1. Support Vector Machines

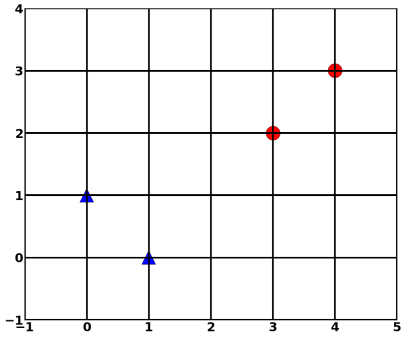
a. We typically frame an SVM problem as trying to *maximize* the margin. Explain intuitively why a bigger margin will result in a model that will generalize better, or perform better in practice.

> **Solution:** One intuition is that if points are closer to the border, we are less certain about their class. Thus, it would make sense to create a boundary where our "certainty" is highest about all the training set points.
>
> Another intuition involves thinking about the process that generated the data we are working with. Since it's a noisy process, if we drew a boundary close to one of our training points of some class, it's very possible that a point of the same class will be generated across the boundary, resulting in an incorrect classification. Therefore it makes sense to make the boundary as far away from our training points as possible.

b. Show that the width of an SVM slab with linearly separable data is $\frac{2}{\|w\|}$.

> **Solution:** The width of the margin is defined by the points that lie on it, also called support vectors. Let's say we have a point, $\vec{x}'$, which is a support vector. The distance between $\vec{x}'$ and the separating hyperplane can be calculated by projecting the vector starting at the plane and ending at $\vec{x}$ onto the plane's unit normal vector. The equation of the plane is $\vec{w}^T\vec{x} + b = 0$. Since $\vec{w}$ by definition is orthogonal to the hyperplane, we want to project $\vec{x}' - \vec{x}$ onto the unit vector normal to the hyperplane, $\frac{\vec{w}}{\|\vec{w}\|}$.
>
> $$\frac{\vec{w}^T}{\|\vec{w}\|}(\vec{x}' - \vec{x}) = \frac{1}{\|\vec{w}\|}(\vec{w}^T\vec{x}' - \vec{w}^T\vec{x}) = \frac{1}{\|\vec{w}\|}(\vec{w}^T\vec{x}' + b - \vec{w}^T\vec{x} - b)$$
>
> Since we set $\vec{w}^T\vec{x}' + b = 1$ (or -1) and by definition, $\vec{w}^T\vec{x} + b = 0$, this quantity just turns into $\frac{1}{\|\vec{w}\|}$, or $\vec{1}\|\vec{w}\|$, so the distance is the absolute value, $\frac{1}{\|\vec{w}\|}$.
>
> Since the margin is half of the slab, we double it to get the full width of $\frac{2}{\|\vec{w}\|}$.

c. You're presented with the following set of data (triangle = +1, circle = -1):



Find the equation (by hand) of the hyperplane $\vec{w}^T x + b = 0$ that would be used by an SVM classifier. Which points are support vectors?

Solution: The equation of the hyperplane will pass through point $(2, 1)$, with a slope of -1. The equation of this line is $x_1 + x_2 = 3$. We know that from this form, $w_1 = w_2$. We also know that the at the support vectors, $w^T x + b = \pm 1$. This gives us the equations:

$$1w_1 + 0w_2 + b = 1$$

$$3w_1 + 2w_2 + b = -1$$

Solving this system of equations, we get $\vec{w} = [-\frac{1}{2}, -\frac{1}{2}]^T$ and $b = \frac{3}{2}$.
The support vectors are $(1, 0), (0, 1)$, and $(3, 2)$.

2. What's the difference between the perceptron algorithm and the hard-margin SVM algorithm?

Solution: The perceptron algorithm seeks to find any hyperplane that will separate the training data. The hard-margin SVM will try to find the separating hyperplane that maximizes the distance (margin) from the hyperplane to the closest training point on either side.

3. What's the difference between the hard-margin and the soft-margin SVM? How does the hyperparameter C affect the solution to the soft-margin SVM?

Solution: The hard-margin SVM will only work if the data is separable. C determines how strictly the margin constraint can be violated. A high value of C places a high penalty on violating the margin constraint, whereas a lower value allows more points to violate the margin constraint. The hard-margin SVM places an infinitely high penalty on violating the margin constraints, i.e. $C = \infty$.

4. Matrix Calculus (Linear Regression)

Let $X \in \mathbb{R}^{n \times d}$ be a data matrix and $y \in \mathbb{R}^n$ be the corresponding vector of labels. What is the weight vector $\theta \in \mathbb{R}^d$ that minimizes the quadratic loss between the predicted labels $X\theta$ and the actual labels y?

Solution: The loss function is

$$\begin{aligned} J(\theta) &= ||X\theta - y||^2 \\ &= (X\theta - y)^\top (X\theta - y) \\ &= \theta^\top X^\top X\theta - y^\top X\theta - \theta^\top X^\top y + y^\top y \\ &= \theta^\top X^\top X\theta - 2y^\top X\theta + y^\top y \end{aligned}$$

To minimize the loss function, we take the derivative with respect to θ and set it to 0.

$$\begin{aligned} \frac{\partial J}{\partial \theta} &= 2X^\top X\theta - 2X^\top y = 0 \\ \hat{\theta} &= (X^\top X)^{-1} X^\top y \end{aligned}$$