Linear Classifiers

Carlos C. Rodríguez
http://omega.albany.edu:8008/

October 12, 2004

Classification with Hyperplanes

Assume we have observed n examples of labeled data $(X_1, Y_1), \ldots, (X_n, Y_n)$ but now let us take the labels $Y \in \{1, -1\}$. This convention simplifies some of the formulas below and it is the standard for support vector machines.

We'll be interested in classification rules of the form,

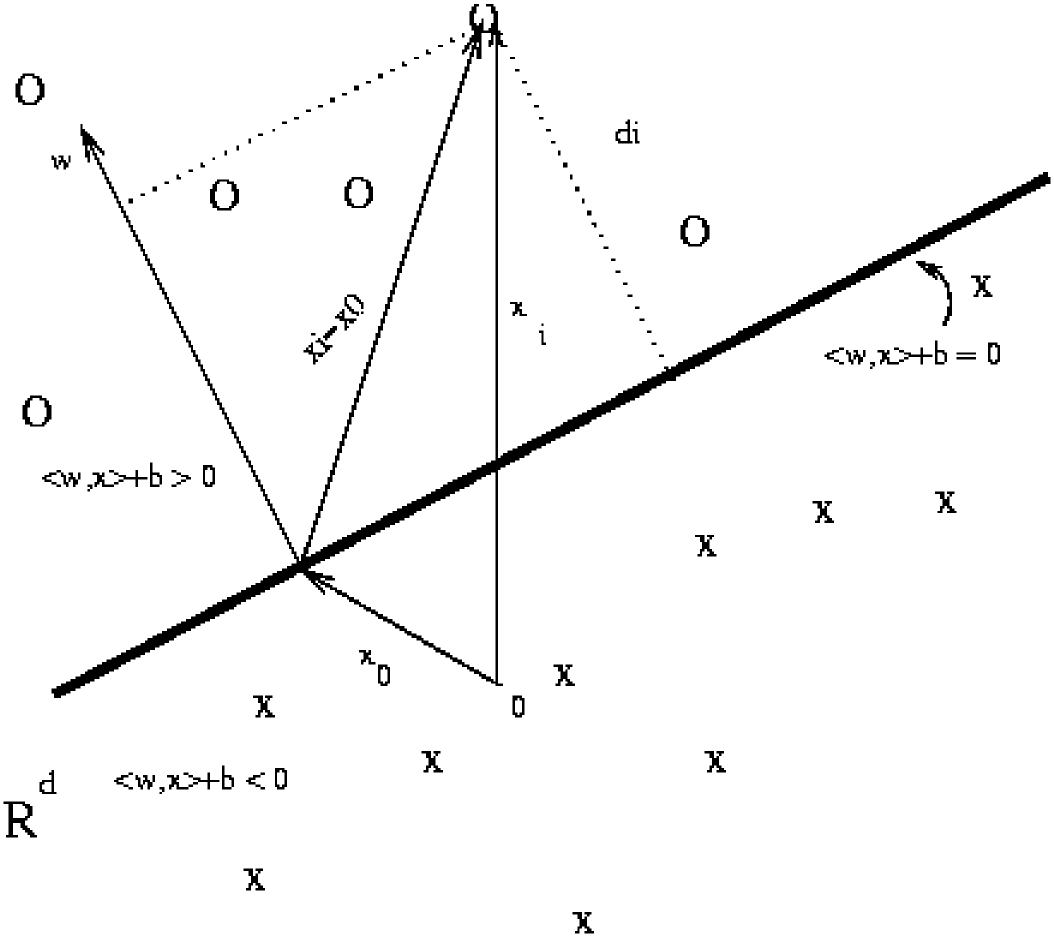
$$g(x) = \text{sgn }\ (< w, x > +b)$$

where $<, >$ denotes the (euclidean) inner product in R^d and $w \in R^d$ and $b \in R$ parametrize all the rules of this type. These rules are called linear. The class $\mathcal{G}$ of linear classifiers is about the simplest possible way of separating two kinds of objects in R^d. A classifier in $\mathcal{G}$ assigns the label $+1$ or -1 to a data point $x \in R^d$ depending on weather we arrive to x from the plane by going in the direction of the normal w or its oposite $-w$.

Recall that the set of vectors,

$$\{x :< w, x > +b = 0\}$$

correspond to the hyperplane perpendicular to w that goes through every point x_0 such that, $< w, x_0 >= -b$. For example the following picture shows such an hyperplane that perfectly separates the O's (with label $+1$) from the X's (with label -1).



The hyperplane with parameters (w, b) is the same as the hyperplane with parameters (cw, cb) for any non-zero scalar c. To fix the scale we say that a plane is in *canonical position* with respect to a set of points $\mathcal{X} = \{x_1, \ldots, x_n\}$ if,

$$\min_{i \leq n} | < w, x_i > +b| = 1.$$

Clearly, the perpendicular distance from x_i to the hyperplane with parameters (w, b) is the magnitud of the projection of $(x_i - x_0)$ onto the unit normal direction $w/|w|$. It is given by,

$$d_i = \frac{| < w, (x_i - x_0) > |}{|w|} = \frac{| < w, x_i > +b|}{|w|}$$

as it can be seen from the above picture. Thus, if we assume that the plane is in canononical position w.r.t. $\{x_i : i \leq n\}$ then,

$$\rho = \min_{i \leq n} d_i = \frac{1}{|w|}.$$

This minimun distance ρ is known as the *margin* between a separating hyperplane and the points w.r.t. which that plane is in general position.

Maximum Margin Classifiers

When there is a hyperplane with parameters (w, b) that separates the points $\{x_i : y_i = 1, i \leq n\}$ from the points $\{x_i : y_i = -1, i \leq n\}$ then, if we take the plane in canonical position w.r.t. $\mathcal{X} = \{x_1, \ldots, x_n\}$ we have,

$$y_i(< w, x_i > +b) \geq 1 \text{ for all } i \leq n \ .$$

There could be many such planes but it is geometrically obvious that the one farthest from the points should be prefered over all the others. The intuition behind this is the fact that the farther the separating boundary is from the observations the more likely it seems for future data to end up being correctly classified by this plane. In other words, we expect better generalization power for boundaries with larger margin, i.e., for planes with small $|w|$ (recall that $\rho = 1/|w|$. This intuition is confirmed by the following theorem.

Theorem Consider hyperplanes through the origin in canonical position w.r.t. $\mathcal{X} = \{x_1, \ldots, x_n\}$. Then, the set of linear classifiers, $g_w(x) = \text{sgn} < w, x >$ with $|w| \leq \Lambda$ has VC dimension,

$$V \leq \min\{R^2\Lambda^2, R^d\Lambda^d\}$$

where R is the radius of the smallest sphere around the origin containing $\mathcal{X}$

Before proving it notice that in fact the length of w controls the VC dimension so that the larger the margin the larger the generalization power as expected.

Proof: We need to show both, $V \leq R^d\Lambda^d$ and $V \leq R^2\Lambda^2$. The first inequality follows from the fact that the maximum number of spheres of radius $1/\Lambda$ that still fit inside a sphere of radius R is at most $\text{vol}(R)/\text{vol}(1/\Lambda) = (R\Lambda)^d$, where $\text{vol}(r) = C_d r^d$ is the volume of the sphere of radius r in R^d. More than this number of points is imposible to shatter with margin of at least $1/\Lambda$. Thus, $V \leq (R\Lambda)^d$. For the other inequality again we show that the number of points that can be shattered with margin $1/\Lambda$ is at most $(R\Lambda)^2$. If n points can be shattered then for all possible choices of $y_i \in \{1, -1\}$ there is an hyperplane with parameter w with $|w| \leq \Lambda$, in canonical position w.r.t. $\{x_1, \ldots, x_n\}$ separating $+1$'s from -1's, i.e., satisfying,

$$y_i < w, x_i > \geq 1 \text{ for all } i = 1, \ldots, n.$$

But this together with Cauchy-Schwartz inequality and the fact that $|w| \leq \Lambda$ imply,

$$
\begin{aligned}
n &\leq |< w, \sum_{i=1}^{n} y_i x_i > | \leq |w|| \sum_{i=1}^{n} y_i x_i| \\
&\leq \Lambda | \sum_{1}^{n} y_i x_i |.
\end{aligned}
$$

To bound the sum consider the case when the labels y_i are Rademacher variables, i.e. iid with $P\{y_i = 1\} = 1 - P\{y_i = -1\} = 1/2$. From the fact that $E\{y_i y_j\} = 0$ when $i \neq j$ and that $y_i^2 = 1$ for all $i \leq n$, we obtain

$$
\begin{aligned}
E|\sum y_i x_i|^2 &= \sum_{i=1}^{n} E\{< y_i x_i, \sum y_j x_j >\} \\
&= \sum_{i=1}^{n} \left\{ \sum_{j \neq i} E[< y_i x_i, y_j x_j >] + E[< y_i x_i, y_i x_i >] \right\} \\
&= \sum_{i=1}^{n} |x_i|^2 \leq n R^2.
\end{aligned}
$$

If the bound is true for the expectation when Rademacher variables are used, then there must exist y_i's for which,

$$
|\sum y_i x_i|^2 \leq n R^2.
$$

Squaring the initial inequality and using the above bound, we obtain

$$
n^2 \leq \Lambda^2 n R^2
$$

and the result follows after deviding through by n.●