CS 6140: Machine Learning **Spring 2015**
College of Computer and Information Science
Northeastern University
Lecture TBD **TBD, 26**
Instructor: Bilal Ahmed **Scribe: Bingyu Wang, Virgil Pavlu**

Support Vector Machines I

1 What's SVM

The original SVM algorithm was invented by Vladimir N. Vapnik[1] and the current standard incarnation (soft margin) was proposed by Corinna Cortes[2] and Vapnik in 1993 and published in 1995.

A support vector machine(SVM) constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier.[3] In this notes, we will explain the intuition and then get the primal problem, and how to translate the primal problem to dual problem. We will apply kernel trick and SMO algorithms to solve the dual problem and get the hyperplane we want to separate the dataset. Give general idea about SVM and introduce the goal of this notes, what kind of problems and knowledge will be covered by this node.

In this note, one single SVM model is for two labels classification, whose label is $y \in \{-1, 1\}$. And the hyperplane we want to find to separate the two classes data set is h, for which classifier, we use parameters w, b and we write our classifier as

$$h_{w,b}(x) = g(w^T x + b)$$

Here, $g(z) = 1$ if $z \geq 0$, and $g(z) = -1$ otherwise.

2 Margins

Following Andrew Ng[4], we will start the by talking about margins, which can give us the "confidence" of our predictions.

Consider logistic regression, where the probability $p(y = 1|x; w)$ is modeled by $h_w(x) = g(w^T x)$.We would then predict "1" on an input x if and only if $h_w(x) \geq 0.5$, or equivalently, if and only if $w^T x \geq 0$. Consider a positive training example ($y = 1$). The larger $w^T x$ is, the larger also is $h_w(x) = p(y = 1|x; w, b)$, and thus also the higher our degree of "confidence" that the label is 1. Thus informally we can think of our prediction as being a very confident one that $y = 1$ if $w^T x \gg 0$. Similarly, we think of logistic regression as making a very confident prediction of $y = 0$, if $w^T x \ll 0$. Given a training set, again informally it seems that we'd have found a good fit to the training data if we can find w so that $w^T x_i \gg 0$ whenever $y_i = 1$, and $w^T x_i \ll 0$ whenever $y_i = 0$, since this would reflect a very confident (and correct) set of classifications for all the training examples. This seems to be a nice goal to aim for, and we?ll soon formalize this idea using the notion of functional margins.

For a different type of intuition, consider the following Figure 1, in which x's represent positive training examples, o's denote negative training examples, a decision boundary (this is the line given by the equation $w^T x = 0$, and is also called the **separating hyperplane**) is also shown, and three points have also been labeled A, B and C.

Notice that the point A is very far from the decision boundary. If we are asked to make a prediction for the value of y at A, it seems we should be quite confident that $y = 1$ there. Conversely, the point C is

[1] http://en.wikipedia.org/wiki/Vladimir_Vapnik

[2] http://en.wikipedia.org/wiki/Corinna_Cortes

[3] http://en.wikipedia.org/wiki/Support_vector_machine

[4] CS229 Lecture notes, Part V Support Vector Machines

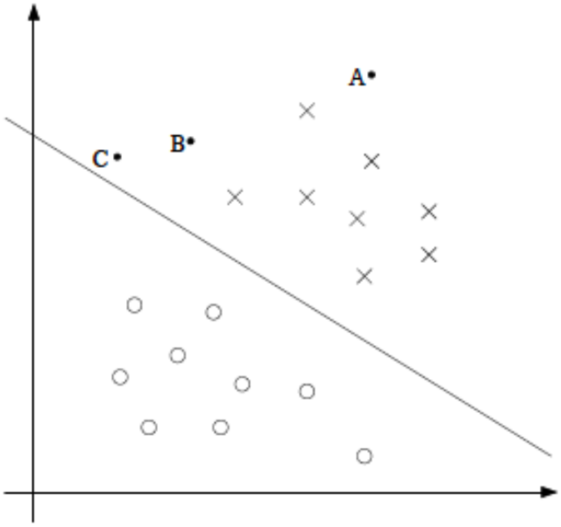


Figure 1: Confident Example, linearly separable.

very close to the decision boundary, and while it's on the side of the decision boundary on which we would predict $y = 1$, it seems likely that just a small change to the decision boundary could easily have caused our prediction to be $y = 0$. Hence, we're much more confident about our prediction at A than at C. The point B lies in-between these two cases, and more broadly, we see that if a point is far from the separating hyperplane, then we may be significantly more confident in our predictions. Again, informally we think it'd be nice if, given a training set, we manage to find a decision boundary that allows us to make all correct and confident (meaning far from the decision boundary) predictions on the training examples.

In another word, if we could find a decision boundary, who can give us a larger margin, it will be better than the one give us a smaller margin. From the following Figure 2, we can tell that the black decision boundary is better than the green decision boundary, because the black one gives us a larger margin than the green one.

2.1 Functional and Geometric Margins

Lets now formalize the margin intuition into notions of the functional and geometric margins. Given a training example (x_i, y_i), we define the functional margin of (w, b) with respect to the training example

$$\hat{\gamma}_i = y_i(w^T x + b)$$

Note that if $y_i = 1$, then for the functional margin to be large (i.e., for our prediction to be confident and correct), we need $w^T x + b$ to be a large positive number. Conversely, if $y_i = -1$, then for the functional margin to be large, we need $w^T x + b$ to be a large negative number. Moreover, if $y_i(w^T x + b) > 0$, then our prediction on this example (x_i, y_i) is correct. Hence, a large functional margin represents a confident and a correct prediction.

Given a training set $S = \{(x_i, y_i); i = 1, 2, \ldots, m\}$, we also define the function margin of (w, b) with respect to S to be the smallest of the functional margins of the individual training examples. Denoted by $\hat{\gamma}$, this can therefore be written:

$$\hat{\gamma} = \min_{i=1,\ldots,m} \hat{\gamma}_i$$

Functional margins can represent a confident and a correct prediction. The larger functional margins, the classifier better. However, by scaling w, b, we can make the functional margin arbitrarily large without really changing anything meaningful. Typically for a linear classifier, the final prediction is made by applying the sign function g to the linear score:

$$g(z) = \begin{cases} 1 & \text{if } z \geq 0 \\ -1 & \text{if } z < 0 \end{cases}$$

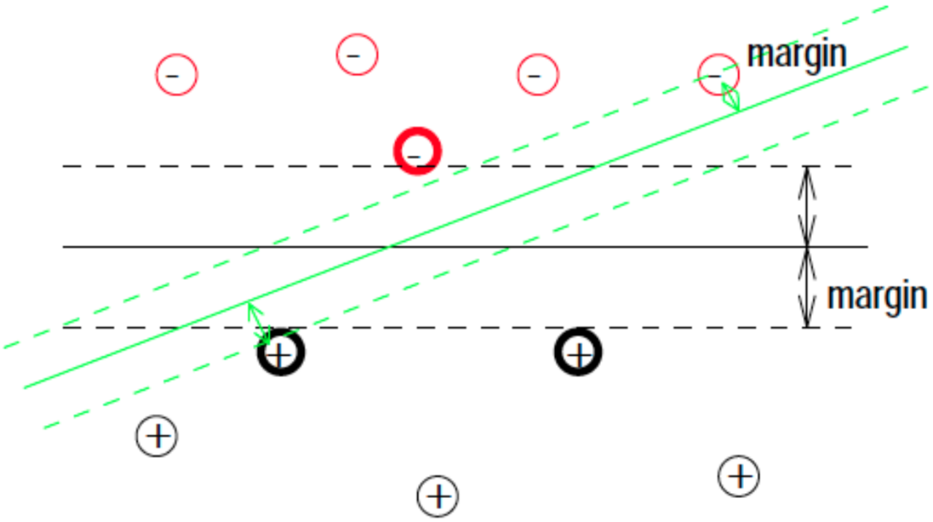
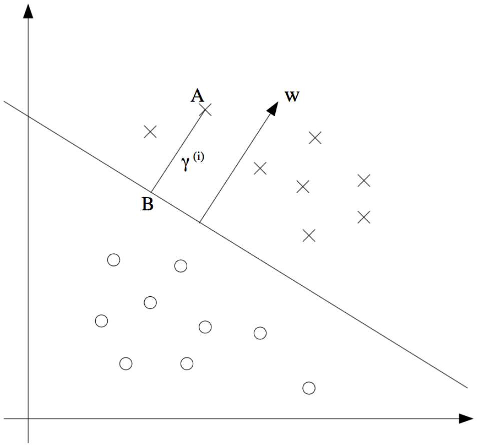


Figure 2: Margin Example. The black separating plane is better than the green one, because it has larger margins (sits more "in the middle"). A mechanical analogy: if the separating plane is free to rotate but constrained to be separator, when the points start pushing force towards the plane, the plane will settle in an equilibrium "middle" position - thats where the black separator is.

We note for any scalar c we can replace w with cw and b with cb, and have that $g(w^Tx+b) = g(cw^Tx+cb)$, this would not change the prediction $h_{w,b}(x) = g$ at all. I.e., g, and hence also $h_{w,b}(x)$, depends only one the sign, but not on the magnitude of $w^T + b$. However, replacing (w, b) with (cw, cb) also results in multiplying our functional margin by a factor of c. Thus, it seems that by exploiting our freedom to scale w and b, we can make the functional margin arbitrarily large without really changing anything meaningful. We can make a reference decision on scale, and will choose the scale such that minimum functional margin is $y(w^Tx+b) = 1$.

2.1.1 Geometric Margins

In order to solve the above problem, we introduce the geometric margins. Consider the picture similar to Figure 1 as below:



The decision boundary corresponding to (w, b) is shown, along with the vector w. Note that w is orthogonal to the separating hyperplane. Consider the point at A, which represents the input x_i of some training example with label $y_i = 1$. Its distance to the decision boundary, γ_i , is given by the line segment AB.How can we find the value of γ_i? Well, $w/||w||$ is a unit-length vector pointing in the same direction as w. Since A represents x_i, we therefore find that the point B is given by $x_i - \gamma_i * w/||w||$. But this point lies on the decision boundary, and all points x on the decision boundary satisfy the equation $w^T x + b = 0$. Hence,

$$w^T (x_i - \gamma_i * \frac{w}{||w||}) + b = 0$$

Solving for γ_i, we have

$$\gamma_i = \frac{w^T x_i + b}{||w||} = (\frac{w}{||w||})^T x_i + \frac{b}{||w||}$$

This was worked out for the case of a positive training example at A in the figure, where being on the "positive" side of the decision boundary is good. More generally, we define the geometric margin of (w, b) with respect to a training example (x_i, y_i) to be

$$\gamma_i = y_i((\frac{w}{||w||})^T x_i + \frac{b}{||w||})$$

Note that if $||w|| = 1$, then the functional margin equals the geometric margin—this thus gives us a way of relating these two different notions of margin. Also, the geometric margin is invariant to rescaling of the parameters; i.e., if we replace w with $2w$ and b with $2b$, then the geometric margin does not change. This will in fact come in handy later. Specifically, because of this invariance to the scaling of the parameters, when trying to fit w and b to training data, we can impose an arbitrary scaling constraint on w without changing anything important; for instance, we can demand that $||w|| = 1$, or $|w_1| = 5$, or $|w_1 + b| + |w_2| = 2$, and any of these can be satisfied simply by rescaling w and b.Finally, given a training set $S = (x_i, y_i); i = 1, ..., m$, we also define the geometric margin of (w, b) with respect to S to be the smallest of the geometric margins on the individual training examples:

$$\gamma = \min_{i=1,...,m} \gamma_i$$