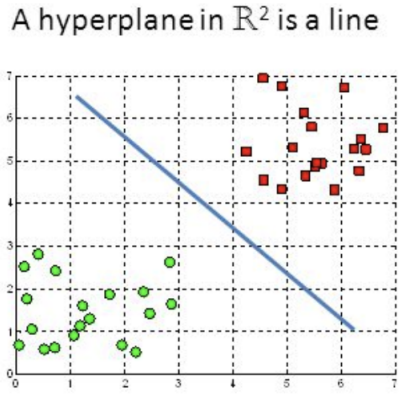
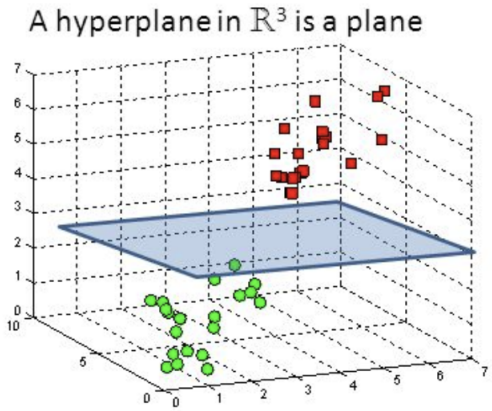
CS 181 Spring 2021 Section 5

Margin-Based Classification, SVMs

1 Motivation

In the past, with binary linear classifiers, we found a hyperplane that separated the data (or in cases when this was not possible separated a large amount of the data).

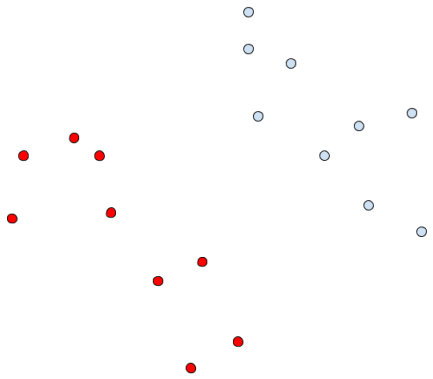




Source: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

However, if data is linearly separable, there may be many boundaries that work, so how can we know which is best?

Concept Question: Draw the "best" decision boundary for the data below (using your own definition of best). What would be a worse decision boundary? Why?



As your intuition told you above, the idea for Support Vector Machines is that, for all the linear hyperplanes that exist, we want one that will create the largest distance, or "margin", with the training data. At a high level, we define the marign as the minimum distance between a point and our boundary. Larger margins tend to improve generalization error.

Now we'll put this into math. To find a mathematical formula for the margin, we consider a hyperplane of the form

$$\mathbf{w}^\top \mathbf{x} + w_0 = 0$$

For two points $\mathbf{x}_1$ and $\mathbf{x}_2$ on the hyperplane, consider the projection with $\mathbf{w}$:

$$\mathbf{w}^\top(\mathbf{x}_1 - \mathbf{x}_2) = \mathbf{w}^\top \mathbf{x}_1 - \mathbf{w}^\top \mathbf{x}_2 = -w_0 - (-w_0) = 0$$

Therefore, $\mathbf{w}$ is orthogonal to the hyperplane. So to get the distance from a hyperplane and an arbitrary example $\mathbf{x}$, we just need the length in the direction of $\mathbf{w}$ between the point and the hyperplane (this gives us the perpendicular distance between $\mathbf{x}$ and the hyperplane–why do we want the perpendicular distance?). We let r signify the distance between a point and the hyperplane. Then $\mathbf{x}_p$ is the projection of the point onto the hyperplane, so that we can decompose a point $\mathbf{x}$ as

$$\mathbf{x}_p + r\frac{\mathbf{w}}{||\mathbf{w}||} = \mathbf{x}$$

Left multiply by $\mathbf{w}^\top$:

$$\mathbf{w}^\top \mathbf{x}_p + r\frac{\mathbf{w}^\top \mathbf{w}}{||\mathbf{w}||} = \mathbf{w}^\top \mathbf{x} \Rightarrow r = \frac{\mathbf{w}^\top \mathbf{x} + w_0}{||\mathbf{w}||}$$

Scalar r then gives the signed, normalized distance between a point and the hyperplane. For correctly classified data, we have $y_i = +1$ when this distance is positive and $y_i = -1$ when it is negative. Based on this, we can obtain a positive distance for both kinds of examples by multiplying by y_i (and y_i will not change the magnitude). We define the margin of the dataset as the minimum such distance over all examples:

$$\min_i \frac{y_i(\mathbf{w}^\top \mathbf{x}_i + w_0)}{||\mathbf{w}||}$$

Concept Question: Before optimizing a model, it's important to make sure you understand how it works. Give a new data point $\mathbf{x}$, what would a SVM predict (assume you know w^*)?
We now want to figure out how to find the optimal $\mathbf{w}$. As we discussed in motivating this model, we want the $\mathbf{w}$ and w_0 that maximize the margin:

$$\arg\max_{\mathbf{w}, w_0} \frac{1}{||\mathbf{w}||} \min_i y_i(\mathbf{w}^\top \mathbf{x}_i + w_0)$$

When discussing margin classifiers, we consider both hard and soft margin classifiers. In the hard-margin training problem, we know that the data is linearly separable and therefore any margin (including the optimal) must be greater than 0 (in the soft-margin problem, we do not make this assumption; this case is more complex so we deal with it later):

$$\min_i \frac{y_i(\mathbf{w}^\top \mathbf{x}_i + w_0)}{||\mathbf{w}||} > 0$$

We can observe that $\mathbf{w}$ and w_0 are invariant to changes of scale. Because of this, it is without loss of generality to impose $\min_i \frac{y_i(\mathbf{w}^\top \mathbf{x}_i + w_0)}{||\mathbf{w}||} > 1$ (we prove this in the exercises of these notes). This lets us write the optimization problem as:

$$\arg\max_{\mathbf{w}, w_0} \frac{1}{||\mathbf{w}||} \quad \text{s.t. } \forall i \; y_i(\mathbf{w}^\top \mathbf{x}_i + w_0) \geq 1$$

We can invert $\mathbf{w}$ to change the max to a min:

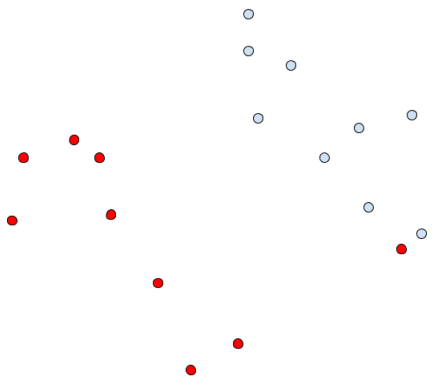
$$\arg\min_{\mathbf{w}, w_0} \frac{1}{2}||\mathbf{w}||^2 \quad \text{s.t. } \forall i \; y_i(\mathbf{w}^\top \mathbf{x}_i + w_0) \geq 1$$

Informally, this is the same because the constraint is binding for the examples closest to the decision boundary. For these examples we have $y_i(\mathbf{w}^\top \mathbf{x}_i + w_0) = 1$. The distance on these examples is $1/||\mathbf{w}||$, and is maximized by minimizing $||\mathbf{w}||^2$. Mathematically, the min of $\frac{1}{||\mathbf{w}||}$ must be the max of $\frac{1}{2}||\mathbf{w}||^2$ since otherwise, we could just pick the proposed better maximum of $\mathbf{w}^*$ and find something eve smaller than our preexisting min. We discuss exactly how to solve this problem in the third part of these notes, but first we discuss the soft margin problem.

2 Soft Margin Formulation

For the hard margin formulation, we have been assuming that the data is linearly separable. However, this is not always true, and even if the data is linearly separable, it may not be best to find a separating hyperplane. In optimizing generalization error, there is a tradeoff between the size of the margin and the number of mistakes on the training data.

Concept Question: Let's illustrate the tradeoff between size of the margin and number of mistakes on training data. Assuming that you have sufficient basis transformations to draw any boundary (since a boundary linear on the transformed data may be nonlinear on the original dimensions), draw what you feel would be the most generalizable SVM classifier for the below data?



For the soft margin formulation, we introduce a slack variable $\xi_i \geq 0$ for each i to relax the constraints on each example.

$$\xi_i \begin{cases} = 0 & \text{if correctly classified} \\ \in (0, 1] & \text{correctly classified but inside margin region} \\ > 1 & \text{if incorrectly classified} \end{cases}$$

We can now rewrite the training problem for a soft margin formulation to be

$$\arg\min_{\mathbf{w}, w_0} \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{n} \xi_i$$

$$\text{s.t. } \forall i \; y_i(\mathbf{w}^\top \mathbf{x}_i + w_0) \geq 1 - \xi_i$$

$$\xi_i \geq 0$$

We add a regularization parameter C, that controls how much we penalize violating the hard margin constraints. A large C penalizes these violations and thus "respects" the data closely and has small regularization. A small C does not penalize the sum of slack variables as heavily, relaxing the constraint. This is increasing the regularization.

3 Practice Problems (cover some but not all at section)

1. **Removing Support Vectors and Retraining (Berkeley, Fall '11)**

   Suppose that we train two SVMs, the first containing all of the training data and the second trained on a data set constructed by removing some of the support vectors from the first training set. How does the size of the optimal margin change between the first and second training data? What is a downside to doing this?

2. **Proof that margin is invariant to scalar multiplication**

   In reformulating our max margin

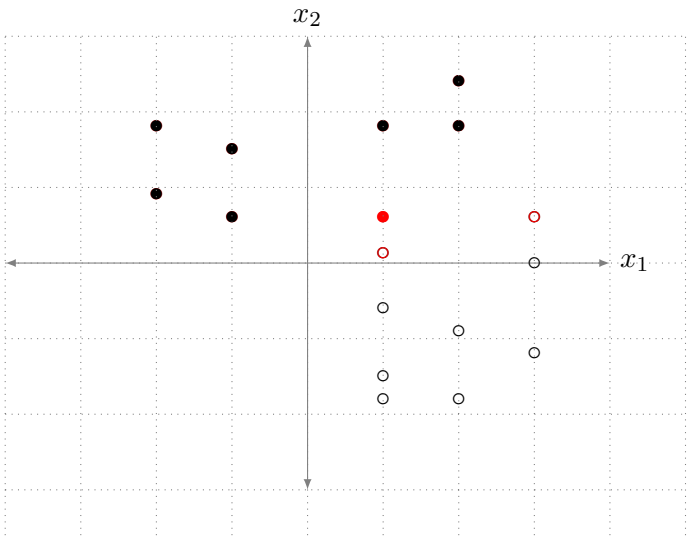
   $$\frac{y_i(\mathbf{w}^\top \mathbf{x}_i + w_0)}{||\mathbf{w}||}$$

   training problem, we use the fact that the margin is invariant to multiplying $(\mathbf{w}, w_0)$ by any $\beta > 0$. Show that this property is true.

3. **Draw Margin Boundary**

i) For the first example, you can assume the hard margin formulation. Draw the decision boundary as well as the two margin boundaries given that the red examples represent points with minimum margin.
ii) For the second example, you can assume the soft margin formulation and that all points are correctly classified with the optimal decision boundary. The decision boundary is already given. Draw the two margin boundaries given the information about the ξ_n slack values for the different examples shown. Assume the other points are outside of the margin region.

i)



ii)