SVM Margins

When talking about the margin, it is important to differentiate between the *functional margin* and the *geometric margin*. The functional margin is just the margin from each point to the boundary that "comes from linear algebra", and is defined by

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

$\mathbf{w}^\top \mathbf{x}_i$ is just the projection onto $\mathbf{w}$, w_0 is what moves the origin (the 0) of the number line around (i.e. it moves the boundary away from the origin, which is equivalent to moving the 0 of the number line created by the projection), and y_i is just a hack that allows us to obtain a positive margin if the data point is in the negative region (thus $\mathbf{w}^\top \mathbf{x}_i + w_0$ would be less than 0 but $y_i = -1$; remember that $\mathcal{Y} = \{1, -1\}$).

The issue is that this margin can be arbitrarily large due to the projection because we can scale $\mathbf{w}$ (the equality $\mathbf{w}^\top \mathbf{x}_i = \|\mathbf{w}\| \hat{\mathbf{w}}^\top \mathbf{x}_i$ makes it obvious) and w_0 (which we could also scale by $\|\mathbf{w}\|$ to keep things proportional). Thus, we must place some constraint on the size of $\mathbf{w}$. This brings us to the geometric margin. A detailed derivation is provided here, but definition is similar to the functional margin except we divide $\mathbf{w}^\top \mathbf{x}_i + w_0$ by $\|\mathbf{w}\|$, thus

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

This margin actually corresponds to the margin in the geometry of the data, and is invariant to the size of $\mathbf{w}$ (as is clear from the division by its norm). Because of this, we will want to maximize the geometric margin. At this point, it is useful to note that the functional or geometric margin of a *point* are as defined previously, and the functional or geometric margin of the *classifier* or *data set* is twice as much as the corresponding margin for a support vector (since we have one for each support vector on each region). Thus, our goal is to maximize the geometric margin of the data set.

This goal must be subject to the constraints of having all data end up in the appropriate regions. For this, we can use the functional margin, and say that $y_i(\mathbf{w}^\top \mathbf{x}_i + w_0) \geq 1$. In other words, we want the functional margin of all points to be at least one (remember that it will be a positive number if we get the right regions due to the y_i hack), and for support vectors, we want this inequality to be be an equality (we want it to equal 1 for at least one point - although by consequence I believe this would make it equal to 1 for all support vectors). The reason why we choose 1 for the margin is because it would make the geometric margin equal to $\frac{2}{\|\mathbf{w}\|}$. This follows from the definition of the geometric margin, since we have that, for support vectors, $y_i(\mathbf{w}^\top \mathbf{x}_i + w_0) = 1$, so

$$2 \cdot y_i \frac{\mathbf{w}^\top \mathbf{x}_i + w_0}{\|\mathbf{w}\|} = 2 \cdot \frac{y_i \mathbf{w}^\top \mathbf{x}_i + w_0}{\|\mathbf{w}\|} = 2 \cdot \frac{1}{\|\mathbf{w}\|} = \frac{2}{\|\mathbf{w}\|}$$

Note that this also means the functional margin equals 2. Having the 2 makes things easier to derive later on (I believe since we will eventually want to minimize $\frac{1}{2}\mathbf{w}^\top \mathbf{w}$).