CIS 419/519 Introduction to Machine Learning
Fall 2014 Midterm

Instructions:

- Please turn off your cell phone.
- This examination is closed-book and closed-notes. Calculators and any other external resources are prohibited.
- 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 either (1) erase the bubble completely, or (2) cross it out via single X and fill in the correct bubble.
- 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 120 points total. Students in CIS 419 will be graded out of 110 points; points earned beyond 110 will not count as extra credit. Students in CIS 519 will be graded out of 120 points.
- I wish each of you the best of luck.

~~Important Grading Note~~

~~In an effort to discourage guessing, this exam will be scored in the following manner:~~

- ~~Correct answers will increase the total score by the number of points for that question~~
- ~~Incorrect answers will *decrease* the total score by $\frac{1}{4}$ of the total number of points for that question~~
- ~~Omitted answers will not change the total score~~

~~For example, an incorrect answer on a question worth 4 points would change the total score by -1 point. Therefore, if you are not certain of a question, it may be better to leave the answer blank than to guess.~~

Please note that negative grading will not be used,
so please disregard this section.

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.

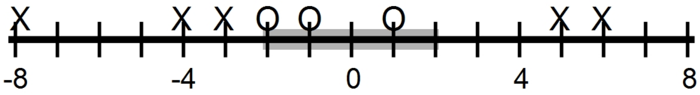
II General Knowledge (18 pts total)

Choose the correct answer for each question below:

1.) (1 pts) I ♡ CIS 419/519 because...

(A) I get points on exams for checking that the exam has all 12 pages (hint hint!)
(B) Machine learning is just kinda awesome!
(C) I'm super excited about the final project
(D) Although it is hard work, I'm learning a lot
(E) All of the above

2.) (3 pts) Consider the following one-dimensional integer classification problem, with the shaded region of the number line specifying the *true* region that should be labeled "O" and the marked points denoting the given training data.



For what value of k will the k-nearest neighbor classifier achieve 100% generalization accuracy on this problem? (Assume ties are broken randomly.)
(A) 1 (B) 2 (C) 3 (D) 4 (E) None of the above

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

3.) (1 pts) The training error of 1-Nearest Neighbor is 0.

4.) (1 pts) The accuracy of a classifier on the training data is a poor estimate of the generalization error. Instead, cross-validation can be used to produce a more reliable estimate of the generalization error.

5.) (1 pts) Naïve Bayes is a two-level Bayesian network, with the class node influencing each attribute node, and the attribute nodes conditionally independent of each other given the class.

6.) (1 pts) Maximizing the likelihood of linear regression yields multiple local optimums; the specific optimum you reach depends on your initial parameter values.

7.) (1 pts) 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.

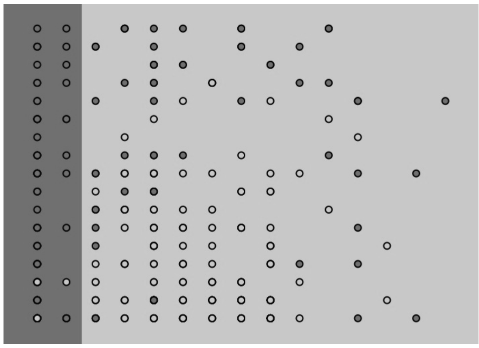
8.) (1 pts) In computer vision, asking a user to outline objects in the training image that would best improve the learned model would be an example of *active learning*.

(1 pt each) For each plot below, name the machine learning algorithm that *most likely* generated the decision surface given the training data. Answers may be used more than once. Your possible choices are:

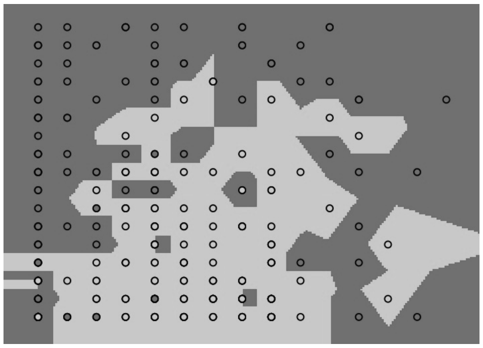
(A)(B) C4.5 decision tree	(B)(D) perceptron
(A)(C) decision stump	(B)(E) SVM with a linear kernel
(A)(D) 1-nearest neighbor	(C)(D) SVM with a polynomial kernel
(A)(E) k-nearest neighbor (with $k > 1$)	(C)(E) SVM with a radial basis function kernel
(B)(C) logistic regression	(D)(E) AdaBoost with decision stumps

Note that you must fill in TWO bubbles for each answer. (For example, you must fill in both bubbles (A) and (B) to specify a decision tree.)

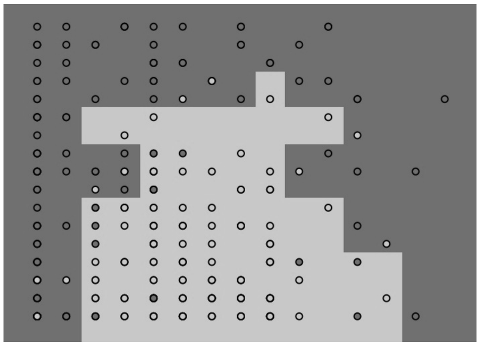
9.)



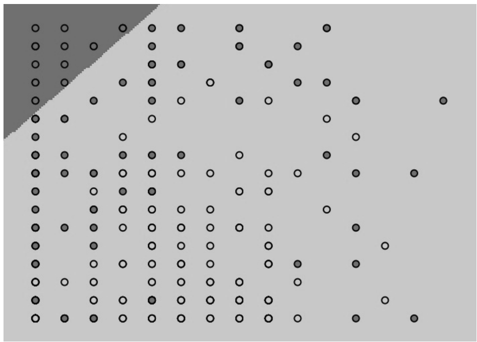
13.)



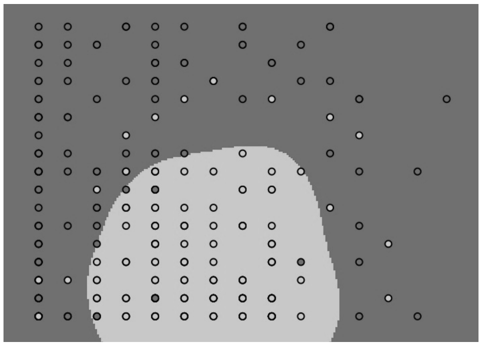
10.)



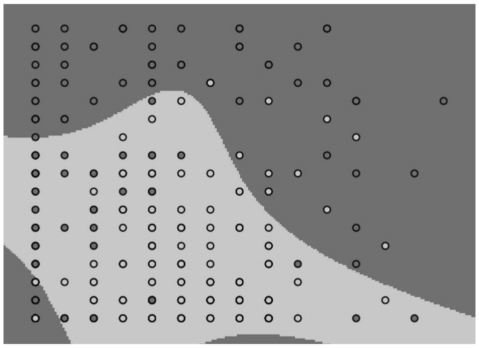
14.)



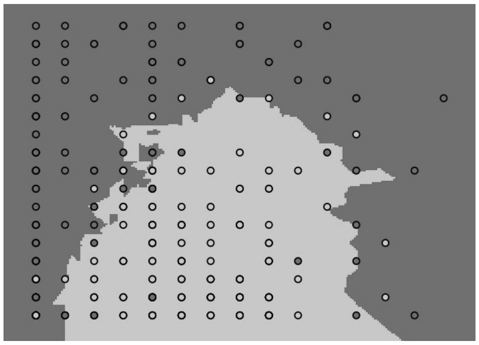
11.)



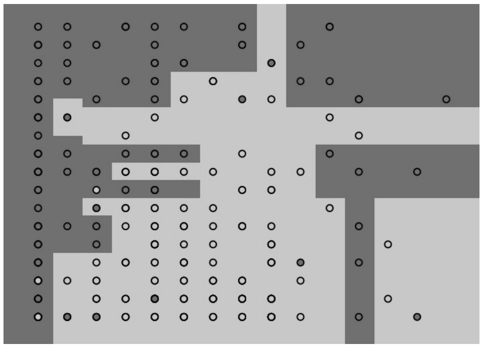
15.)



12.)



16.)



III Bias/Variance (20 pts total)

(2 pts each) Each plot below shows learning curves for a regression problem. The learning curves were generated using the same data set with different amounts of regularization.

For each learning curve plot, choose whether the regressor has:

(A) high bias / high variance

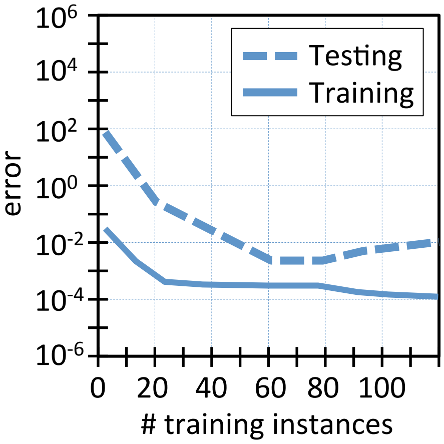
(B) high bias / low variance

(C) a good balance between bias and variance

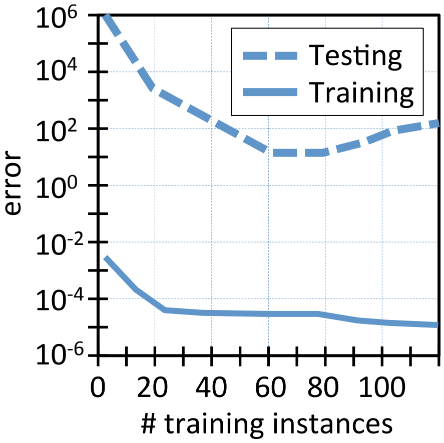
(D) low bias / high variance

(E) low bias / low variance

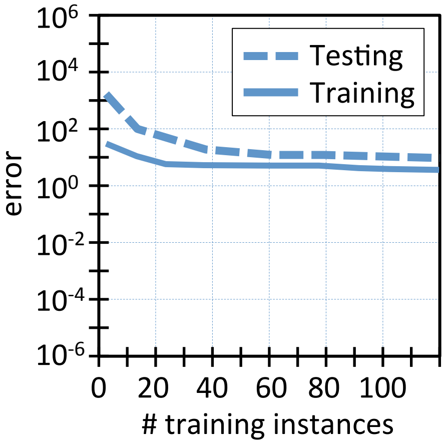
17.)



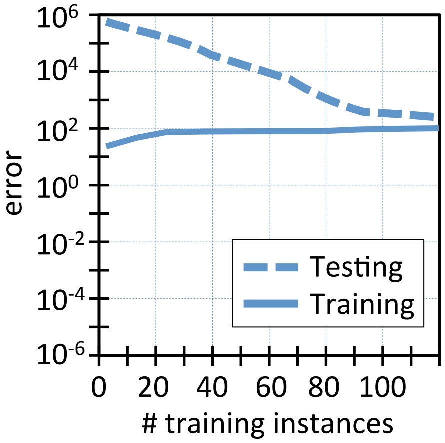
18.)



19.)



20.)



Often, we can control the bias and variance of the model by varying classifier parameters. For each parameter below, state its general effect on bias/variance.

(2 points) For k-nearest neighbor, what is the effect on the bias/variance as k increases?

21.) (A) Bias decreases (B) Bias remains unchanged (C) Bias increases

22.) (A) Variance decreases (B) Variance remains unchanged (C) Variance increases

(2 points) For unpruned decision trees, what is the effect on the bias/variance as the maximum allowable depth of the tree increases?

23.) (A) Bias decreases (B) Bias remains unchanged (C) Bias increases

24.) (A) Variance decreases (B) Variance remains unchanged (C) Variance increases

(2 points) For logistic regression, what is the effect on the bias/variance as the number of training instances increases?

25.) (A) Bias decreases (B) Bias remains unchanged (C) Bias increases

26.) (A) Variance decreases (B) Variance remains unchanged (C) Variance increases

(2 points) For SVMs with a Gaussian kernel and $C \approx 0$, what is the effect on the bias/variance as the Gaussian kernel bandwidth σ increases?

27.) (A) Bias decreases (B) Bias remains unchanged (C) Bias increases

28.) (A) Variance decreases (B) Variance remains unchanged (C) Variance increases

(2 points) For SVMs with a Gaussian kernel and an extremely large C value (i.e., $C \to \infty$), what is the effect on the bias/variance as the Gaussian kernel bandwidth σ increases?

29.) (A) Bias decreases (B) Bias remains unchanged (C) Bias increases

30.) (A) Variance decreases (B) Variance remains unchanged (C) Variance increases

(2 points) For AdaBoost with decision stumps, what is the effect on the bias/variance as the number of boosting iterations T increases?

31.) (A) Bias decreases (B) Bias remains unchanged (C) Bias increases

32.) (A) Variance decreases (B) Variance remains unchanged (C) Variance increases

IV Experimental Protocol (14 pts total)

For each scenario below, tell whether the experimental setup is (A) okay or (B) problematic.

33.) (1 pts) A company claims great success after achieving 98% classification accuracy on detecting fraudulent transactions (a binary classification task where fraud is rare). Their data consisted of 50 positive examples of fraud and 5,000 negative examples.
(A) Okay (B) Problematic

34.) (1 pts) A project team split their data into training and test sets. Using cross-validation over their training data, they chose the best model parameters. They built a model using these parameters and their training data, and then report their error on test data.
(A) Okay (B) Problematic

35.) (1 pts) A government organization performed a feature selection procedure on the full data and reduced their large feature set to a smaller set. Then, they split the data into test and training portions, and report their test error.
(A) Okay (B) Problematic

36.) (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.
(A) Okay (B) Problematic

(8 pts total) Imagine you're writing a program to compute the learning curves for two classifiers (A and B), averaged over 20 trials of 10-fold cross-validation. Put the following pseudocode steps in their correct order:

(A)(B) Load the data set
(A)(C) Shuffle the entire data set
(A)(D) Shuffle each individual fold
(A)(E) Split data into 10 folds
(B)(C) For each testing fold, do:
(B)(D) For each percent of the training data, do:
(B)(E) For each trial, do:
(C)(D) For each training instance, do:
(C)(E) Train model A, then evaluate on test data
(D)(E) Train model B, then evaluate on test data

There are 8 steps total. Note that you must fill in TWO bubbles per answer. Some pseudocode steps may be used more than once, and some may not be used at all. Artificially, we require that model A be trained sometime before model B in the order.

Any steps after a looping step ("For . . . , do:") will be considered to occur *inside* the loop body. E.g., choosing AB as the 1^{st} step, CD as the 2^{nd} step, and CE as the 3^{rd} step would correspond to:

```
Load the data set
For each training instance, do:
    Train model A, then evaluate on test data
```

37.) What is the first step?

38.) What is the second step?

39.) What is the third step?

40.) What is the fourth step?

41.) What is the fifth step?

42.) What is the sixth step?

43.) What is the seventh step?

44.) What is the eighth step?

45.) (2 pts) [Make sure you saw #41–44 above!] 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

V Decision Trees (14 pts total)

46.) (1 pts) Consider a decision stump and a decision tree of depth 3, both trained on the same data. Which has higher bias?
(A) The decision stump (B) The decision tree of depth 3 (C) Both have equal bias

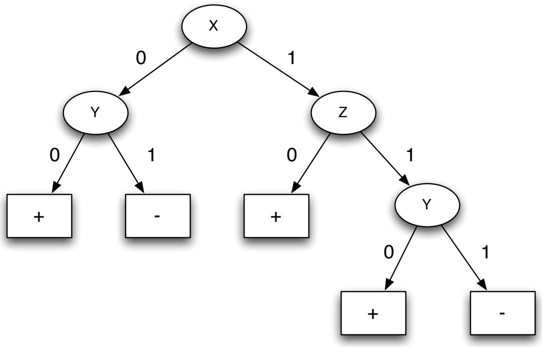
47.) (1 pts) If an event is certain, its entropy is:
(A) 0 (B) between 0 and 1/2 (C) 1/2 (D) between 1/2 and 1 (E) 1

48.) (1 pts) If two events have equal probability, their entropy is:
(A) 0 (B) between 0 and 1/2 (C) 1/2 (D) between 1/2 and 1 (E) 1

SIZE	SMELLY?	HAIR?	COLOR	WUMPUS?
large	yes	yes	brown	yes
medium	yes	yes	brown	yes
small	yes	yes	brown	no
large	no	yes	brown	no
medium	yes	yes	purple	yes
large	yes	yes	purple	yes
medium	no	no	green	no
small	yes	no	green	no

49.) (5 pts) Given the data above, what is the information gain for splitting on the "COLOR" attribute? (Use "WUMPUS?" as the class attribute.)
(A) 0.125 (B) 0.25 (C) 0.5 (D) 0.75 (E) None of the above

Reduced Error Pruning The next three questions concern the unpruned decision tree and validation set below. What would be the effect on the decision tree of reduced error pruning using the validation set?



Validation Set:

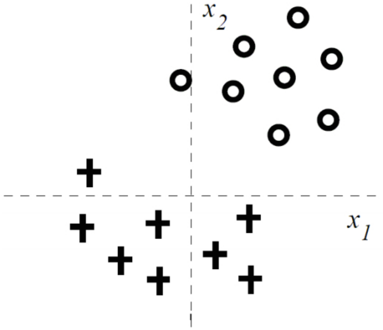
X	Y	Z	*class*
0	0	1	+
0	1	0	−
1	1	0	−
1	1	1	+
1	0	1	+

50.) (2 pts) Would node Y on the LEFT side of the tree be pruned?
(A) Yes (B) No

51.) (2 pts) Would node Z be pruned?
(A) Yes (B) No

52.) (2 pts) Would node Y on the RIGHT side of the tree be pruned?
(A) Yes (B) No

VI Logistic Regression (9 pts)



Consider the two-dimensional classification problem depicted above. Points labeled as '+' correspond to class $y = 1$ and points labeled as '∘' correspond to class $y = 0$.

We can solve this problem with a simple linear regression model

$$P_{\boldsymbol{\theta}}(y = 1 \mid \mathbf{x}) = \frac{1}{1 + \exp(-\theta_0 - \theta_1 x_1 - \theta_2 x_2)} \ ,$$

where x_j denotes the j^{th} feature of $\mathbf{x}$. Since the data is linearly separable, we can obtain a training error of zero. Now, consider a regularized form of the problem, where we try to maximize:

$$\sum_{i=1}^{n} \log P_{\boldsymbol{\theta}}(y_i \mid \mathbf{x}_i) - \lambda \theta_j^2 \ ,$$

where $(\mathbf{x}_i, y_i)$ is the i^{th} training instance. In this case, we'll use a very large value for λ.

Unlike in previous forms of this problem, the regularization term is given by $-\lambda\theta_j^2$ for $j \in \{0, 1, 2\}$. In other words, only ONE of the parameters is regularized in solving the problem. Given the training data above, how will the training error change with regularization of each parameter θ_j using a very large value for λ?

53.) (3 pts) How will the training error change if we regularize θ_0 (i.e., set $j = 0$)?

(A) It will stay the same

(B) It will increase, misclassifying one or more '+' points

(C) It will increase, misclassifying one or more '∘' points

(D) It will increase, misclassifying one or more '+' points and one or more '∘' points

(E) It may stay the same or increase, depending on the initial value for $\boldsymbol{\theta}$.

54.) (3 pts) How will the training error change if we regularize θ_1 (i.e., set $j = 1$)?

(A) It will stay the same

(B) It will increase, misclassifying one or more '+' points

(C) It will increase, misclassifying one or more '∘' points

(D) It will increase, misclassifying one or more '+' points and one or more '∘' points

(E) It may stay the same or increase, depending on the initial value for $\boldsymbol{\theta}$.

55.) (3 pts) How will the training error change if we regularize θ_2 (i.e., set $j = 2$)?

(A) It will stay the same

(B) It will increase, misclassifying one or more '+' points

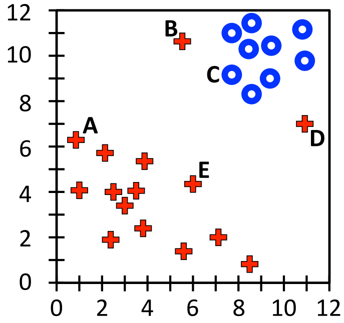
(C) It will increase, misclassifying one or more '∘' points

(D) It will increase, misclassifying one or more '+' points and one or more '∘' points

(E) It may stay the same or increase, depending on the initial value for $\boldsymbol{\theta}$.

VII Support Vector Machines (18 pts total)

The data set below contains two classes of points ('+' and '∘'); five instances in the data set have been labeled as points A–E.



Assume that we are training an SVM with a quadratic kernel (polynomial kernel of degree 2) on the data above. The location of the separating hyperplane will be determined by the slack penalty C.

For each question below, you may need to fill in multiple bubbles; the question will be marked as correct only if ALL correct bubbles are filled in.

56.) (2 pts) Consider the decision boundary for very large values of C (i.e., as $C \to \infty$). Which points A–E would be support vectors? (Other points besides A–E may be support vectors too.)

57.) (2 pts) Consider the decision boundary for very large values of C (i.e., as $C \to \infty$). Which points A–E would be classified correctly by this decision boundary?

58.) (2 pts) Consider the decision boundary for $C \approx 0$. Which points A–E would be support vectors? (Other points besides A–E may be support vectors too.)

59.) (2 pts) Consider the decision boundary for $C \approx 0$. Which points A–E would be classified correctly by this decision boundary?

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 $\phi(\mathbf{x}) = [1, x_1, x_2, x_1 x_2]$, 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.

60.) (4 pts) 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

61.) (2 pts) What kernel most closely corresponds to the feature transformation defined by ϕ?
(A) linear (B) quadratic (C) cubic (D) Gaussian (E) sigmoid

62.) (4 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 ∞

VIII Probabilistic Models (16 pts total)

Given the following data, construct a naïve Bayes classifier with Laplace smoothing to identify whether a creature is a Snark.

FEATHERS?	COLOR	SNARK?
yes	gold	yes
yes	gold	yes
yes	gold	yes
yes	gold	no
yes	purple	yes
yes	purple	no
no	green	no
no	green	no

63.) (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

64.) (1 pts) What is $P(\text{Feathers?} = \textit{yes} \mid \text{Snark?} = \textit{yes})$?
(A) 1/2 (B) 3/4 (C) 5/6 (D) 1 (E) None of the above

65.) (1 pts) What is $P(\text{Feathers?} = \textit{yes} \mid \text{Snark?} = \textit{no})$?
(A) 2/5 (B) 1/2 (C) 3/5 (D) 3/4 (E) None of the above

66.) (1 pts) What is $P(\text{Color?} = \textit{gold} \mid \text{Snark?} = \textit{yes})$?
(A) 1/2 (B) 2/3 (C) 3/4 (D) 4/5 (E) None of the above

67.) (1 pts) What is $P(\text{Color?} = \textit{gold} \mid \text{Snark?} = \textit{no})$?
(A) 1/4 (B) 1/3 (C) 2/5 (D) 1/2 (E) None of the above

68.) (1 pts) What is $P(\text{Color?} = \textit{green} \mid \text{Snark?} = \textit{yes})$?
(A) 0 (B) 1/8 (C) 1/5 (D) 1/4 (E) None of the above

69.) (5 pts) What is the probability that a purple creature with feathers is a Snark, according to the naïve Bayes model?
(A) 5/36 (B) 5/18 (C) 1/2 (D) 5/8 (E) None of the above

70.) (3 pts) Under what conditions is the maximum a posteriori (MAP) hypothesis equivalent to the maximum likelihood hypothesis (MLE)? (Fill in one or more bubbles to choose all that apply) (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

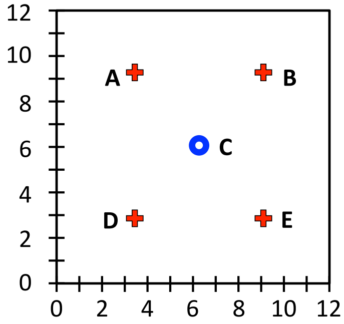
(C) When the posterior probabilities are equal

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

(E) None of the above

IX Ensemble Techniques (11 pts total)

The data set below contains two classes of points ('+' and '∘'); the instances are labeled A–E.



Consider training AdaBoost with decision stumps to solve the classification problem.

71.) (3 pts) Which instances will have their weight increased at the end of the first boosting iteration? (Fill in one or more corresponding bubbles)

72.) (3 pts) What is the **minimum** number of iterations that the ensemble could take to achieve zero training error?
(A) 2 (B) 3 (C) 4 (D) 5 (E) The ensemble will never achieve zero training error

73.) (1 pts) Is it possible to add another example to the data set to guarantee that boosting achieves a zero training error in just two iterations?
(A) Yes (B) No

74.) (1 pts) If you were to add another example to the data set to guarantee that boosting achieves a zero training error in just two iterations, where would you place that point?

(A) Near instance A

(B) Midway between instances A and B

(C) Near instance C

(D) Midway between instances A and C

(E) Nowhere; it isn't possible to achieve zero training error in just two boosting iterations

75.) (3 pts) A colleague informs you that they've created an ensemble learning method that achieves a cross-validated test error of 0.001% on a large data set where the attributes and labels contain 5% noise. Which do you do?

(A) Congratulate your colleague. This is an amazing discovery.

(B) Understand that this behavior is typical of boosting methods, which repeatedly focus on mispredicted instances to drive down the ensemble's error without overfitting.

(C) Infer that the ensemble uses feature diversity to compensate for the noise, with each ensemble member focusing on a different subset of features.

(D) Assume that your colleague used stacking to construct the ensemble, leveraging the strengths of each ensemble member to achieve the high performance.

(E) State that this is impossible.

That's it! Relax a bit, check your answers, and check that your name and PennID are on your bubble sheet. Then, quietly turn in your exam. Please be mindful of others around you who are still completing the exam.

We still have a few students who need to take a makeup midterm, so please **DO NOT post questions or comments about the exam on Piazza**.

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

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

Support Vector Machines

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

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

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

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

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

AdaBoost

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

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

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

Answer Key

1.) Any, but preferably E. :-)

2.) C - 3

3.) A - true

4.) A - true

5.) A - true

6.) B - false

7.) B - false

8.) A - true (THROWN OUT DUE TO WORDING - CORRECT ANSWERS COUNT AS EXTRA CREDIT)

9.) AC - Decision Stump

10.) DE - AdaBoost w/ Decision Stumps

11.) CE - SVM RBF kernel

12.) AE - 9-NN

13.) AD - 1-NN (AE - k-NN ALSO ACCEPTED DUE TO ERROR IN FIGURE)

14.) BD - Perceptron

15.) CD - SVM polynomial kernel

16.) AB - Decision Tree

17.) C - good balance

18.) D - low bias / high variance

19.) B - high bias / low variance

20.) A - high bias / high variance (‘B’ ALSO ACCEPTED)

21.) C - Bias increases

22.) A - variance decreases

23.) A - Bias decreases

24.) C - variance increases

25.) B - Bias remains unchanged

26.) A - variance decreases

27.) B - Bias remains unchanged

28.) B - Variance remains unchanged

29.) B - Bias remains unchanged

30.) A - variance decreases

31.) A - bias decreases

32.) A - variance decreases

33.) B - problematic. Just predicting the majority class will yield 99% accuracy.

34.) A - okay.

35.) B - problematic. Performing feature selection on the full data set leaks test information.

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

37.) AB - Load the data set

38.) BE - For each trial, do:

39.) AC - Shuffle the entire data set

40.) AE - Split data into 10 folds

41.) BC - For each testing fold, do:

42.) BD - For each percentage of the training data:

43.) CE - Train model A, then evaluate on test data

44.) DE - Train model B, then evaluate on test data

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

46.) A - the decision stump

47.) A - 0

48.) E - 1

49.) C - 0.5 $Info(D) = 4/8 * lg4/8 + 4/8 * lg4/8 = 1$
$Info(Color, D) = 4/8 * 1 + 2/8 * 0 + 2/8 * 0 = 4/8$
$Gain(Color, D) = 1 - 4/8 = 5/10 = 0.5$

50.) B - No - the first two validation instances that filter to the left are correct

51.) A - Yes - after pruning node Y, always predicting + down this branch, so why not?

52.) A - Yes - last two validation instances that filter down to node Y would be better off if we just predicted +

53.) B - increase, misclassifying some ’+’ points

54.) A - it will stay the same

55.) D - increase, misclassifying some ’+’ and ’◦’

56.) BCDE

57.) ABCDE

58.) BCDE (THROWN OUT, CORRECT ANSWER COUNTED AS EXTRA CREDIT)

59.) ACE

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

61.) B - quadratic

62.) A - $\|\mathbf{w}\|_2 \to 0$

63.) B - 7

64.) C - 5/6

65.) B - 1/2

66.) E - 4/7 (INCORRECT ANSWER B - 2/3 ALSO ACCEPTED DUE TO ERROR IN SLIDES)

67.) E - 2/7 (INCORRECT ANSWER B - 1/3 ALSO ACCEPTED DUE TO ERROR IN SLIDES)

68.) E - 1/6

69.) D - 5/8 Let s = snark, f = feathers, p = purple. Then
$P(s|f,p) \propto P(s)P(f|s)P(p|s) = \alpha * \cancel{3/6} * 5/6 * \cancel{2/7} = 5/6\alpha$
$P(\neg s|f,p) \propto P(\neg s)P(f|\neg s)P(p|\neg s) = \alpha * \cancel{3/6} * 3/6 * \cancel{2/7} = 3/6\alpha$
$\alpha = 1/(5/6 + 3/6) = 8/6$
$P(s|f,p) = 5/6\alpha = 5/8$
$P(\neg s|f,p) = 3/6\alpha = 3/8$

70.) A - when the priors are equal (THROWN OUT, CORRECT ANSWER COUNTS AS EXTRA CREDIT)

71.) C - the first decision stump with least error will always predict positive, and so only the ’o’ point is misclassified.

72.) B - 3. The first iteration will misclassify point C; the 2nd iteration will have a high weight for C, and so will misclassify 2 positive examples; therefore, a 3rd iteration is needed.

73.) B - No. The simplest addition would be at the center, or between any two positive instances, but you’d still have 3 different decision regions, and so it will always take a minimum of 3 iterations.

74.) E - Nowhere (see explanation of previous question)

75.) E - state that this is impossible; 5% label noise fundamentally limits the performance that can be obtained.