinge loss (E) exponential loss

18.) (1 pts) ID3 Decision Tree

19.) (1 pts) Logistic regression

20.) (1 pts) Restricted Boltzmann machine

21.) (1 pts) SVM

22.) (1 pts) K-means

III Frequently Missed Questions from the Midterm (10 pts total)

The following questions from the midterm were frequently answered incorrectly. Here's your second chance! (There are a few more scattered through the rest of the exam too.)

Logistic Regression

23.) (1 pts) For logistic regression, what is the effect on the bias as the number of training instances increases?
(A) Bias decreases (B) Bias remains unchanged (C) Bias increases

24.) (1 pts) For logistic regression, what is the effect on the variance as the number of training instances increases?
(A) Variance decreases (B) Variance remains unchanged (C) Variance increases

25.) (2 pts) Imagine you are writing code to compute the performance of logistic regression with a polynomial basis expansion on a data set, averaged over 20 trials of 10-fold cross-validation. Where would be the *best* place in your program to standardize the data?

(A) Immediately upon loading the data set, before splitting it into training/testing sets

(B) Immediately after cross-validation splits the data into training/testing sets

(C) Within the classifier's constructor

(D) Within the classifier's `fit()` and `predict()` methods

(E) None of the above; you should not standardize data when using logistic regression

Boosting

26.) (1 pts) For AdaBoost with decision stumps, what is the effect on the bias as the number of boosting iterations T increases?
(A) Bias decreases (B) Bias remains unchanged (C) Bias increases

27.) (1 pts) For AdaBoost with decision stumps, what is the effect on the variance as the number of boosting iterations T increases?
(A) Variance decreases (B) Variance remains unchanged (C) Variance increases

Experimental Procedure

28.) (1 pts) You read a paper that uses 100 trials of 10-fold cross-validation to choose the optimal model parameters. For each training set, the authors train multiple models (each with different parameters) on the training data, and choose the parameters with the lowest average test error. Is this scenario okay or problematic?
(A) Okay (B) Problematic

Probabilistic Models

29.) (3 pts) Under what conditions is the maximum a posteriori (MAP) hypothesis equivalent to the maximum likelihood hypothesis (MLE)? (Hint: if you're having trouble, write out equations for the MLE and MAP and figure out when they're equal.)

(A) When the class prior probabilities are equal

(B) When all instances are equally probable

(C) When the posterior probabilities are equal

(D) When the data attributes are conditionally independent given the class label

(E) None of the above

IV SVMs (14 pts total)

Consider the four training instances in $\mathbb{R}^2$ that are given in the figure to the right.

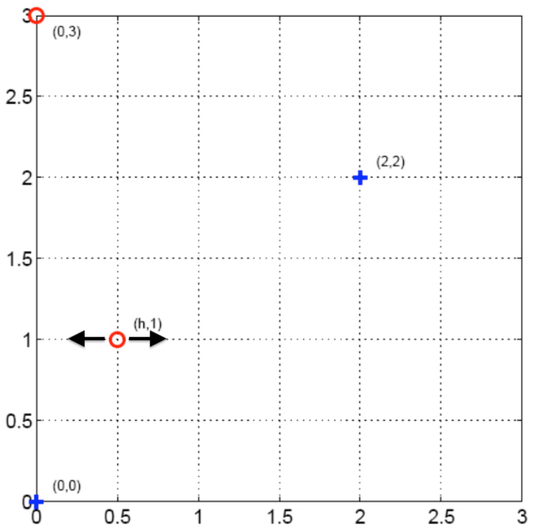
There are positive examples at:

$$\mathbf{x}_1 = [0, 0] \qquad \mathbf{x}_2 = [2, 2]$$

and negative examples at:

$$\mathbf{x}_3 = [h, 1] \qquad \mathbf{x}_4 = [0, 3].$$

Note that $\mathbf{x}_3$ can move horizontally; its horizontal position h is a **variable** such that $0 \leq h \leq 3$.



30.) (2 pts) How large can h be so that the training points are still linearly separable?
(A) $h < 0.5$ (B) $h < 1$ (C) $h < 1.5$ (D) $h \leq 3$ (E) None of the above

31.) (2 pts) When the points are linearly separable, does the orientation (i.e., angle) of the maximum margin decision boundary change as h varies?
(A) Yes (B) No

32.) (3 pts) Assume that we can only observe the 2nd dimension of the input vectors. Without the other component, the labeled training points reduce to (0,+), (2,+), (1,-), and (3,-). What is the **lowest-**order degree d of the polynomial kernel that would allow us to correctly classify these points?

(A) $d = 1$
(B) $d = 2$
(C) $d = 3$
(D) $d = 4$
(E) It is not possible to separate these points with the polynomial kernel

33.) (1 pts) True (A) or False (B): Using the quadratic kernel in an SVM with very large numbers of training instances is much more computationally efficient than precomputing the equivalent basis expansion of each training instance and using a linear kernel.

34.) (3 pts) Consider the two-dimensional data set with positive examples at $(1, 1)$ and $(-1, -1)$, and negative examples at $(1, -1)$ and $(-1, 1)$. The data set is not linearly separable in the original feature space, so we decide to apply the feature transformation $\phi(\mathbf{x}) = [1, x_1, x_2, x_1x_2]$, where x_j is the j^{th} feature value of $\mathbf{x}$. The prediction function is $h(\mathbf{x}) = \boldsymbol{\theta}^\intercal \phi(\mathbf{x})$ in this feature space.

Solve for the coefficients $\boldsymbol{\theta}$ of the maximum-margin separator in the expanded feature space. (You can do this by inspection, without significant computation.) What is the L$_2$ norm of $\boldsymbol{\theta}$?
(A) $\sqrt{2}$ (B) 2 (C) $\frac{2}{\sqrt{2}}$ (D) $2\sqrt{2}$ (E) None of the above

35.) (3 pts) In the standard form of the optimization problem for finding the maximum margin hyperplane, we require that $\mathbf{w} \cdot \mathbf{x} + b = -1$ for negative support vectors and $\mathbf{w} \cdot \mathbf{x} + b = +1$ for positive support vectors. What would happen to $\mathbf{w}$ if we used $-\epsilon$ and $+\epsilon$ instead of -1 and $+1$ as $\epsilon \to 0$? (Hint: It may help to consider a simple one dimensional dataset with two points, one positive and one negative.)

(A) $\|\mathbf{w}\|_2$ will shrink toward 0
(B) $\|\mathbf{w}\|_2$ will shrink, but only slightly
(C) $\|\mathbf{w}\|_2$ will not change
(D) $\|\mathbf{w}\|_2$ will grow, but only slightly
(E) $\|\mathbf{w}\|_2$ will grow toward ∞

V Naïve Bayes Learning (15 pts total)

(That's right, again!)

Snoopy is going on a tropical winter holiday! Given the following data, construct a naïve Bayes classifier with Laplace smoothing to identify whether Snoopy will wear sunglasses when he goes to the beach in Hawaii. You should consider the temperature as a discretized variable. Also, Snoopy only goes to the beach when the temperature is 70–99°, so using only three values for the temperature is sufficient.

SKY	HIGH TEMP	SUNGLASSES?
cloudy	90s	yes
sunny	90s	yes
sunny	90s	yes
cloudy	90s	yes
sunny	80s	yes
cloudy	80s	no
sunny	80s	yes
sunny	70s	no

36.) (3 pts) What is the MINIMUM total number of probabilities the naïve Bayes classifier will need to store? (That is, the total count of numbers in all conditional probability tables in the naïve Bayes model. If you can derive some of the parameters from others, than these parameters are not necessary to store explicitly.)
(A) 5 (B) 7 (C) 8 (D) 12 (E) None of the above

37.) (1 pts) What is $P(\text{Cloudy?} = yes \mid \text{Sunglasses?} = yes)$, assuming Laplace smoothing?
(A) 1/3 (B) 3/8 (C) 3/9 (D) 1/2 (E) None of the above

38.) (1 pts) What is $P(\text{Temp?} = 90s \mid \text{Sunglasses?} = yes)$, assuming Laplace smoothing?
(A) 1/2 (B) 5/9 (C) 5/8 (D) 2/3 (E) None of the above

39.) (1 pts) What is $P(\text{Temp?} = 90s \mid \text{Sunglasses?} = no)$, assuming Laplace smoothing?
(A) 0 (B) 1/6 (C) 1/5 (D) 1/4 (E) None of the above

40.) (1 pts) What is $P(\text{Temp?} = 80s \mid \text{Sunglasses?} = yes)$, assuming Laplace smoothing?
(A) 1/6 (B) 2/9 (C) 1/4 (D) 1/3 (E) None of the above

41.) (3 pts) Will Snoopy will wear sunglasses on a sunny day with a high of 78°?
(A) Yes (B) No

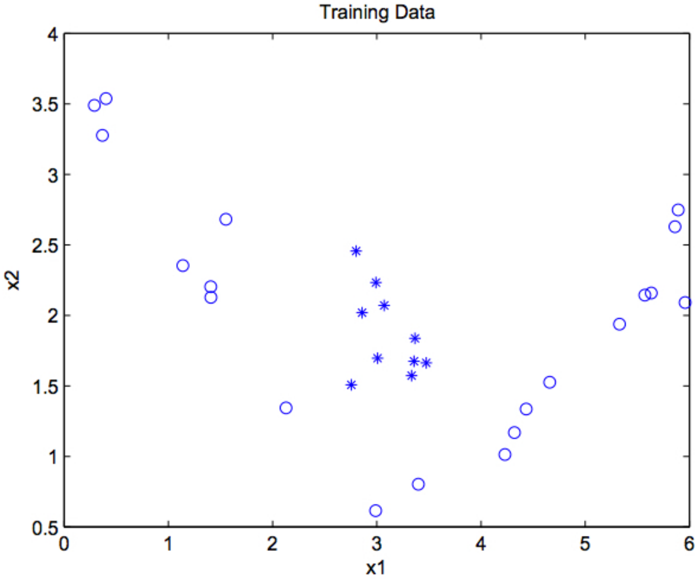
Before he leaves for vacation, however, Snoopy needs to decide how to decorate his dog house for the annual holiday decorating contest. To better inform his design choices, Snoopy has made a different naïve Bayes model to predict whether or not he will win the decorating contest. The parameters for Snoopy's model are given below; all variables are binary.

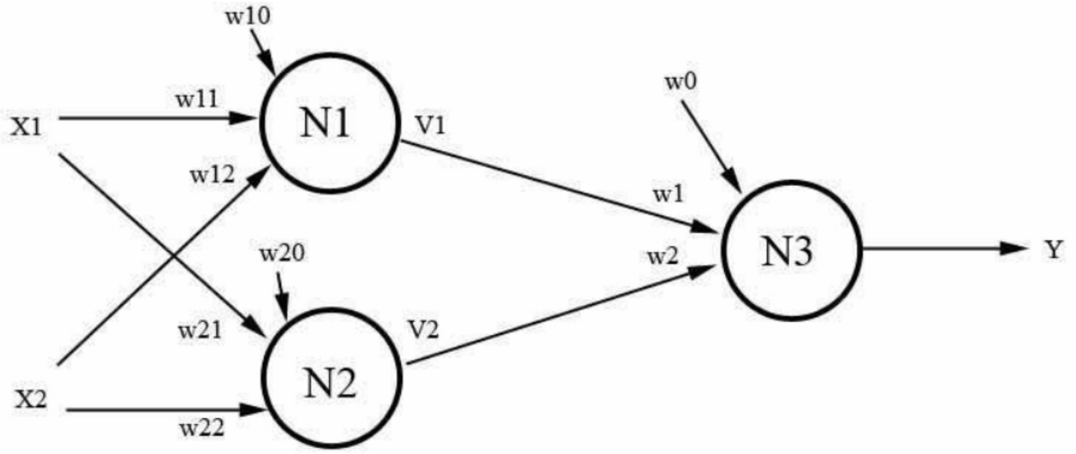
$$\begin{aligned} P(Win) &= 0.5 & P(\neg Win) &= 0.5 \\ P(UseColoredLights \mid Win) &= 0.75 & P(UseColoredLights \mid \neg Win) &= 0.5 \\ P(\neg UseColoredLights \mid Win) &= 0.25 & P(\neg UseColoredLights \mid \neg Win) &= 0.5 \\ P(PlayMusic \mid Win) &= 0.75 & P(PlayMusic \mid \neg Win) &= 0.25 \\ P(\neg PlayMusic \mid Win) &= 0.25 & P(\neg PlayMusic \mid \neg Win) &= 0.75 \end{aligned}$$

42.) (5 pts) What is the probability that Snoopy will win the decorating contest with his colored musical dog house this year?
(A) 9/32 (B) 9/16 (C) 9/13 (D) 9/11 (E) None of the above

VI Neural Networks (9 pts total)

Consider a neural network trained on the binary classification data below. (In this plot, positive instances are denoted by filled points (*), and negative instances are denoted by hollow circles (∘).) Assume that we train the neural network model given below, which uses sigmoid activation functions.

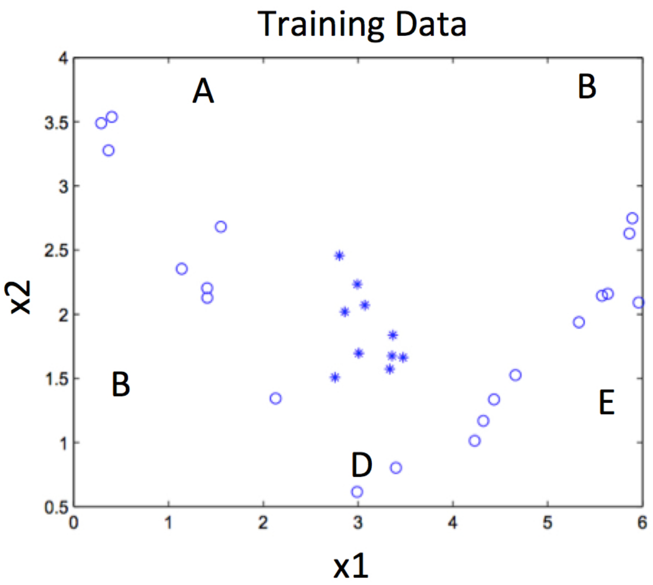




Training the neural network on the data will set the weights (the w's) of the network so that it is capable of correctly classifying this data set. In this problem, you will plot the decision boundaries learned by each of the nodes in the network.

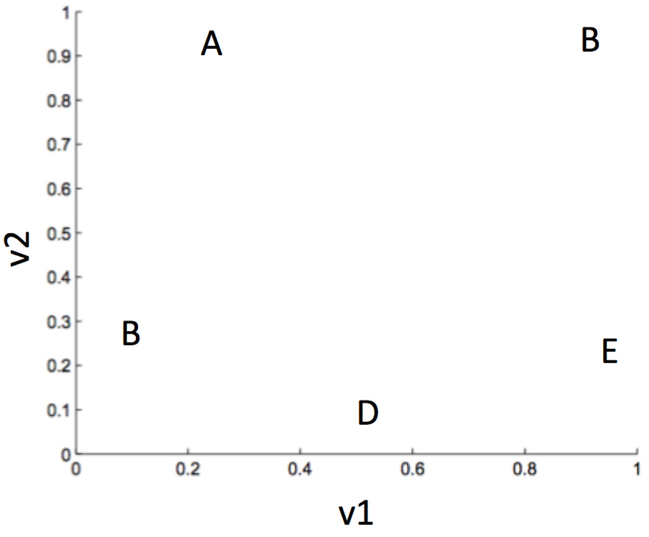
43.) (3 pts) The N1 node takes as input the features of the training data, and so we can visualize the decision boundary learned by this node in the original feature space. Plot N1's decision boundary in the figure below (i.e., for neuron N1, plot the line where w10 + w11 * X1 + w12 * X2 = 0).

When you draw the decision boundary in the figure below, you should notice that it passes through (or close by) two or more of the letters A–E in the figure. Record your answer by filling in the bubbles for the corresponding letters closest to the decision boundary. (E.g., if your decision boundary is a curve that connects D to E to B, then fill in those three bubbles.)



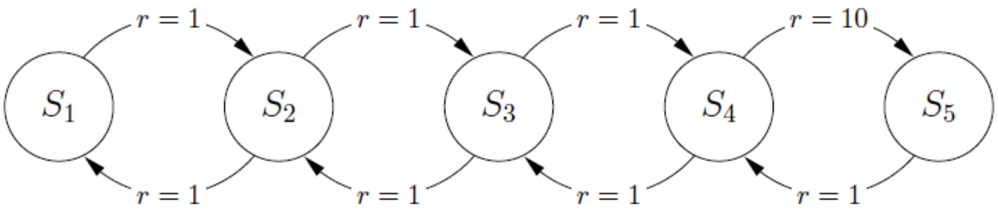
44.) (3 pts) Now, plot the decision boundary for N2 in the figure above, and fill in the letters closest to N2's decision boundary in the figure.

45.) (3 pts) Plotting the decision boundary for N3 is a bit different. N3 takes as input the output of N1 and N2, and so the graph below has axes V1 and V2 to correspond to the outputs of those nodes. Start by plotting (V1(x1,x2), V2(x1,x2)) for a few training points. Once you have a good idea of the input space to N3, draw a decision boundary such that the neural net will correctly classify the data. Once again, specify your answer by filling in the letters closest to that decision boundary in the figure.



VII Reinforcement Learning (15 pts total)

Consider the following Markov Decision Process:



We have states S_1, S_2, S_3, S_4, and S_5. We have actions Left and Right, and the chosen action happens with probability 1. In S_1 the only option is to go back to S_2, and similarly in S_5 we can only go back to S_4. The reward for taking any action is $r = 1$, except for taking action Right from state S_4, which has a reward $r = 10$. For all parts of this problem, assume that $\gamma = 0.8$.

46.) (3 pts) What is the optimal policy for this MDP?

(A) $\pi^* = \{S_1 : \text{Right},\ S_2 : \text{Left},\ S_3 : \text{Right},\ S_4 : \text{Right},\ S_5 : \text{Left}\}$

(B) $\pi^* = \{S_1 : \text{Left},\ S_2 : \text{Left},\ S_3 : \text{Left},\ S_4 : \text{Left},\ S_5 : \text{Right}\}$

(C) $\pi^* = \{S_1 : \text{Right},\ S_2 : \text{Right},\ S_3 : \text{Right},\ S_4 : \text{Right},\ S_5 : \text{Right}\}$

(D) $\pi^* = \{S_1 : \text{Right},\ S_2 : \text{Right},\ S_3 : \text{Right},\ S_4 : \text{Right},\ S_5 : \text{Left}\}$

(E) None of the above

47.) (4 pts) What is $V^*(S_5)$?
(A) 0 (B) 1 (C) 10 (D) 25 (E) ∞

Consider executing Q-learning on this MDP. Assume that the Q values for all (state, action) pairs are initialized to 0, that $\alpha = 0.5$, and that Q-learning uses a greedy exploration policy, meaning that it always chooses the action with maximum Q value. The algorithm breaks ties by choosing Left.

Write down the first 10 (state, action) pairs if our robot learns using Q-learning and starts in state S3 (e.g., "(S3, Left), (S2,Right), (S3,Right), …").

48.) (0.5 pts) What is the 1st state in your list?
(A) S_1 (B) S_2 (C) S_3 (D) S_4 (E) S_5

49.) (0.5 pts) What is the 2nd state in your list?
(A) S_1 (B) S_2 (C) S_3 (D) S_4 (E) S_5

50.) (0.5 pts) What is the 3rd state in your list?
(A) S_1 (B) S_2 (C) S_3 (D) S_4 (E) S_5

51.) (0.5 pts) What is the 4th state in your list?
(A) S_1 (B) S_2 (C) S_3 (D) S_4 (E) S_5

52.) (0.5 pts) What is the 5th state in your list?
(A) S_1 (B) S_2 (C) S_3 (D) S_4 (E) S_5

53.) (0.5 pts) What is the 6th state in your list?
(A) S_1 (B) S_2 (C) S_3 (D) S_4 (E) S_5

54.) (0.5 pts) What is the 7th state in your list?
(A) S_1 (B) S_2 (C) S_3 (D) S_4 (E) S_5

55.) (0.5 pts) What is the 8th state in your list?
(A) S_1 (B) S_2 (C) S_3 (D) S_4 (E) S_5

56.) (0.5 pts) What is the 9th state in your list?
(A) S_1 (B) S_2 (C) S_3 (D) S_4 (E) S_5

57.) (0.5 pts) What is the 10th state in your list?
(A) S_1 (B) S_2 (C) S_3 (D) S_4 (E) S_5

58.) (3 pts) (You saw the 2nd column of questions above, right?) Define $V^\pi(s)$ precisely:

(A) The expected cumulative discounted reward starting at state s and following policy π.

(B) A lower-bound on the cumulative discounted reward starting at state s and following policy π.

(C) The expected immediate discounted reward starting at state s and following policy π.

(D) The expected immediate reward starting at state s and following policy π.

(E) The optimal action to take at state s according to policy π.

VIII Clustering (11 pts total)

59.) (1 pts) True (A) or False (B): K-means is equivalent to fitting a **Gaussian mixture model** with **an identity covariance matrix** (i.e., spherical) for all clusters.

60.) (1 pts) True (A) or False (B): The variance of a Gaussian mixture model **increases** as the number of mixture components k increases.

61.) (3 pts) Each iteration, the k-Means algorithm produces a k-way partitioning of the input data set. Is it possible for k-Means to revisit a partitioning after leaving it?

(A) Yes, and it does this often with any arbitrary data set

(B) Yes, but it is a rare occurrence with any arbitrary data set

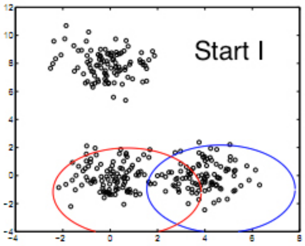
(C) Maybe, depending upon the particular data set

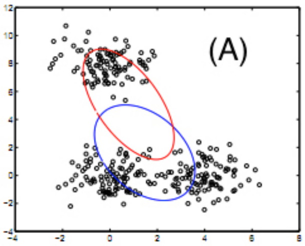
(D) No, this is not possible

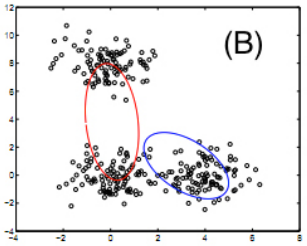
EM Mixup

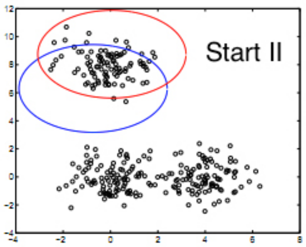
For homework, a student in machine learning was estimating a mixture of two Gaussians based on 2D data shown below. The student randomly initialized the mixture twice, yielding the two figures marked "Start I" and "Start II." The learning then fit the models using EM, generating figures A–D, with each figure showing a different step of the algorithm. However, the figures got mixed up and we need your help to fix them!

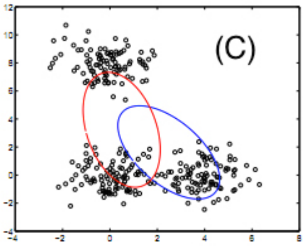
Find the sequences showing the learning process, beginning with each of the starting figures. It may help you to draw an arrow from one figure to another to indicate how they follow from each other (you should draw only four arrows total). Each sequence will begin with one of the starting figures and will contain three figures total; each figure is used once.

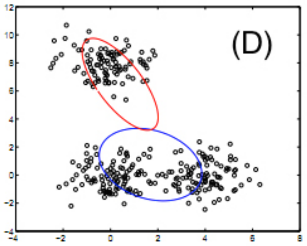












62.) (1.5 pts) What is the second figure in sequence I? (the one immediately after "Start I")

63.) (1.5 pts) What is the third figure in sequence I?

64.) (1.5 pts) What is the second figure in sequence II? (the one immediately after "Start II")

65.) (1.5 pts) What is the third figure in sequence II?

That's it! Relax a bit, check your answers, and check that your name and PennID are on your bubble sheet. Also, remember to fill out the final project survey. Then, quietly turn in your exam. Please be mindful of others around you who are still completing the exam. Best of luck in all your future endeavors!

Reference Page

Decision Trees Let D be the data, C be the class attribute, and A be an attribute (which could be C). The accessor $A.values$ denotes the values of attribute A.

$$\begin{aligned}
\mathit{Info}(D) &= \sum_{i \in C.values} -\frac{|D_i|}{|D|} * \lg\left(\frac{|D_i|}{|D|}\right) \\
\mathit{Info}(A, D) &= \sum_{j \in A.values} \frac{|D_j|}{|D|} * \mathit{Info}(D_j) \\
\mathit{Gain}(A, D) &= \mathit{Info}(D) - \mathit{Info}(A, D) \\
\mathit{GainRatio}(A, D) &= \frac{\mathit{Gain}(A, D)}{\mathit{SplitInfo}(A, D)} \\
\mathit{SplitInfo}(A, D) &= \mathit{Info}(D) \text{ considering } A \text{ as the class attribute } C. \\
&= \sum_{i \in A.values} -\frac{|D_i|}{|D|} * \lg\left(\frac{|D_i|}{|D|}\right)
\end{aligned}$$

Linear Regression $min_{\boldsymbol{\theta}} \ \frac{1}{2n} \sum_{i=1}^{n} \left(h_{\boldsymbol{\theta}}(\mathbf{x}_i) - y_i\right)^2 + \lambda \sum_{j=1}^{d} \theta_j^2$

Perceptron Update Rule $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + y_i \mathbf{x}_i$ if $\mathbf{x}_i$ is misclassified

Support Vector Machines

$$\text{Primal:} \quad \min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|_2^2 \quad \text{s.t.} \quad \forall i \ y_i(\mathbf{w} \cdot \mathbf{x} - b) \geq 1$$

$$\text{Dual:} \quad \min_{\boldsymbol{\alpha}} \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) - \sum_{i=1}^{n} \alpha_i \quad \text{s.t.} \quad \sum_{i=1}^{n} y_i \alpha_i = 0 \text{ and } \forall i \ \alpha_i \geq 0$$

Probability Theory

$$\begin{aligned}
P(A_1, \ldots, A_m \mid B_1, \ldots, B_n) &= \frac{P(B_1, \ldots, B_n \mid A_1, \ldots, A_m) P(A_1, \ldots, A_m)}{P(B_1, \ldots, B_n)} \\
P(A_1, \ldots, A_m \mid B_1, \ldots, B_n) &= \frac{P(A_1, \ldots, A_m, B_1, \ldots, B_n)}{P(B_1, \ldots, B_n)}
\end{aligned}$$

Logistic Regression

$$P(y = 1 \mid \mathbf{x}) = \frac{1}{1 + \exp(-\boldsymbol{\theta}^\intercal \mathbf{x})} \qquad P(y = 0 \mid \mathbf{x}) = 1 - P(y = 1 \mid \mathbf{x})$$

$$\min_{\boldsymbol{\theta}} \ -\sum_{i=1}^{n} \left[y_i \lg P(y_i = 1 \mid \mathbf{x}_i) + (1 - y_i) \lg P(y_i = 0 \mid \mathbf{x}_i)\right]$$

AdaBoost

$$\beta_t = \frac{1}{2} \ln \frac{1 - \epsilon_t}{\epsilon_t} \quad \text{where} \quad \epsilon_t = \sum_{i=1}^{n} w_t(\mathbf{x}_i) \mathbb{1}[y_i \neq h_t(\mathbf{x}_i)]$$

$$w_{t+1}(\mathbf{x}_i) = \frac{w_t(\mathbf{x}_i) \exp(-\beta_t y_i h_t(\mathbf{x}_i))}{Z_t} \quad \text{where} \quad Z_t = \sum_{i=1}^{n} w_t(\mathbf{x}_i) \exp(-\beta_t y_i h_t(\mathbf{x}_i))$$

$$H(\mathbf{x}) = \text{sign}\left(\sum_{t=1}^{T} \beta_t h_t(\mathbf{x})\right)$$

Neural Networks

$$\min_{\Theta} -\frac{1}{n}\left[\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\log(h_\Theta(\mathbf{x}_i))_k + (1-y_{ik})\log\Big(1-(h_\Theta(\mathbf{x}_i))_k\Big)\right] + \frac{\lambda}{2n}\sum_{l=1}^{L-1}\sum_{i=1}^{s_{l-1}}\sum_{j=1}^{s_l}\left(\Theta_{ji}^{(l)}\right)^2$$

Deep Learning

$$\begin{aligned} E(\mathbf{v},\mathbf{h}) &= -\sum_{i,j} v_i h_j \Theta_{ij} \\ P(\mathbf{v},\mathbf{h}) &\propto e^{-E(\mathbf{v},\mathbf{h})} \\ P(\mathbf{v}) &\propto \sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})} \\ \frac{\partial \log P(\mathbf{v})}{\partial \Theta_{ij}} &= \mathbb{E}_0(v_i h_j) - \mathbb{E}_\infty(v_i h_j) \end{aligned}$$

Reinforcement Learning

$$\begin{aligned} R_t &= \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \\ V^\pi(s) &= E_\pi\{R_t | s_t = s\} \end{aligned}$$

Bellman Equations:

$$\begin{aligned} V^*(s) &= \max_a Q^*(s,a) \quad = \max_a \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V^*(s')] \\ Q^*(s,a) &= \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V^*(s')] \end{aligned}$$

TD-update rule:

$$V(s_t) \leftarrow V(s_t) + \alpha\left[r_{t+1} + \gamma V(s_{t+1}) - V(s_t)\right]$$

Q-update rule:

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha\left[r_{t+1} + \gamma Q(s_{t+1},a_{t+1}) - Q(s_t,a_t)\right]$$

Unsupervised Learning

K-means:

$$\arg\min_{\boldsymbol{\mathcal{S}}} \sum_{i=1}^{k}\sum_{\mathbf{x}\in\mathcal{S}_i} \|\mathbf{x}-\boldsymbol{\mu}_i\|_2^2 \ \text{ where } \boldsymbol{\mathcal{S}} = \{\mathcal{S}_1,\ldots,\mathcal{S}_k\} \text{ is a partitioning over } X \text{ and } \boldsymbol{\mu}_i = \text{mean}(\mathcal{S}_i)$$

Gaussian mixture models:

$$P(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^d|\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

Answer Key

1.) Any answer

2.) A - true

3.) B - false

4.) B - false

5.) A - true

6.) B - false

7.) B - false

8.) A - true

9.) A - true

10.) A - true

11.) E - ∞ since it can shatter an arbitrary set of points.

12.) B - local optimum

13.) A - global optimum

14.) B - local optimum

15.) A - global optimum

16.) B - local optimum

17.) A - global optimum

18.) A - zero-one loss

19.) C - log loss

20.) C - log loss

21.) D - hinge loss

22.) B - (within-class) sum of squared error (from mean)

23.) B - Bias remains unchanged

24.) A - variance decreases

25.) D - in fit() and predict()

26.) A - bias decreases

27.) A - variance decreases

28.) B - problematic. Params should be chosen on a validation subset of train data, not test data.

29.) A - when the priors are equal

30.) B - $h < 1$

31.) B - no, since H1, H2, and H3 stay the support vectors

32.) C - 3. We will need a cubic function

33.) B - false

34.) E - The optimal $\boldsymbol{\theta} = [0, 0, 0, 1]$, with $\|\boldsymbol{\theta}\|_2 = 1$

35.) A - $\|\mathbf{w}\|_2 \rightarrow 0$

36.) B - 7

37.) B - 3/8

38.) B - 5/9

39.) C - 1/5

40.) D - 1/3

41.) A - Yes: $P(Sunglasses? = yes|sunny70s) \propto 3/4 * 5/8 * 1/9 = 5/96\alpha(PREDICTION!)$
$P(Sunglasses? = no|sunny70s) \propto 1/4 * 1/2 * 2/5 = 5/100\alpha$.

42.) D - 9/11 $P(win|colormusic) \propto 2/4*3/4*3/4 = 9/32\alpha = 9/11$ $P(\neg win|colormusic) \propto 2/4 * 2/4 * 1/4 = 2/32\alpha = 2/11$

43.) AD or BD

44.) BD or AD

45.) AE – the positive points will fall near (1,1), while the negative points will fall around (0,0), (0,1), and (1,0).

46.) D

47.) D - 25

48.) C

49.) B

50.) A

51.) B

52.) A

53.) B

54.) A

55.) B

56.) A

57.) B

58.) A

59.) A - true

60.) A - true

61.) D - not possible. The partitioning must change each iteration, otherwise the algorithm converges. Also, the mean-squared error decreases monotonically, and so it is impossible to revisit a configuration.

62.) C

63.) B

64.) A

65.) D