Lecture 5: Linear Classification

CS 194-10, Fall 2011

Laurent El Ghaoui

EECS Department
UC Berkeley

September 8, 2011

Outline

Binary Classification

Regularization and Robustness

Outline

Data

We are given a *training* data set:

- Feature vectors: data points $x_i \in \mathbf{R}^p$, $i = 1, \ldots, n$.
- Labels: $y_i \in \{-1, 1\}$, $i = 1, \ldots, n$.

Examples:

Feature vectors	Labels
Companies' corporate info	default/no default
Stock price data	price up/down
News data	price up/down
News data	sentiment (positive/negative)
Emails	presence of a keyword
Genetic measures	presence of disease

Linear classification

Using the training data set $\{x_i, y_i\}_{i=1}^n$, our goal is to find a classification rule $\hat{y} = f(x)$ allowing to predict the label $\hat{y}$ of a new data point x.

Linear classification rule: assumes f is a combination of the sign function and a linear (in fact, affine) function:

$$\hat{y} = \mathbf{sign}(w^T x + b),$$

where $w \in \mathbf{R}^p$, $b \in \mathbf{R}$ are given.

The goal of a linear classification algorithm is to find w, b, using the training data.

Separable data

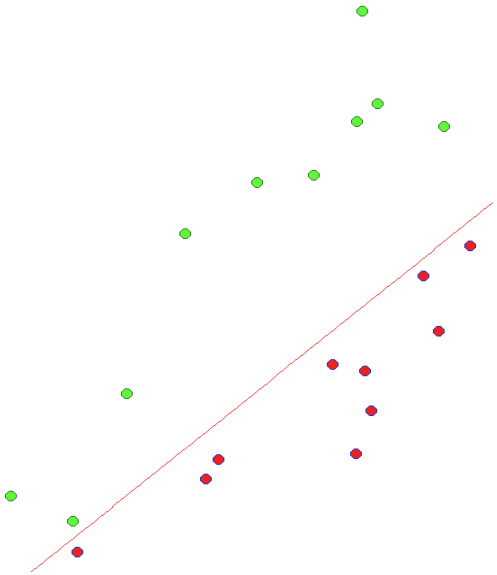
The data is linearly separable if there exist a linear classification rule that makes no error on the training set.

This is a set of linear inequalities constraints on (w, b):

$$y_i(w^T x_i + b) \geq 0, \quad i = 1, \ldots, n.$$

Strict separability corresponds the the same conditions, but with strict inequalities.

Geometry

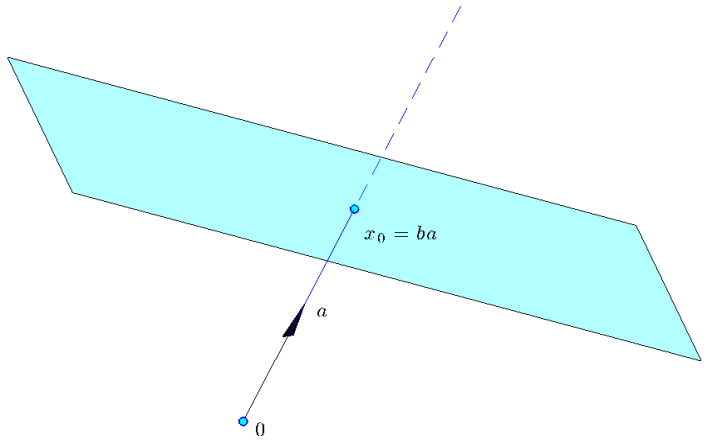


Geometrically: the hyperplane

$$\{x \ : \ w^T x + b = 0\}$$

perfectly separates the positive and negative data points.

Linear algebra flashback: hyperplanes



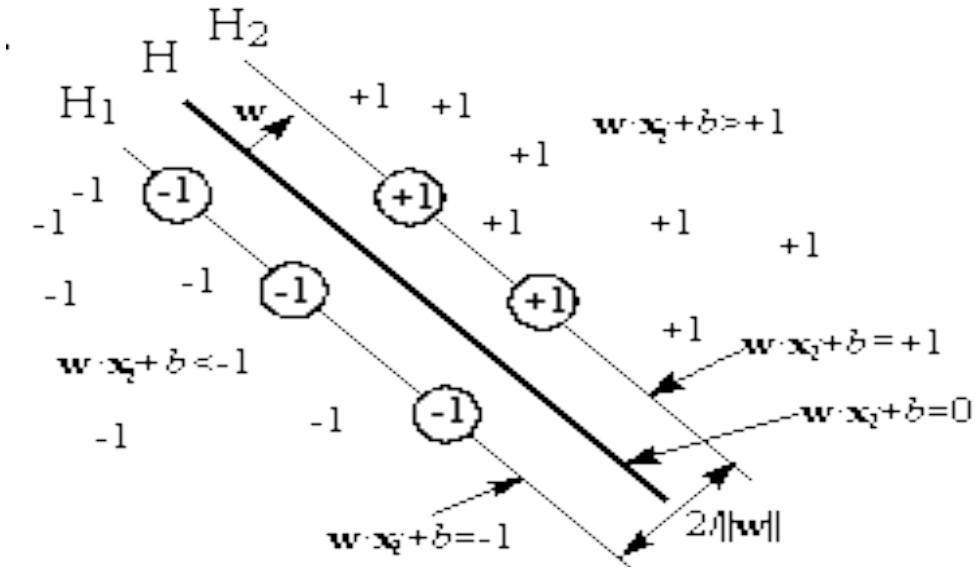
Geometrically, an hyperplane $\mathbf{H} = \{w : a^T x = b\}$, with (WLOG) $\|a\|_2 = 1$, is a translation of the set of vectors orthogonal to a. The direction of the translation is determined by a, and the amount by b.

Geometry (cont'd)

Assuming strict separability, we can always rescale (w, b) and work with

$$y_i(w^T x_i + b) \geq 1, \quad i = 1, \ldots, n.$$

Amounts to make sure that negative (resp. positive) class contained in half-space $w^T x + b \leq -1$ (resp. $w^T x + b \geq 1$).



The distance between the two "± 1" boundaries turns out the be equal to $2/\|w\|_2$. Thus the "margin" $\|w\|_2$ is a measure of how well the hyperplane separates the data apart.

Non-separable data

Separability constraints are homogeneous, so WLOG we can work with

$$y_i(w^T x_i + b) \geq 1, \quad i = 1, \ldots, n.$$

If the above is infeasible, we try to minimize the "slacks"

$$\min_{w,b,s} \sum_{i=1}^{n} s_i \ : \ s \geq 0, \quad y_i(w^T x_i + b) \geq 1 - s_i, \quad i = 1, \ldots, n.$$

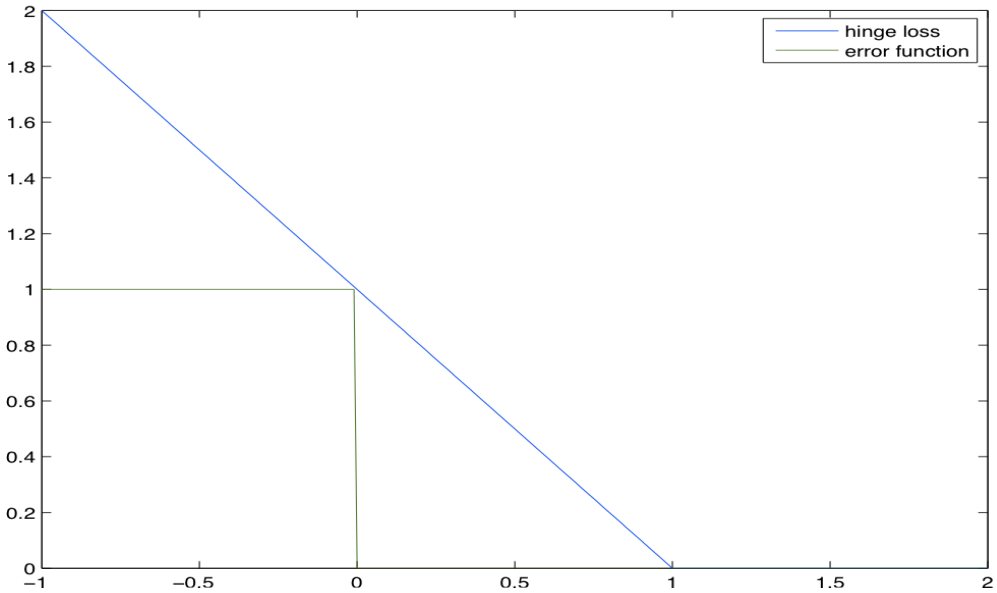
The above can be solved as a "linear programming" problem (in variables w, b, s).

Hinge loss function

The previous LP can be interpreted as minimizing the hinge loss function

$$L(w, b) := \sum_{i=1}^{n} \max(1 - y_i(w^T x_i + b), 0).$$

This serves as an approximation to the number of errors made on the training set:



Outline

Regularization

The solution might not be unique, so we add a regularization term $\|w\|_2^2$:

$$\min_{w,b} \frac{1}{2}\|w\|_2^2 + C \cdot L(w,b)$$

where $C > 0$ allows to trade-off the accuracy on the training set and the prediction error (more on why later). This makes the solution unique.

The above model is called the *Support Vector Machine* . It can be reliably solved using special fast algorithms that exploit its structure.

If C is large, and data is separable, reduces to

$$\min_{w,b} \frac{1}{2}\|w\|_2^2 \ : \ y_i(w^T x_i + b) \geq 1, \quad i = 1, \ldots, n.$$

Robustness interpretation

Return to separable data. The set of constraints

$$y_i(w^T x_i + b) \geq 0, \quad i = 1, \ldots, n,$$

has many possible solutions (w, b).

We will select a solution based on the idea of robustness (to changes in data points).

Maximally robust separating hyperplane

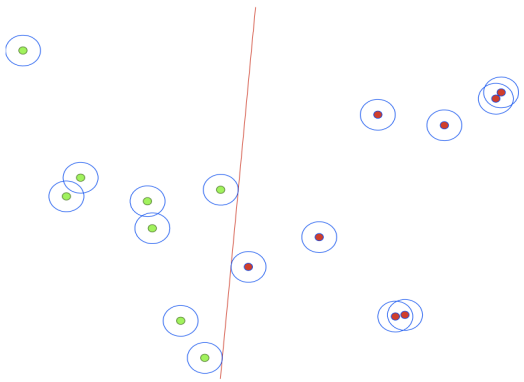
Spherical uncertainty model: assume that the data points are actually unknown, but bounded:

$$x_i \in \mathcal{S}_i := \{\hat{x}_i + u_i \ : \ \|u_i\|_2 \leq \rho\},$$

where $\hat{x}_i$'s are known, $\rho > 0$ is a given measure of uncertainty, and u_i is unknown.

Robust counterpart: we now ask that the separating hyperplane separates the spheres (and not just the points):

$$\forall x_i \in \mathcal{S}_i \ : \ y_i(w^T x_i + b) \geq 0, \quad i = 1, \ldots, n.$$



For separable data we can try to separate spheres around the given points. We'll grow the spheres' radius until sphere separation becomes impossible.

Robust classification

We obtain the equivalent condition

$$y_i(w^T \hat{x}_i + b) \geq \rho \|w\|_2, \quad i = 1, \ldots, n.$$

Now we seek (w, b) which maximize ρ subject to the above.

By homogeneity we can always set $\rho\|w\|_2 = 1$, so that problem reduces to

$$\min_w \ \|w\|_2 \ : \ y_i(w^T \hat{x}_i + b) \geq 1, \quad i = 1, \ldots, n.$$

This is exactly the same problem as the SVM in separable case.

Dual problem

Denote by $\mathcal{C}^+$ (resp. $\mathcal{C}^-$) the set of points x_i with $y_i = +1$ (resp. -1).

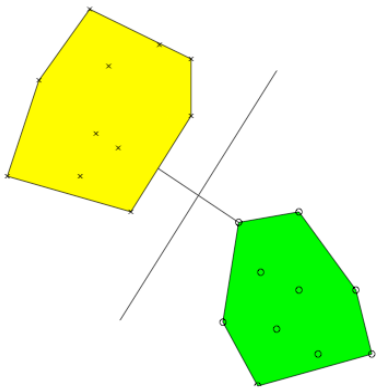
It can be shown that the SVM problem can be expressed as:

$$\min_{x_+, x_-} \|x_+ - x_-\|_2 \ : \ x_+ \in \mathbf{Co}\mathcal{C}^+, \quad x_- \in \mathbf{Co}\mathcal{C}^-,$$

where $\mathbf{Co}\mathcal{C}$ denotes convex hull of set $\mathcal{C}$, that is:

$$\mathbf{Co}\mathcal{C} = \left\{ \sum_{i=1}^{q} \lambda_i x_i \ : \ x_i \in \mathcal{C}, \quad \lambda \geq 0, \quad \sum_{i=1}^{q} \lambda_i = 1 \right\}.$$

Geometry



Dual problem amounts to find the smallest distance between the two classes, each represented by the convex hull of its points. The optimal hyperplane sits at the middle of the line segment joining the two closest points.

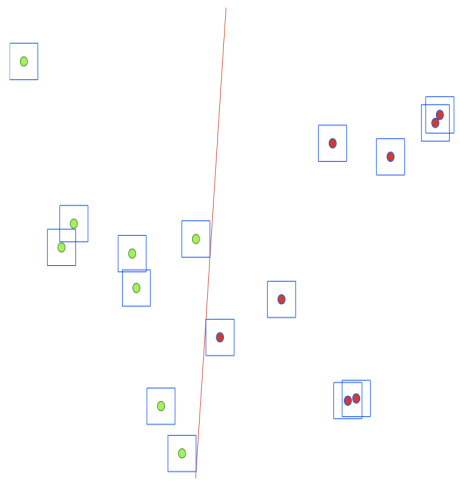
Separating boxes instead of spheres

We can use a box uncertainty model:

$$x_i \in \mathcal{B}_i := \{\hat{x}_i + u_i \ : \ \|u_i\|_\infty \leq \rho\} \, .$$

This leads to

$$\min_w \ \|w\|_1 \ : \ y_i(w^T \hat{x}_i + b) \geq 1, \quad i = 1, \ldots, n.$$



Classifiers found that way tend to be sparse. In 2D, the boundary line tends to be vertical or horizontal.