CS 189 Introduction to Machine Learning
Spring 2023 Jonathan Shewchuk

DIS1

In this discussion, we'll review linear classifiers and develop some intuition for the hard-margin support vector machine (SVM) optimization problem:

$$\min_{w,\alpha} \ \|w\|^2 \ \text{ subject to } \ y_i(X_i \cdot w + \alpha) \geq 1, \ \ \forall i \in \{1, \ldots, n\}.$$

1 Linear Decision Rules

A *decision rule* is a function $r : \mathbb{R}^d \to \pm 1$ that maps a feature vector (test point) to $+1$ ("in class") or -1 ("not in class"). Many classifiers compute a *decision function*, f, which is also known as a *predictor function* or *discriminant function*. The decision rule for the classifier is then defined as

$$r(x) = \begin{cases} +1 & \text{if } f(x) \geq 0 \\ -1 & \text{otherwise} \end{cases}.$$

The *decision boundary* is the boundary chosen by the classifier to separate items in different classes. For a decision function, f, the decision boundary is

$$\{x \in \mathbb{R}^d : f(x) = 0\}.$$

For a d-dimensional feature space, linear decision functions have the form

$$f(x) = x \cdot w + \alpha,$$

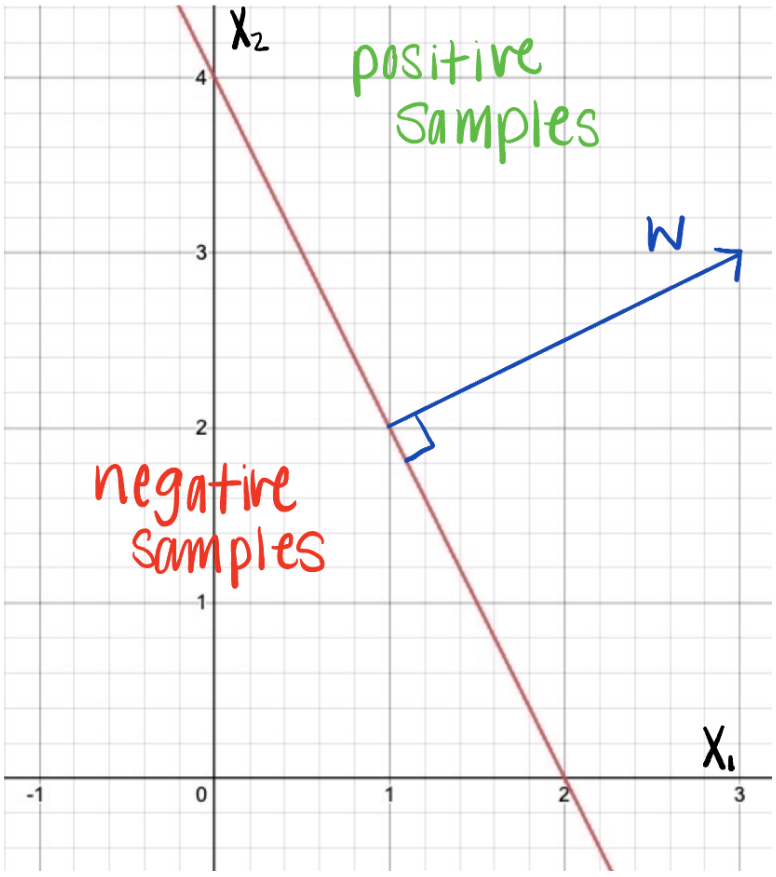
where $w \in \mathbb{R}^d$ is a weight vector and $\alpha \in \mathbb{R}$ is a bias term. For this linear decision function, the decision boundary is the hyperplane

$$\mathcal{H} = \{x \in \mathbb{R}^d : x \cdot w + \alpha = 0\}.$$

(a) Draw a figure depicting the hyperplane $\mathcal{H} = \{x \in \mathbb{R}^2 : x \cdot w + \alpha = 0\}$ with $w = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$ and $\alpha = -4$. Include in your figure the vector w, drawn relative to $\mathcal{H}$.

(b) Indicate in your figure the region in which data points would be classified as $+1$ (in class). Do the same for data points that would be classified as -1 (not in class).

Solution:


X_2
positive samples
w
negative samples
X_1
4
3
2
1
0
-1
1
2
3

2 Maximum Margin Classifier

Consider a data set of n d-dimensional sample points, $\{X_1, \ldots, X_n\}$. Each sample point, $X_i \in \mathbb{R}^d$, has a corresponding label, y_i, indicating to which class that point belongs. For now, we will assume that there are only two classes and that every point is either in the given class ($y_i = 1$) or not in the class ($y_i = -1$). Consider the linear decision boundary defined by the hyperplane

$$\mathcal{H} = \{x \in \mathbb{R}^d : x \cdot w + \alpha = 0\}.$$

The *maximum margin classifier* maximizes the distance from the linear decision boundary to the closest training point on either side of the boundary, while correctly classifying all training points.

(a) An in-class sample point is correctly classified if it is on the positive side of the decision boundary, and an out-of-class sample is correctly classified if it is on the negative side. Write a set of n constraints to ensure that all n points are correctly classified.

Solution: We can begin by writing the set of constraints

$$\begin{cases} X_i \cdot w + \alpha \geq 1 & \text{if } y_i = 1 \\ X_i \cdot w + \alpha \leq -1 & \text{if } y_i = -1 \end{cases} \quad \text{for } i = 1, \ldots, n.$$

Note that we could replace ± 1 in the inequalities above with $\pm c$, where c is any non-negative constant. We can combine these two sets of constraints into the n constraints

$$y_i(X_i \cdot w + \alpha) \geq 1 \text{ for } i = 1, \ldots, n.$$

(b) The maximum margin classifier aims to maximize the distance from the training points to the decision boundary. Derive the distance from a point X_i to the hyperplane $\mathcal{H}$.

Solution: Let $\hat{X}_i$ denote the projection of point X_i onto the hyperplane, $\mathcal{H}$. We know that $\hat{X}_i$ must lie on the hyperplane, so $\hat{X}_i \cdot w + \alpha = 0$. We also know that the vector $(X_i - \hat{X}_i)$ must be perpendicular to the hyperplane. Because w is the normal vector for $\mathcal{H}$, $(X_i - \hat{X}_i)$ must lie in the direction of w. Therefore, there exists some scalar $\eta \in \mathbb{R}$ such that $(X_i - \hat{X}_i) = \eta w$. With these two observations, we can determine the value of η as follows.

$$(X_i - \hat{X}_i) \cdot w = (\eta w) \cdot w$$

$$X_i \cdot w - \hat{X}_i \cdot w = \eta(w \cdot w)$$

$$X_i \cdot w + \alpha = \eta \|w\|^2$$

$$\eta = \frac{X_i \cdot w + \alpha}{\|w\|^2}.$$

The distance from the point X_i to the hyperplane $\mathcal{H}$ is thus

$$d_i = \|X_i - \hat{X}_i\| = \|\eta w\| = |\eta| \, \|w\| = \left| \frac{X_i \cdot w + \alpha}{\|w\|^2} \right| \|w\| = \frac{|X_i \cdot w + \alpha|}{\|w\|}.$$

(c) Assuming all the points are correctly classified, write an inequality that relates the distance of sample point X_i to the hyperplane $\mathcal{H}$ in terms of only the normal vector w.

Solution: From part (a), we know that if the points are correctly classified,

$$y_i(X_i \cdot w + \alpha) \geq 1 \text{ for } i = 1, \ldots, n.$$

Because y_i is either 1 or -1, these inequalities imply that

$$|X_i \cdot w + \alpha| \geq 1 \text{ for } i = 1, \ldots, n.$$

Therefore, we obtain the following inequality for the distance of X_i to the hyperplane.

$$d_i = \frac{|X_i \cdot w + \alpha|}{\|w\|} \geq \frac{1}{\|w\|}.$$

(d) For the maximum margin classifier, the training points closest to the decision boundary on either side of the boundary are referred to as *support vectors*. What is the distance from any support vector to the decision boundary?

Solution: A support vector X_+ in the given class (i.e. a positive sample) must satisfy

$$X_+ \cdot w + \alpha = 1.$$

A support vector X_- not in the given class (i.e. a negative sample) must satisfy

$$X_- \cdot w + \alpha = -1.$$

Therefore, every support vector X_i must satisfy

$$|X_i \cdot w + \alpha| = 1.$$

Hence the distance from the closest point on either side of the decision boundary is

$$d_i = \frac{|X_i \cdot w + \alpha|}{\|w\|} = \frac{1}{\|w\|}.$$

(e) Using the previous parts, write an optimization problem for the maximum margin classifier.

Solution: The distance of any point to the hyperplane can never be less than $\frac{1}{\|w\|}$. Therefore, to maximize the margin, we want to maximize $\frac{1}{\|w\|}$, which is equivalent to minimizing $\|w\|$. This leads us to the maximum margin classification problem

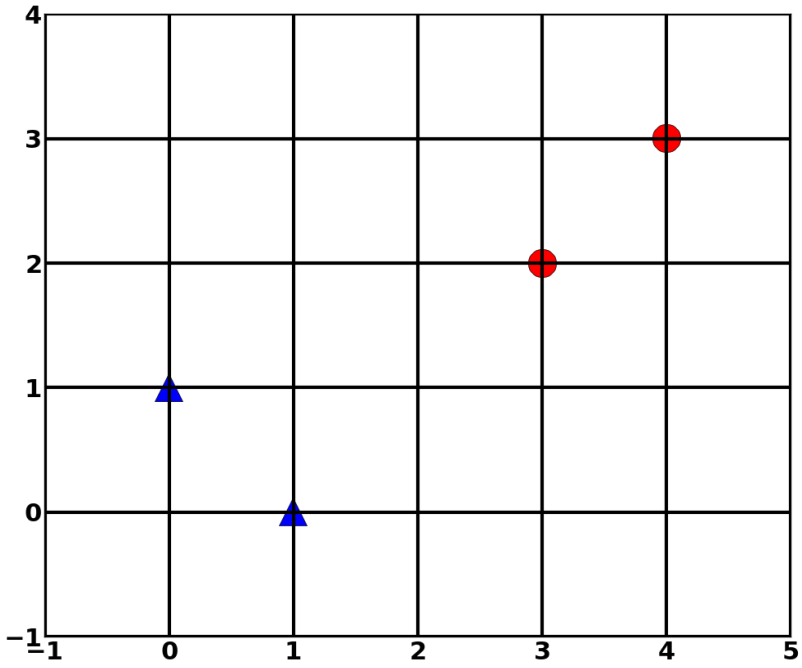
$$\min_{w,\alpha} \ \|w\| \ \text{ subject to } \ y_i(X_i \cdot w + \alpha) \geq 1, \ \ \forall i \in \{1, \ldots, n\}.$$

We prefer a smooth objective function, and $\|w\|$ is not smooth. (It is pointed at $w = 0$.) So we equivalently express this problem as

$$\min_{w,\alpha} \ \|w\|^2 \ \text{ subject to } \ y_i(X_i \cdot w + \alpha) \geq 1, \ \ \forall i \in \{1, \ldots, n\}.$$

3 Hard-Margin SVM by Hand

You are given the sample points shown in the figure below. The blue triangles are positive samples (in the given class), and the red circles are negative examples (not in the given class).



Find (by hand) the equation of the hyperplane $\mathcal{H} = \{x \in \mathbb{R}^2 : x \cdot w + \alpha = 0\}$ that a hard-margin SVM classifier would learn. Draw the decision boundary and its margins.

Solution:

The maximum margin hyperplane bisects and is perpendicular to the line segment joining the closest points in the convex hulls of the two sets. In this example, the convex hulls are the line segment joining the negative points and the line segment joining the positive points, so the two points that are closest are $(3, 2)$ and $(1, 0)$. Therefore, the hyperplane will pass through the point $(2, 1)$ with a slope of -1. The equation of this line is $x_1 + x_2 = 3$.

To determine the weight vector, w, and bias term, α, for the desired hyperplane, notice first that the point $(2, 1)$ must lie on the hyperplane. Therefore,

$$2w_1 + 1w_2 + \alpha = 0.$$

Next, we can also see that $(0, 1)$ is a positive support vector. Therefore,

$$0w_1 + 1w_2 + \alpha = 1.$$

Finally, notice that $(3, 2)$ is a negative support vector. Therefore,

$$3w_1 + 2w_2 + \alpha = -1.$$

Now we have three equations and three unknowns. Solving this systems of linear equations, we obtain the parameters of the linear boundary determined by the SVM.

$$w = \left[-\frac{1}{2}, -\frac{1}{2}\right]^T \quad \text{and} \quad \alpha = \frac{3}{2}.$$