Problem Set 6

Problems 1-5 correspond to "A simple linear classifier"

Problem 1

3/3 points (graded)

A linear classifier on $\mathbb{R}^2$ is specified by $w = (-1, 3)$ and $b = -6$.

a) At what point does the decision boundary intersect the x_1-axis? (Just give the x_1-intercept, a real number.)

-6

✓ **Answer:** -6

-6

b) At what point does the decision boundary intersect the x_2-axis? (Just give the x_2-intercept.)

2

✓ **Answer:** 2

2

c) What label, 1 or -1, is assigned to the point $(1, 1)$?

-1

✓ **Answer:** -1

-1

Hint (1 of 2): Recall that the decision boundary consists of all points x for which $w \cdot x + b = 0$.

Hint (2 of 2): The label assigned to any point is the sign of $w \cdot x + b$.

Next Hint

Submit

Answers are displayed within the problem

Problem 2

1/1 point (graded)

A particular line in $\mathbb{R}^2$ passes through the points $(0,1)$ and $(2,0)$ and is specified by equation $w \cdot x + b = 0$, where $b = -2$ and $w \in \mathbb{R}^2$. What is w?

○ $w = (0,1)$

○ $w = (0,2)$

● $w = (1,2)$

○ $w = (2,-1)$

✓

Explanation

The formula for the line is $w_1 x_1 + w_2 x_2 - 2 = 0$ and we need to figure out $w = (w_1, w_2)$. Plugging in the point $(0,1)$, we get $w_2 = 2$. Plugging in $(2,0)$, we get $w_1 = 1$.

Submit

Answers are displayed within the problem

Problem 3

1/1 point (graded)

The Perceptron algorithm makes an update whenever it encounters a data point (x, y) that is "misclassified" by the current w, b. What does this mean, precisely? Choose the best option from this list.

○ $y\,(w \cdot x + b) = 0$

○ $y\,(w \cdot x + b) < 0$

● $y\,(w \cdot x + b) \leq 0$

○ $y\,(w \cdot x + b) > 0$

✓

Explanation

A point (x, y) is *correctly* classified if and only if y equals the sign of $w \cdot x + b$, that is, if and only if $y\,(w \cdot x + b) > 0$.

Submit

Answers are displayed within the problem

Problem 4

1/1 point (graded)

A particular data set of n points is randomly permuted and then the Perceptron algorithm is run on it, repeatedly cycling through the points until convergence. It converges after k updates. Which of the following must be true? Select all that apply.

□ $n \geq k$

☑ If this process were repeated with a different random permutation, then the algorithm would again converge.

□ If this process were repeated with a different random permutation, then the algorithm would again make k updates before convergence.

☑ The data is linearly separable.

✓

Explanation

The Perceptron algorithm will converge if and only if the data is linearly separable; and this is true regardless of what random permutation is chosen. Hence the second and fourth options are correct.
For the first option, it is not necessarily true that $n \geq k$ since the algorithm might make multiple updates on the same point. For the third, we saw in lecture how different random permutations can lead to different numbers of updates and different final classifiers.

Submit

ℹ Answers are displayed within the problem

Problem 5

1/1 point (graded)

The Perceptron algorithm is run on a data set, and converges after performing $p + q$ updates. Of these updates, p are on data points whose label is -1 and q are on data points whose label is $+1$. What is the final value of parameter b?

● $q - p$

○ $p + q$

○ $p - q$

○ q

✓

Explanation

Parameter b starts off at zero. On any update, it is incremented if $y = 1$ and decremented if $y = -1$. Therefore its final value is: (the number of points with label 1) minus (the number of points with label -1).

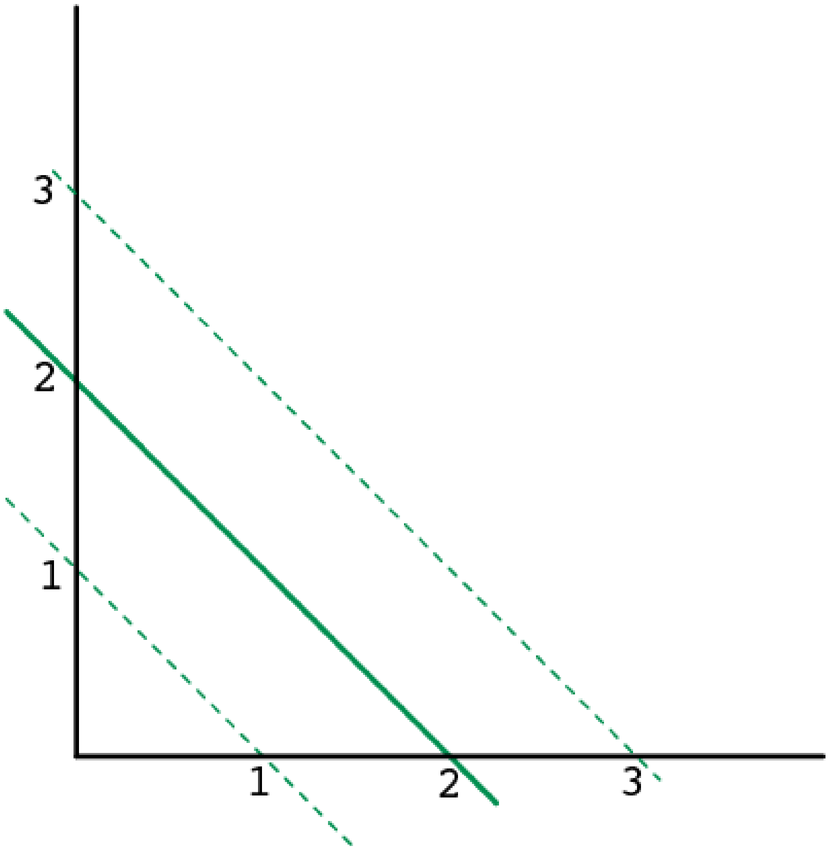
Submit

Answers are displayed within the problem

Problems 6-8 correspond to "Support vector machines I"

Problem 6

1/1 point (graded)
The figure below shows a two-dimensional linear separator $w \cdot x + b = 0$, along with the parallel lines $w \cdot x + b = -1$ and $w \cdot x + b = 1$.



What is the margin of this classifier?

● A number between 0.5 and 1.

○ 1.

○ A number between 1 and 2.

○ 2.

○ A number greater than 2.

✓

Explanation
Using the hint, we can picture a right triangle whose hypotenuse (longest edge) is 1 and whose shorter edges both have length m, the margin. From Pythagoras' theorem, we then have $2m^2 = 1$, so $m = 1/\sqrt{2}$.

? **Hint (1 of 1):** Note that these lines are at a 45-degree angle. Therefore the margin is the leg of an isoceles right triangle whose hypotenuse is 1.

Next Hint

Submit

Answers are displayed within the problem

Problem 7

5/5 points (graded)
A support vector machine classifier is learned for a data set in $\mathbb{R}^2$. It is given by $w = (3, 4)$ and $b = -12$.

a) What is the x_1-intercept of the decision boundary?

4

✓ **Answer:** 4

4

b) What is the x_2-intercept of the decision boundary?

3

✓ **Answer:** 3

3

c) What is the margin of this classifier?

1/5

✓ **Answer:** .2

$\frac{1}{5}$

d) It turns out that the data set has two distinct support vectors of the form $(1, ?)$. What are they?

(give the missing x_2 coordinates for the support vectors with the smaller x_2 value first)

2

✓ **Answer:** 2

2

10/4

✓ **Answer:** 2.5

$\frac{10}{4}$

? **Hint (1 of 4):** It might be helpful to plot this line on a piece of graph paper.

Next Hint

Hint (2 of 4): The decision boundary is $3x_1 + 4x_2 - 12 = 0$. The margin-boundary on the positive side is $3x_1 + 4x_2 - 12 = 1$. The margin-boundary on the negative side is $3x_1 + 4x_2 - 12 = -1$.

Hint (3 of 4): Recall from lecture that the margin is given by the formula $1/\|w\|$. In this case, w is a two-dimensional vector. What is its norm?

Hint (4 of 4): For part (d), one of the support vectors will lie on the positive margin-boundary and one will lie on the negative margin-boundary. Plug into the equations for each of these margins.

Submit

Answers are displayed within the problem

Problem 8

4/4 points (graded)

Consider the following small data set in $\mathbb{R}^2$:

Points $(1,2), (2,1), (2,3), (3,2)$ have label -1.

Points $(4,5), (5,4), (5,6), (6,5)$ have label $+1$.

Now, suppose (hard margin) SVM is run on this data.

a) What is the x_1-intercept of the decision boundary?

7

✓ **Answer:** 7

7

b) What is the x_2-intercept of the decision boundary?

7 ✓ **Answer:** 7

7

c) What is w?

○ $w = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$

○ $w = \begin{pmatrix} -1 \\ -1 \end{pmatrix}$

○ $w = \begin{pmatrix} 2 \\ 1/2 \end{pmatrix}$

● $w = \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix}$

✓

d) What is b?

-3.5 ✓ **Answer:** -7/2

-3.5

? **Hint (1 of 1):** Start by plotting the eight points on a piece of graph paper. At that point, the decision boundary should become clear.

Next Hint

Submit

ℹ Answers are displayed within the problem

Problems 9-12 correspond to "Support vector machines II"

Problem 9

4/4 points (graded)

Here is the optimization problem for the soft-margin SVM.

$$\begin{gathered}\min_{w\in\mathbb{R}^d, b\in\mathbb{R}, \xi\in\mathbb{R}^n} \quad \|w\|^2 + C\sum_{i=1}^{n}\xi_i \\ \text{s.t.:}\quad y^{(i)}(w\cdot x^{(i)} + b) \geq 1 - \xi_i \quad \text{for all } i = 1, 2, \ldots, n \\ \xi \geq 0\end{gathered}$$

a) How many slack variables are there? The answer should be a function of n and/or d.

n

✓ **Answer:** n

b) What setting of C will recapture the hard-margin SVM?

○ Very small C

● Very large C

○ There is no value of C that will do this

✓

Answer

Correct: Larger C imposes a heavier penalty on slack.

c) As C is increased, what happens to the margin of the linear classifier that is returned?

○ The margin gets larger.

● The margin gets smaller.

○ The margin is unchanged.

○ The way in which the margin changes is unpredictable.

✓

Answer

Correct:
As C grows, the optimization problem places more emphasis on classifying the training data correctly and less on having a big margin.

d) Suppose we have a data set that is linearly separable and we use it to train both a hard-margin SVM (w_H, b_H) and a soft-margin SVM (w_S, b_S). Which of the following statements is true? Select all that necessarily apply.

□ Both linear classifiers have zero training error.

☑ $\|w_H\| \geq \|w_S\|$

□ $\|w_H\| \leq \|w_S\|$

☑ The margin achieved by (w_H, b_H) is at most the margin achieved by (w_S, b_S).

✓

Explanation

Part (d) is a direct consequence of part (c). We can think of the hard-margin SVM as having been obtained by a larger value of C and thus we would expect it to have a smaller margin. Since the margin of w equals $1/\|w\|$, the second option also holds.

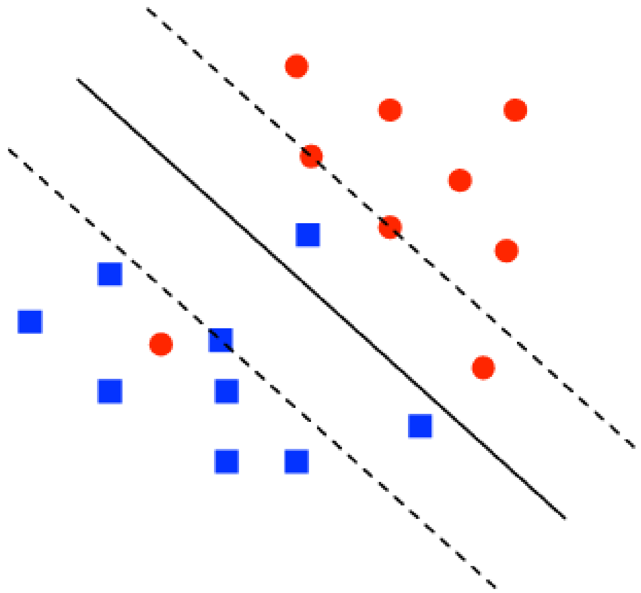
Submit

Answers are displayed within the problem

Problem 10

4/4 points (graded)

The picture below shows the decision boundary obtained upon running soft-margin SVM on a small data set of blue squares and red circles.



a) How many support vectors are there?

7

✓ **Answer:** 7

7

b) What is the largest slack value on a red (circle-shaped) point, roughly?

2.25

✓ **Answer:** [2.1, 2.6]

2.25

c) What is the largest slack value on a blue (square-shaped) point, roughly?

1.5

✓ **Answer:** [1.1, 1.5]

1.5

d) Suppose the factor C in the soft-margin SVM optimization problem were increased. Would you expect the margin to `increase` or `decrease`?

decrease ⌄ ✓ **Answer:** decrease

? **Hint (1 of 3):** The support vectors consist of: positive points on the positive margin-boundary; negative points on the negative margin-boundary; and any point on which slack is used.

Hint (2 of 3): If it is unclear how to obtain the slack values, try reviewing the lecture on soft-margin SVM.

Hint (3 of 3): The effect of increasing C is to place less emphasis on having a large margin and more emphasis on reducing slack.

Next Hint

Submit

Answers are displayed within the problem

Problem 11

1/1 point (graded)

Would it ever make sense to use the soft-margin SVM on a linearly separable data set? Select all that apply.

☐ No, unless you are unsure whether the data is linearly separable.

☑ Yes, because it may lead to a larger margin and better generalization.

☐ No, because it might fail to perfectly separate the training set.

✓

? **Hint (1 of 1):** The soft-margin SVM is the version that is typically used in practice, regardless of whether the data is linearly separable or not.

Next Hint

Submit

Answers are displayed within the problem

Problem 12

1/1 point (graded)

The soft-margin SVM involves a constant C that needs to be set. Which of the following is an appropriate way of setting it? Select all that apply.

□ The output of the SVM is not very sensitive to the choice of C, so it doesn't really matter how C is set.

□ Try various settings, and pick the one that yields the smallest training error.

☑ Try various settings, and pick the one that yields the smallest cross-validation error.

□ Try various settings, and pick the one that yields the largest margin.

✓

Submit

Answers are displayed within the problem

Problems 13-16 correspond to "Duality"

Problem 13

1/1 point (graded)

The dual form of the Perceptron algorithm is run on a data set of four points $x \in \mathbb{R}^2$ with labels $y \in \{-1, 1\}$. The very first update takes place on the first data point, $x^{(1)} = (3, 2)$, which has label -1. What are the values of α and b right after this first update?

○ $\alpha = (-1, 0, 0, 0)\,,\ b = 1$

○ $\alpha = (1, 0, 0, 0)\,,\ b = 1$

● $\alpha = (1, 0, 0, 0)\,,\ b = -1$

○ $\alpha = (0, 0, 0, 0)\,,\ b = -1$

✓

Explanation

The dual form of the Perceptron encodes w using a vector α that has a coordinate for each data point. When an update is made on a data point, that point's entry in α is incremented. Thus the effect of doing just one update, on the first data point, is to produce $\alpha = (1, 0, 0, 0)$. And b is either incremented or decremented on each update, depending on whether the label of the point is positive or negative. In this case, it is negative.

Submit

Answers are displayed within the problem

Problem 14

4/4 points (graded)

The dual form of the Perceptron algorithm is used to learn a binary classifier, based on n training points. It converges after k updates, and returns a vector α. For each of the following statements, indicate whether it is necessarily true or possibly false.

a) Each α_i is either 0 or 1.

possibly false ✓ **Answer:** possibly false

b) $\sum_i \alpha_i = k$.

necessarily true ✓ **Answer:** necessarily true

c) α has at most k nonzero coordinates.

necessarily true ⌄ ✓ **Answer:** necessarily true

d) The training data must be linearly separable.

necessarily true ⌄ ✓ **Answer:** necessarily true

Explanation
First of all, the Perceptron algorithm converges if and only if the data is linearly separable. The vector $\alpha = (\alpha_1, \ldots, \alpha_n)$ is set so that α_j is the number of updates made on the jth data point. Thus immediately implies that parts (b) and (c) are true. There could be multiple updates on any given point; thus part (a) is possibly false.

Submit

ℹ Answers are displayed within the problem

Problem 15

1/1 point (graded)
The dual form of the hard-margin SVM returns a vector α. Which data points $x^{(i)}$ are the support vectors in this solution?

○ Those with $\alpha_i = 0$

● Those with $\alpha_i > 0$

○ Those with $\alpha_i \geq 0$

○ The support vectors cannot be determined simply by looking at α

✓

Explanation
The support vectors are *by definition* those that contribute to w, that is, those with $\alpha_i > 0$.

Submit

Answers are displayed within the problem

Problem 16

1/2 points (graded)

Consider the primal and dual forms of the soft-margin SVM for binary classification. Suppose they are used on a training set of n points, where each point is d-dimensional.

a) How many real-valued variables are there in the primal optimization problem? (Don't use spaces in your expression.)

b) How many real-valued variables are there in the dual optimization problem?

n ✓

Hint (1 of 2): Recall that the primal variables are w, b, ξ. What are their dimensions?

Next Hint

Hint (2 of 2): The dual variables are just α. What is its dimension?

Submit