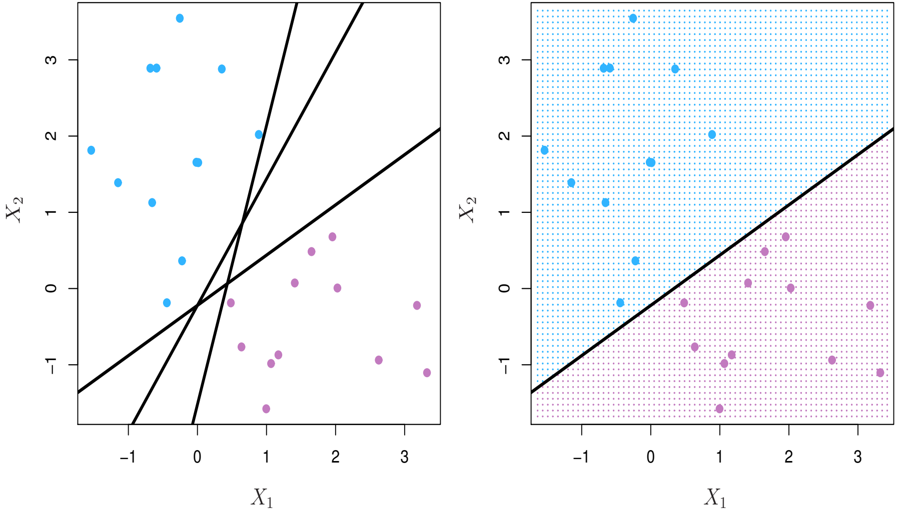
STATS 415: Support Vector Machines

Prof. Liza Levina

Department of Statistics, University of Michigan

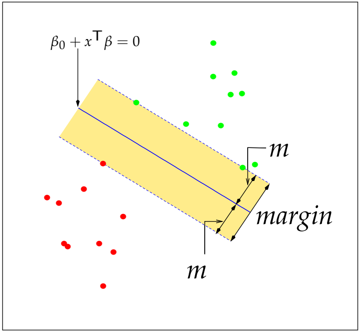
Separating Hyperplanes

- Consider a two class classification problem with two predictors X_1 and X_2
- In 2d, a hyperplane is simply a line
- Suppose that the two classes are "linearly separable", i.e., one can draw a straight line in which all points on one side belong to one class and points on the other side to the other class.
- Then a natural approach is to find the line that gives the biggest separation between the classes, i.e., the points are as far from the line as possible.
- This is the basic idea of a classifier called support vector machine.



X_2
X_1
−1
0
1
2
3

Maximum Margin Classifier



- m is the minimum perpendicular distance between each point and the separating line.
- We find the line which maximizes m.
- This line is called the "optimal separating hyperplane."
- The classification of a point is determined by which side of the line it falls on.

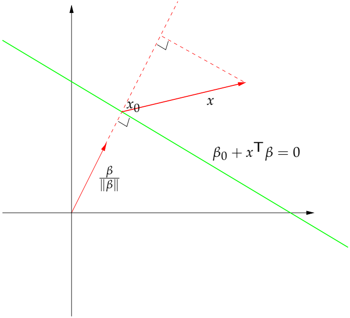
SVM terminology

- Separating hyperplane: a linear boundary between the classes
- Margin: distance from the separating hyperplane to each class (need to maximize)
- Support vectors: data points on the margin boundary

More Than Two Predictors

- This idea works just as well with more than two predictors.
- Three predictors: find the plane that produces the largest separation between the classes.
- More than three predictors: a hyperplane. It becomes hard to visualize a plane but it still exists.

Properties of hyperplanes



- Hyperplane is defined by

$$F = \{x : \beta_0 + x^T\beta = 0\} \ .$$

- The vector β is perpendicular to F
- For any point x_0 in the hyperplane,

$$x_0^T\beta = -\beta_0.$$

- Signed distance from point x to F is

$$\frac{1}{\|\beta\|}(x^T\beta + \beta_0) = \left\langle \frac{\beta}{\|\beta\|}, x - x_0 \right\rangle$$

where x_0 is any point in the plane.

Mathematical Formulation of SVM

- Maximize the minimum distance (Vapnik, 1995)

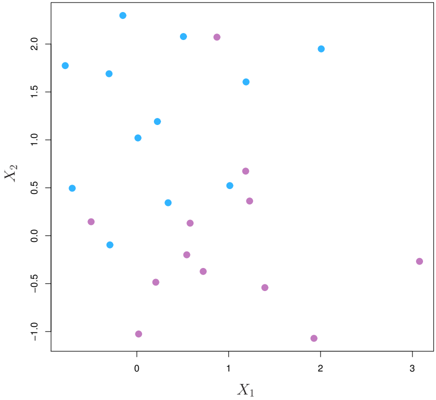
$$\max_{\beta_0, \beta} \quad m$$
$$\text{subject to} \quad \frac{1}{\|\beta\|} y_i(\beta_0 + x_i^\intercal \beta) \geq m \; i = 1, \dots, n$$

- Equivalently (showing this requires knowing optimization theory), solve a quadratic programming problem

$$\min_{\beta_0, \beta} \quad \frac{1}{2} \|\beta\|^2$$
$$\text{subject to} \quad y_i(\beta_0 + x_i^\intercal \beta) \geq 1, \; i = 1, \dots, n$$

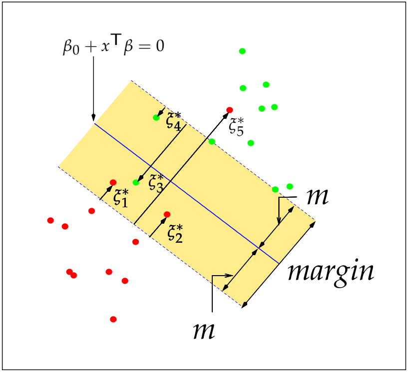
Overlapping Classes

- Of course, it is not always possible to find a hyperplane that perfectly separates two classes.
- That is, for any straight line/plane you can draw, there will always be some points on the wrong side.

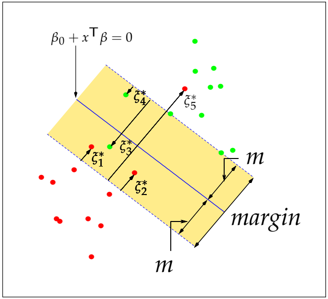


Overlapping Classes

- Then we look for the plane that gives the best separation between the correctly classified points, while keeping the points on the wrong side not too far from the line.
- Let ξ_i represent the amount that the ith point is on the wrong side of the margin (the dashed lines).



Mathematical Formulation



$$\max_{\beta_0,\beta} \quad m$$

$$\text{subject to} \quad \frac{y_i}{\|\beta\|}(\beta_0 + x_i^\mathsf{T}\beta) \geq m(1-\xi_i)$$

$$\xi_i \geq 0, \quad \sum_i \xi_i \leq B$$

- ξ_i: slack variables
- B: tuning parameter

- Points outside the margin, classified correctly: $\xi_i = 0$
- Points inside the margin but classified correctly: $0 < \xi_i < 1$
- Misclassified points: $\xi_i > 1$

Quadratic programming formulation

Equivalently (requires knowing optimization theory)

$$\min_{\beta_0,\beta,\xi_i} \quad \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{n}\xi_i$$
$$\text{subject to} \quad y_i(\beta_0 + x_i^\intercal \beta) \geq 1 - \xi_i,\ \xi_i \geq 0$$

- C is a tuning parameter controlling the amount of “slack”
- Larger $C \to$ smaller margins

Solution

- The solution is expressed in terms of fitted Lagrange multipliers (again from optimization theory), which are just some constants $\hat{\alpha}_i$:

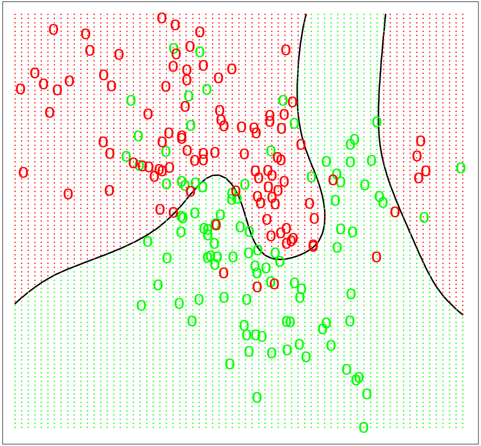
$$\hat{\beta} = \sum_{i=1}^{n} \hat{\alpha}_i y_i x_i$$

- Some of the $\hat{\alpha}_i$ are exactly zero (from optimization theory); the x_i for which $\hat{\alpha}_i \neq 0$ are called support points $\mathscr{S}$. The fitted model is

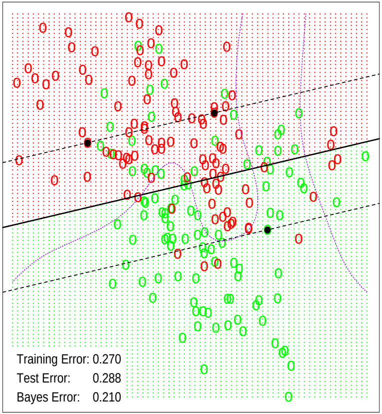
$$\hat{f}(x) = \hat{\beta}_0 + x^\top \hat{\beta} = \hat{\beta}_0 + \sum_{i \in \mathscr{S}} \hat{\alpha}_i y_i \langle x, x_i \rangle$$

- Important consequence: to classify a new point, all we need is its inner products with the support points

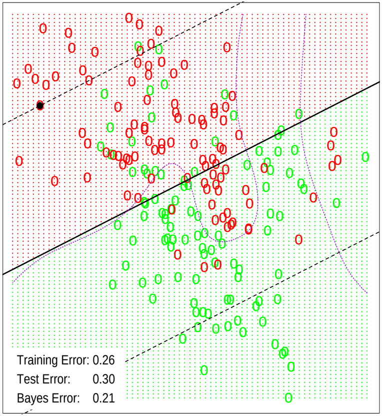
Example



Linear SVMs



$C = 10000$



$C = 0.01$

Why do we need another linear classifier?

- Already have LDA (optimal for Gaussian), logistic regression
- With the same predictors, SVM does not have much advantage
- For example, on the iris data:
  LDA – 3 errors,
  logistic regression – 2 errors,
  SVM – 4 errors
- Question: can we create linear separability?

Main motivation for SVM

- Embed the data in a higher-dimensional space
- For example, given p predictors

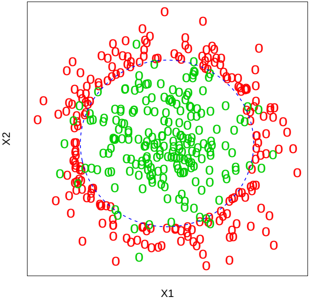
$$X_1, X_2, \cdots, X_p$$

add new variables

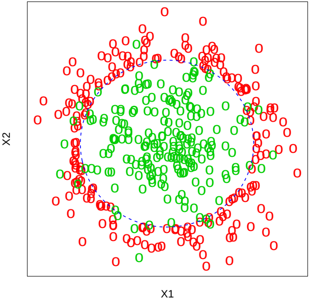
$$X_1^2, X_2^2, \cdots, X_p^2$$

- Then apply a linear method for separating the two classes
- The higher the dimension, the easier it is to find a separating hyperplane

Example: nested spheres data set



Two predictors X_1, X_2. Which methods we know will perform well? Which ones will fail?



Now consider four predictors X_1, X_2, X_1^2, X_2^2. What is a good linear rule?

Nonlinear SVM

- Recall that we extended linear regression to non-linear regression using a basis function, i.e.

$$y = \beta_0 + \beta_1 b_1(x) + \beta_2 b_2(x) + \cdots + \beta_q b_q(x) + \varepsilon$$

- For example, polynomials are a linear rule in the basis 1, x, x^2, etc.
- Same idea: take some basis functions $b_1(x)$, $b_2(x)$, $\cdots$, $b_q(x)$ and find the optimal hyperplane in the space spanned by $b_1(x)$, $b_2(x)$, $\cdots$, $b_q(x)$.
- This approach produces a linear plane in the transformed space but a non-linear decision boundary in the original space.

Example

- Two original predictors, x_1 and x_2.
- Would like the classifier to take the form of 2nd degree polynomial:

$$(ax_1 + bx_2 + c)^2$$

- We choose the following basis: (any constants would work)

$$\begin{aligned}
b_1(x) &= 1 \\
b_2(x) &= \sqrt{2}x_1 \\
b_3(x) &= \sqrt{2}x_2 \\
b_4(x) &= x_1^2 \\
b_5(x) &= x_2^2 \\
b_6(x) &= \sqrt{2}x_1x_2
\end{aligned}$$

Kernels

- The function $h(x)$ is involved only through inner products

$$K(x, x^*) = \langle b(x), b(x^*) \rangle$$

- If we could find a function $K(x, x^*)$ to compute this inner product directly from x and x^*, we would not need to construct the basis vectors $b(x)$ at all.

$$\hat{f}(x) = \hat{\beta}_0 + \sum_{i \in S} \hat{\alpha}_i y_i K(x, x_i)$$

- K is called a kernel function

Back to the 2nd degree polynomial example

If we choose

$$K(x, x') = (1 + \langle x, x' \rangle)^2$$

then

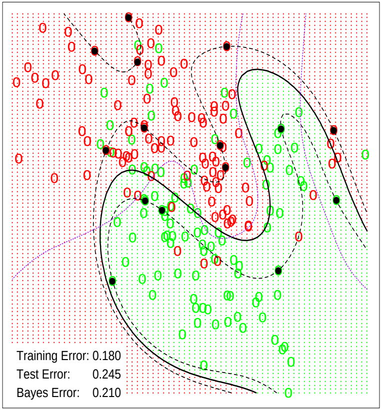
$$\begin{aligned} K(x, x') &= (1 + x_1 x_1' + x_2 x_2')^2 \\ &= 1 + 2x_1 x_1' + 2x_2 x_2' + (x_1 x_1')^2 \\ &\quad + (x_2 x_2')^2 + 2x_1 x_1' x_2 x_2' \\ &= \langle b(x), b(x') \rangle \end{aligned}$$

Popular Kernels

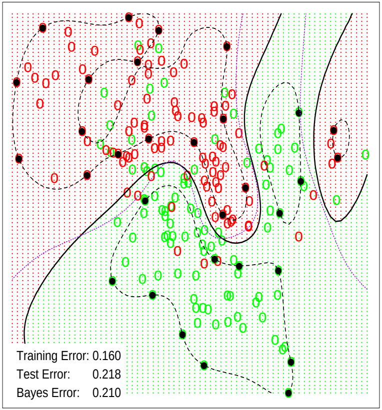
- dth degree polynomial: $K(x,x') = (1 + \langle x, x' \rangle)^d$
- radial basis: $K(x,x') = \exp(-\|x - x'\|^2/\sigma^2)$
- A general kernel function $K(x,x')$ just needs to satisfy two conditions:
  - Symmetric:
    $$K(x,x') = K(x',x)$$
  - Positive (semi-)definite:
    $$\sum_{i,i'=1}^{n} a_i a_{i'} K(x_i, x_{i'}) \geq 0$$
    for every n, and every set of real numbers $a_1, a_2, \ldots, a_n$ and $x_1, x_2, \ldots, x_n$
- Can fit very flexible non-linear boundaries with (nonlinear) Kernel SVMs

Nonlinear SVMs

SVM - Degree-4 Polynomial in Feature Space



SVM - Radial Kernel in Feature Space



The SVM summary

- The main advantage of SVM is its ability to expand into higher-dimensional spaces where separation is easier
- This expansion is made computationally feasible via the use of kernels
- SVMs are not immune to the curse of dimensionality and will suffer if there are many uninformative features