CSC411 Midterm Winter 2019
Machine Learning and Data Mining
Friday, Feburary 15, 2019

Name: ______________________________

Student number: ____________________

This is a closed-book test. It is marked out of 15 marks. Please answer ALL of the questions. Here is some advice:

- The questions are NOT arranged in order of difficulty, so you should attempt every question.
- Questions that ask you to "briefly explain" something only require short (1-3 sentence) explanations. Don't write a full page of text. We're just looking for the main idea.
- None of the questions require long derivations. If you find yourself plugging through lots of equations, consider giving less detail or moving on to the next question.

.

TF: ______ / 4

MC: ______ / 4

Q9: ______ / 1

Q10: ______ / 1

Q11: ______ / 2

Q12: ______ / 1

Q13: ______ / 2

Final mark: ______ / 15

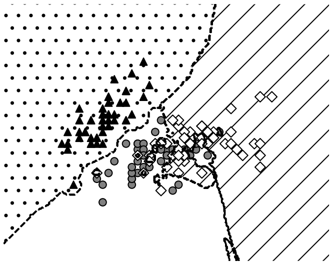
1. True **FALSE** For all real differentiable functions $f : \mathbb{R}^n \mapsto \mathbb{R}$ with at least one local minimum and given any initial point $x \in \mathbb{R}^n$, there exists a learning rate sequence such that the gradient descent algorithm converges to a local minimum of f.

2. **TRUE** False Bob has a magical learning algorithm which returns the true labelling function regardless of the training set. He claims his algorithm has low bias, since its predictions are always correct, but high variance, since its predictions are quite different for different datapoints. Is Bob correct about bias?

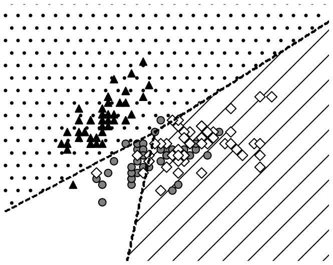
3. True **FALSE** Is Bob correct about variance?

4. True **FALSE** SVM maximizes the objective $\frac{1}{2}\|w\|_2^2$, subject to $y^{(i)}t^{(i)} \geq 1$ for all i, where w is the weights of the decision boundary, $y^{(i)} = w^\top x^{(i)} + b$ is the ith prediction and $t^{(i)} \in \{\pm 1\}$ is the ith label.

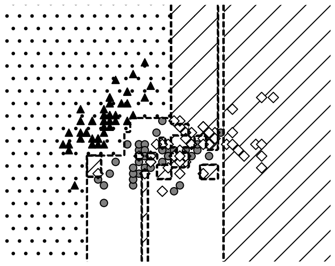
5. (1 point) Which of the following decision boundaries is most likely to be generated by a k-NN?



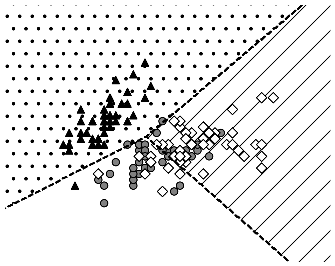
A.



B.



C.



D.

6. (1 point) Which of the following statements about ensemble methods is true?

   A. Combining weak learners using bagging is good since it can reduce the variance.

   B. Combining strong learners using boosting is good since it can reduce the bias.

   C. Combining weak learners using boosting is good since it can reduce the variance.

   D. Combining strong learners using bagging is good since it can reduce the variance.

7. (1 point) Consider the sigmoid function $f(x) = \frac{1}{1+e^{-x}}$. The derivative $f'(x)$ is

   A. $f(x) \log f(x) + (1 - f(x)) \log(1 - f(x))$

   B. $f(x)(1 - f(x))$

   C. $f(x) \log f(x)$

   D. $f(x)(1 + f(x))$

8. (1 point) In soft margin SVMs, the slack variables $\xi^{(i)}$ defined in the constraints $y^{(i)}(w^\top x^{(i)}) \geq 1 - \xi^{(i)}$ have to be

   A. < 0

   B. ≤ 0

   C. > 0

   D. ≥ 0

9. (1 point) Recall that $f : \mathbb{R}^m \to \mathbb{R}$ is convex if for all $x_1, x_2 \in \mathbb{R}^m$ and $\lambda \in [0, 1]$ the following inequality holds:

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2)$$

Suppose $f : \mathbb{R}^m \to \mathbb{R}$ is convex and $a : \mathbb{R}^n \to \mathbb{R}^m$ is linear. Prove that the composition $f \circ a$ is convex.

Grading Notes

1 point: Full marks if (1) used linearity of a (implicit okay, as long as clear), and (2) used convexity property of f to prove the right thing: $f(a(\lambda x_1 + (1 - \lambda)x_2)) \leq \lambda f(a(x_1)) + (1 - \lambda)f(a(x_2)))$. See model answer. Full marks awarded for proving the more general case where a is affine.

0.5 points: Used linearity and proved almost the right thing, or had correct proof overall but for a minor error.

0 points:

- Proved the wrong thing: $f(\lambda(a(x_1)) + (1 - \lambda)a(x_2)) \leq \lambda f(a(x_1)) + (1 - \lambda)f(a(x_2))$, or did not use linearity of a (most common)
- Did not make sense / clearly erroneous logic; e.g., tried to use the fact that composition of two convex functions is convex (but this is not true in general!)
- Blank; not a proof (e.g., just stated conclusion)

Model Answer

We have:

$$\begin{aligned} f(a(\lambda x_1 + (1 - \lambda)x_2)) &= f(\lambda a(x_1) + (1 - \lambda)a(x_2)) \\ &\leq \lambda f(a(x_1)) + (1 - \lambda)f(a(x_2)), \end{aligned}$$

where the first equality holds by linearity of a and the second holds by convexity of f.

10. (1 point) Suppose binary-valued random variables X and Y have the following joint distribution:

	$Y = 0$	$Y = 1$
$X = 0$	2/8	4/8
$X = 1$	1/8	1/8

Determine the information gain $IG(Y|X)$. You may write your answer as a sum of logarithms.

Grading Notes

0.5 for computing the entropy $H[Y]$ correctly,
0.5 for computing the conditional entropy $H[Y|X]$ correctly.

Model Answer

The information gain $IG[Y|X] = H[Y] - H[Y|X]$.

$$H[Y] = -\sum_{y} p(y) \log p(y) = -(\frac{3}{8} \log \frac{3}{8} + \frac{5}{8} \log \frac{5}{8})$$

$$H[Y|X] = -\sum_{x,y} p(x,y) \log p(y|x) = -(\frac{2}{8} \log \frac{1}{3} + \frac{4}{8} \log \frac{2}{3} + \frac{1}{8} \log \frac{1}{2} + \frac{1}{8} \log \frac{1}{2})$$

Note that $H[Y|X] = \sum_x p(x) H[Y|X = x]$. Some students used an extra minus sign.

11. (2 points) Consider the classification problem with the following dataset:

x_1	x_2	x_3	t
0	0	0	1
1	0	0	0
0	1	1	0
1	0	1	1

Your job is to find a linear classifier with weights w_1, w_2, w_3, and b which correctly classifies all of these training examples. None of the examples should lie on the decision boundary.

1. Give the set of linear inequalities the weights and bias must satisfy.

   Grading Notes

   1 point: They only get the full mark when they have the right equations as following.
   1 point: The figure should look like regards to their different c.

   Model Answer

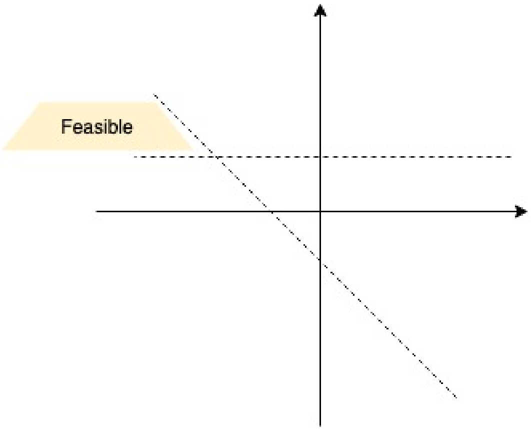
$$
\begin{aligned}
w_0 &> c \\
w_0 + w_1 &< c \\
w_0 + w_2 + w_3 &< c \\
w_0 + w_1 + w_3 &> c
\end{aligned}
$$

2. Shade the feasible region in the two-dimensional slice of weight-space resulting from $b = 5, w_1 = -10$. Place w_2 on the x-axis and w_3 on the y-axis.

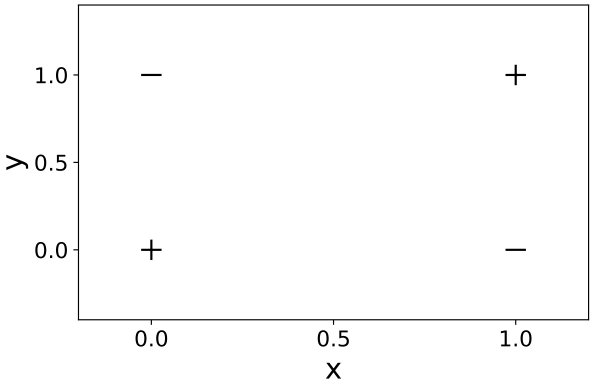
   Grading Notes

   1 point: The figure should look like following regards to their different c.

   Model Answer

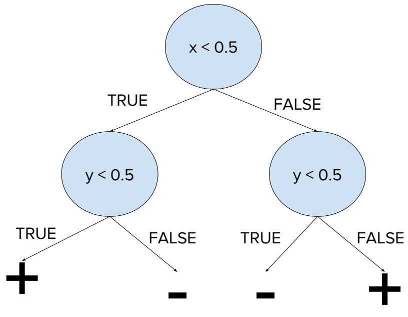

Feasible

12. (1 point) The drawing below shows a dataset. Each example in the dataset has two inputs features x and y and may be classified as a positive example (labelled $+$) or a negative example (labelled $-$). Draw a decision tree which correctly classifies each example in the dataset.

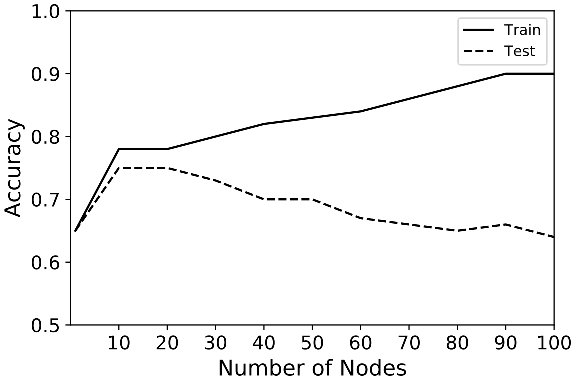


Grading Notes

Model Answer



13. (2 points) The plot below shows training and test accuracies for decision trees of different sizes, when the same finite set of training data is used to train each tree.



1. Describe in one sentence how the training curve would change if the amount of training data used approached infinity.

   Grading Notes

   1 point: If their answer states the training curve goes down (see model answer)

   Model Answer

   The training curve will go down, since, for any number of nodes, it is harder to model the training set accurately as the number of training examples increases.

2. Describe in one sentence how the test curve would change if the amount of training data used approached infinity.

   Grading Notes

   1 point: For stating the test curve will go up and converge to the training curve. 0.5 points for stating something correct (e.g. the test curve will go up) without reference to the convergence of training and test curves

   Model Answer

   The test curve will go up and converge to the training curve, since the training set will capture the entire data distribution as the number of examples approaches infinity.