CS 221, Fall 2020
Quiz 1 Solutions

Multiple Choice

1. **[0 points]** You are allowed to consult online lecture notes, personal notes, and use a basic calculator during this quiz. You may not use the internet to access any websites other than Gradescope or the course website. You may not communicate with others to give or receive aid during or after the quiz. **I agree to abide by these rules and the Stanford Honor Code.**

   (a) Agree

   (b) Disagree

   Answer: (a)

2. **[2 points]** When performing optimization to fit a model in a standard machine learning setup, the loss function is maximized.

   (a) True

   (b) False

   Answer: (b) When performing optimization to fit a model in a standard machine learning setup, the loss function is minimized.

3. **[2 points]** Consider a binary classifier that outputs a score of -6 on a particular input x with label $y = 1$. If the classifier's weight vector was changed so that it instead output a score of -7 on the same input x with label $y = 1$, then the classifier would be:

   (a) *More* `confident in its prediction and its margin would` *increase* `on this example.`

   (b) *Less* `confident in its prediction and its margin would` *increase* `on this example.`

   (c) *More* `confident in its prediction and its margin would` *decrease* `on this example.`

   (d) *Less* `confident in its prediction and its margin would` *decrease* `on this example.`

   Answer: (c) The classifier becomes *more* confident in its prediction of -1, but is incorrect, so its margin *decreases*.

4. **[2 points]** When performing least squares linear regression on a given dataset (by finding the parameters that minimize the squared loss on the training data), the **training loss** will always be 0.

   (a) True

   (b) False

   Answer: (b) The training loss can only be 0 when the training data is colinear.

5. **[2 points]** Select **all** statements that are true.

   (a) A larger step size improves the stability of SGD.

   (b) A larger step size improves the convergence speed of SGD under certain conditions.

   (c) It is common to increase the step size periodically when performing SGD.

   (d) A step size of 0 makes learning impossible.

   Answer: (b), (d) A larger step size may hurt the stability of SGD, but otherwise improves convergence speed in general. It is common to decrease the step size periodically when performing SGD. When the step size is 0 updates do not change the weight vector, so learning is impossible.

6. **[2 points]** Compute the gradient of the linear regression loss function

$$\text{Loss}(x, y, \mathbf{w}) = \begin{cases} \frac{1}{2}(\mathbf{w} \cdot \phi(x) - y)^2 & \text{for } |\mathbf{w} \cdot \phi(x) - y| \leq 1, \\ |\mathbf{w} \cdot \phi(x) - y| - \frac{1}{2} & \text{for } |\mathbf{w} \cdot \phi(x) - y| > 1. \end{cases}$$

What is the gradient, $\nabla_{\mathbf{w}}\text{Loss}(x, y, \mathbf{w})$, evaluated with weight vector $\mathbf{w} = [1, 1]$ at example (x, y) with $\phi(x) = [4, 10]$ and $y = 16$?

(a) $[-4, -10]$

(b) $[4, 10]$

(c) $[-8, -20]$

(d) $[8, 20]$

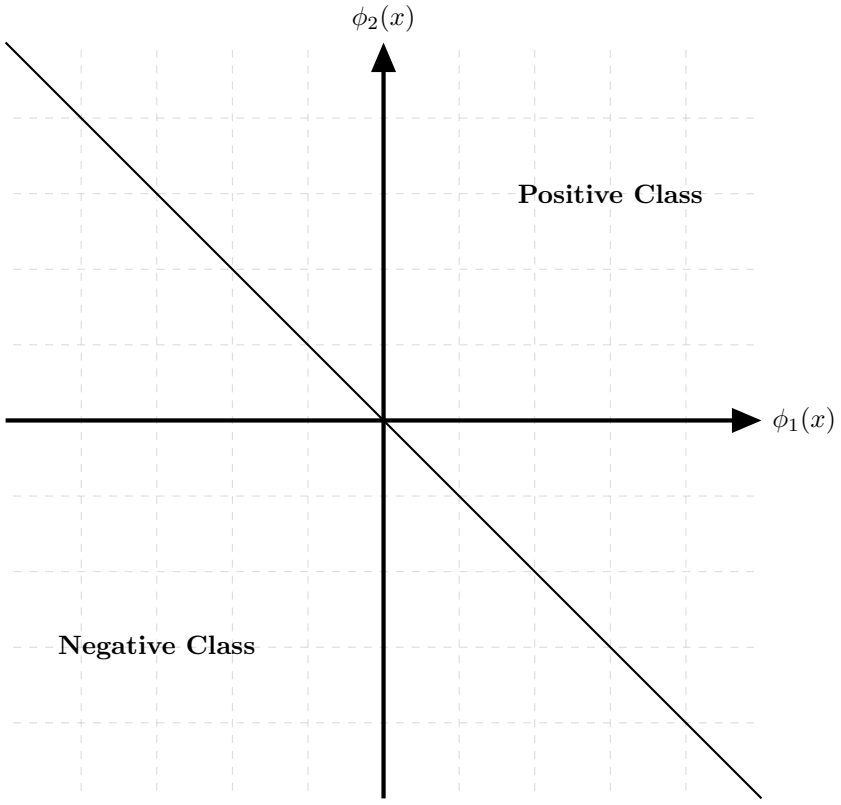
Answer: (a)

$$\nabla_{\mathbf{w}}\text{Loss}(x, y, \mathbf{w}) = \begin{cases} (\mathbf{w} \cdot \phi(x) - y)\phi(x) & \text{for } |\mathbf{w} \cdot \phi(x) - y| \leq 1, \\ \phi(x) & \text{for } \mathbf{w} \cdot \phi(x) - y > 1 \\ -\phi(x) & \text{for } \mathbf{w} \cdot \phi(x) - y < -1 \end{cases}$$

Short Answer

7. **[2.5 points] Linear Binary Classification:** Consider the task of binary classification. Given the linear classifier $f_{\mathbf{w}}(x) = \text{sign}(\mathbf{w} \cdot \phi(x))$ that operates on 2-dimensional feature vectors, draw the decision boundary corresponding to the weight vector $\mathbf{w} = [1, 1]$ and shade (or crosshatch) the region corresponding to positive predictions.

   Answer:



8. **[2.5 points] Gradient Descent:** In this question we will walk through two steps of gradient descent (note, *not* stochastic gradient descent) for linear classification using the hinge loss. The hinge loss is defined to be

$$\text{Loss}_{\text{hinge}}(x, y, \mathbf{w}) = \max\{1 - (\mathbf{w} \cdot \phi(x))y, 0\}$$

Perform two steps of gradient descent with training dataset $\mathcal{D}_{\text{train}} = \{(9, -1), (-1, 1)\}$, feature vector $\phi(x) = [x]$, step size $\eta = 0.1$, and initial weight vector $\mathbf{w} = [0]$. Remember that each gradient descent update requires normalizing by $1/|\mathcal{D}_{\text{train}}|$.

What is the updated weight vector after these two steps are complete? (Enter the 1-dimensional vector as a scalar. Any answer within $\pm.01$ of the exact value will receive full credit.)

Answer: $\mathbf{w} = [-.55]$

$$\nabla_{\mathbf{w}}\text{Loss}_{\text{hinge}}(x, y, \mathbf{w}) = \begin{cases} -\phi(x)y & \text{for } (\mathbf{w} \cdot \phi(x))y < 1, \\ 0 & \text{for } (\mathbf{w} \cdot \phi(x))y > 1. \end{cases}$$

$$\begin{aligned} \mathbf{w^1} &= \mathbf{w^0} - \frac{\eta}{|\mathcal{D}_{\text{train}}|} \sum_{(x,y)\in\mathcal{D}_{\text{train}}} \nabla_{\mathbf{w}}\text{Loss}_{\text{hinge}}(x, y, \mathbf{w}) \\ &= [0] - \frac{0.1}{2}[9 + 1] = [-.5] \\ \mathbf{w^2} &= \mathbf{w^1} - \frac{\eta}{|\mathcal{D}_{\text{train}}|} \sum_{(x,y)\in\mathcal{D}_{\text{train}}} \nabla_{\mathbf{w}}\text{Loss}_{\text{hinge}}(x, y, \mathbf{w}) \\ &= [-.5] - \frac{0.1}{2}[0 + 1] = [-.55] \end{aligned}$$