SB2b Statistical Machine Learning
Hilary Term 2017

Seth Flaxman

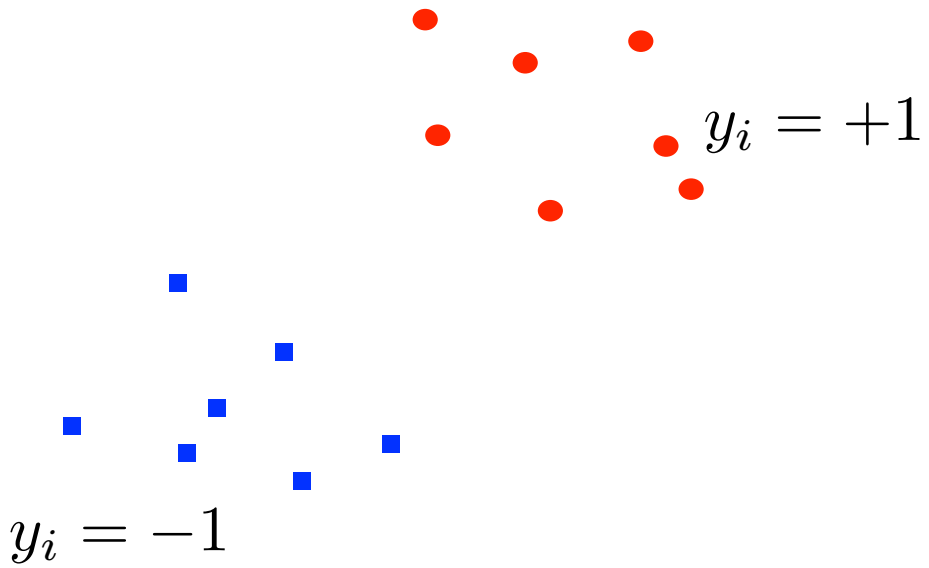
http://www.stats.ox.ac.uk/~flaxman/course_ml.html

Support Vector Machines

These slides are based on Arthur Gretton's UCL course on Advanced Topics in Machine Learning

Linearly separable points

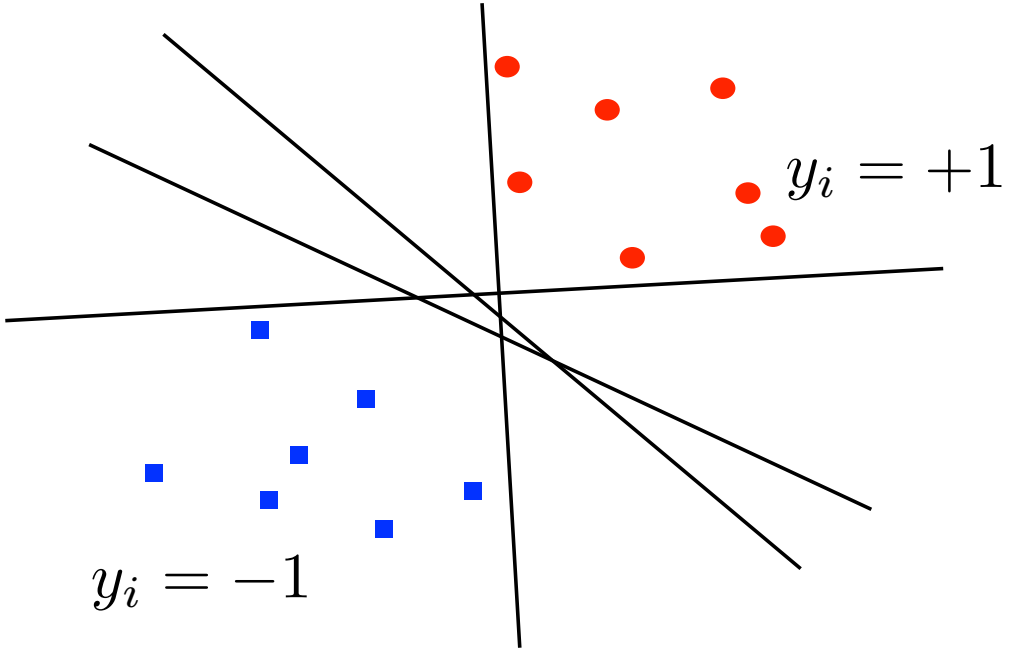
Classify two clouds of points, where there exists a hyperplane which linearly separates one cloud from the other without error.



Data given by $\{x_i, y_i\}_{i=1}^n$, $x_i \in \mathbb{R}^p$, $y_i \in \{-1, +1\}$

Linearly separable points

Classify two clouds of points, where there exists a hyperplane which linearly separates one cloud from the other without error.

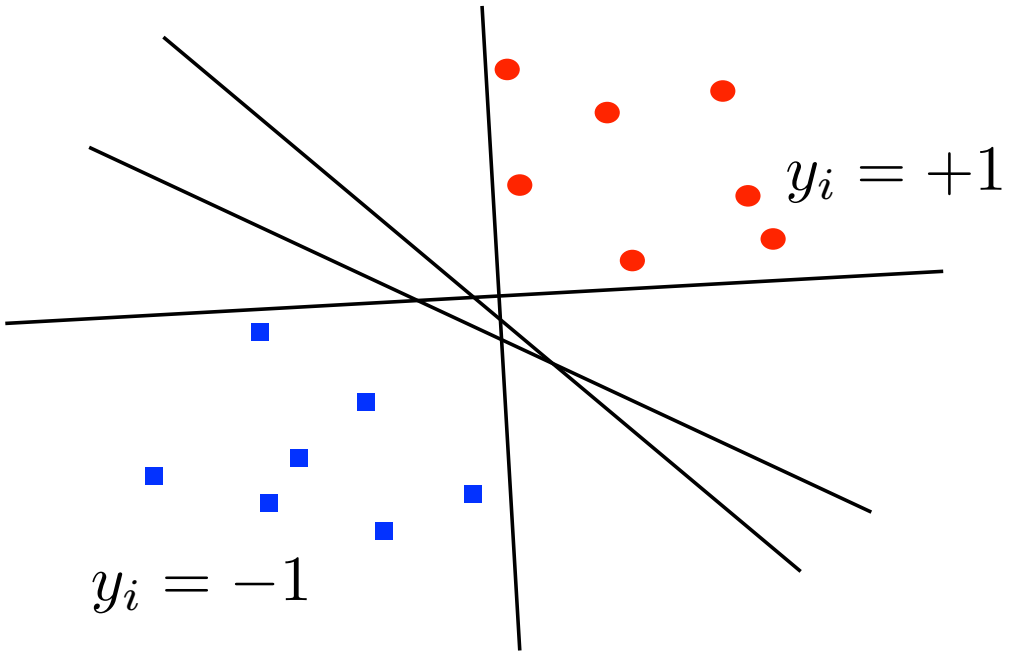


Hyperplane equation $w^\top x + b = 0$. Linear discriminant given by

$$\hat{y}(x) = \text{sign}(w^\top x + b)$$

Linearly separable points

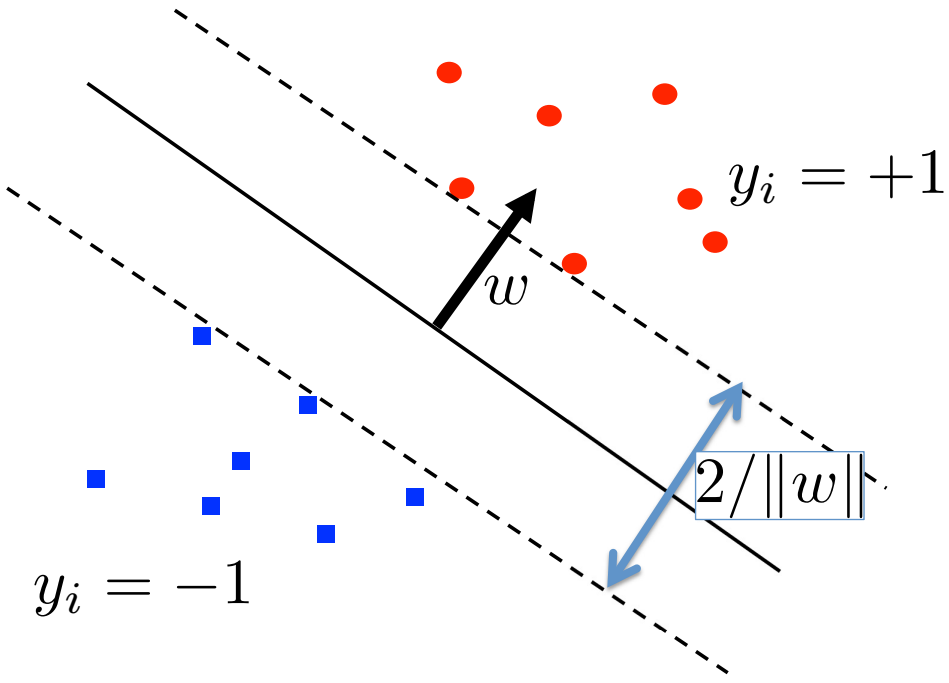
Classify two clouds of points, where there exists a hyperplane which linearly separates one cloud from the other without error.



For a datapoint close to the decision boundary, a small change leads to a change in classification. Can we make the classifier more robust?

Linearly separable points

Classify two clouds of points, where there exists a hyperplane which linearly separates one cloud from the other without error.



Smallest distance from each class to the separating hyperplane $w^\top x + b$ is called the **margin**.

Maximum margin classifier, linearly separable case

This problem can be expressed as follows:

$$\max_{w,b} (\text{margin}) = \max_{w,b} \left(\frac{1}{\|w\|} \right)$$

subject to

$$\begin{cases} w^\top x_i + b \geq 1 & i : y_i = +1, \\ w^\top x_i + b \leq -1 & i : y_i = -1. \end{cases}$$

The resulting classifier is

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

We can rewrite to obtain a **quadratic program**:

$$\min_{w,b} \frac{1}{2} \|w\|^2$$

subject to

$$y_i(w^\top x_i + b) \geq 1.$$

Maximum margin classifier: with errors allowed

Allow "errors": points within the margin, or even on the wrong side of the decision boundary. Ideally:

$$\min_{w,b} \left(\frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \mathbb{I}\left[y_i \left(w^\top x_i + b\right) < 0\right] \right),$$

where C controls the tradeoff between maximum margin and loss. Replace with **convex upper bound**:

$$\min_{w,b} \left(\frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} h\left(y_i \left(w^\top x_i + b\right)\right) \right).$$

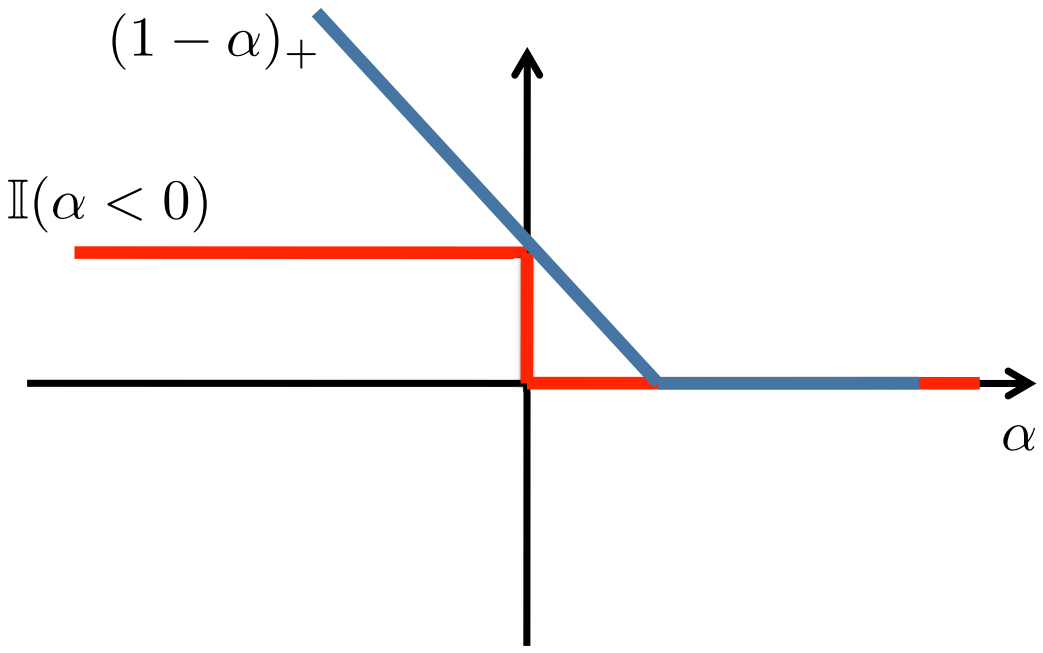
with hinge loss,

$$h(\alpha) = (1-\alpha)_+ = \begin{cases} 1-\alpha, & 1-\alpha > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Hinge loss

Hinge loss:

$$h(\alpha) = (1-\alpha)_+ = \begin{cases} 1-\alpha, & 1-\alpha > 0 \\ 0, & \text{otherwise.} \end{cases}$$



Support vector classification

Substituting in the hinge loss, we get a standard regularised empirical risk minimisation problem - where regularisation naturally arises from the margin penalty.

$$\min_{w,b} \left(\frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} h\left(y_i \left(w^\top x_i + b\right)\right) \right).$$

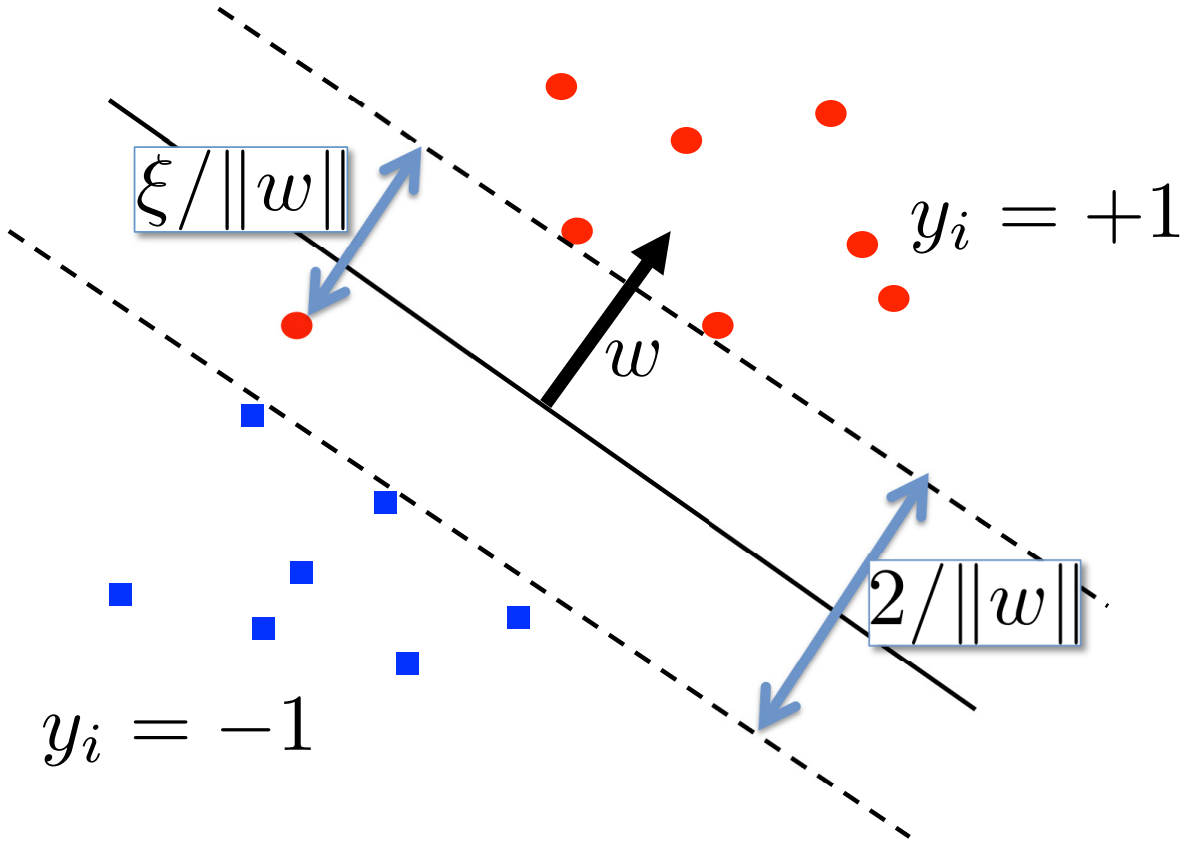
Using substitution $\xi_i = h\left(y_i \left(w^\top x_i + b\right)\right)$, we obtain an equivalent formulation (standard C-SVM):

$$\min_{w,b,\xi} \left(\frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \xi_i \right)$$

subject to

$$\xi_i \geq 0 \qquad y_i \left(w^\top x_i + b\right) \geq 1 - \xi_i$$

Support vector classification



Duality

As a convex constrained optimization problem with affine constraints in w, b, ξ, strong duality holds.

$$\begin{aligned} \text{minimize} \quad & f_0(w,b,\xi) := \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \\ \text{subject to} \quad & f_i(w,b,\xi) := 1 - \xi_i - y_i\left(w^\top x_i + b\right) \le 0,\ i = 1,\dots,n \\ & f_{n+i}(w,b,\xi) := -\xi_i \le 0,\ i = 1,\dots,n. \end{aligned}$$

Support vector classification: Lagrangian

The Lagrangian: $L(w, b, \xi, \alpha, \lambda) =$

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i\left(1 - \xi_i - y_i\left(w^\top x_i + b\right)\right) + \sum_{i=1}^{n}\lambda_i(-\xi_i)$$

with dual variable constraints

$$\alpha_i \geq 0, \qquad \lambda_i \geq 0.$$

Minimize wrt the primal variables w, b, and ξ.
Derivative wrt w:

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{n}\alpha_i y_i x_i = 0 \qquad w = \sum_{i=1}^{n}\alpha_i y_i x_i.$$

Derivative wrt b:

$$\frac{\partial L}{\partial b} = \sum_i y_i \alpha_i = 0.$$

Support vector classification: Lagrangian

Derivative wrt ξ_i:

$$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \lambda_i = 0 \qquad \alpha_i = C - \lambda_i.$$

Since $\lambda_i \geq 0$,

$$\alpha_i \leq C.$$

Now use complementary slackness:

Non-margin SVs (margin errors): $\alpha_i = C > 0$:

1. We immediately have $y_i \left(w^\top x_i + b\right) = 1 - \xi_i$.
2. Also, from condition $\alpha_i = C - \lambda_i$, we have $\lambda_i = 0$, so $\xi_i \geq 0$

Margin SVs: $0 < \alpha_i < C$:

1. We again have $y_i \left(w^\top x_i + b\right) = 1 - \xi_i$.
2. This time, from $\alpha_i = C - \lambda_i$, we have $\lambda_i > 0$, hence $\xi_i = 0$.

Non-SVs (on the correct side of the margin): $\alpha_i = 0$:

1. From $\alpha_i = C - \lambda_i$, we have $\lambda_i > 0$, hence $\xi_i = 0$.
2. Thus, $y_i \left(w^\top x_i + b\right) \geq 1$

The support vectors

We observe:

1. The solution is sparse: points which are neither on the margin nor "margin errors" have $\alpha_i = 0$
2. The support vectors: only those points on the decision boundary, or which are margin errors, contribute.
3. Influence of the non-margin SVs is bounded, since their weight cannot exceed C.

Support vector classification: dual function

Thus, our goal is to maximize the dual,

$$
\begin{aligned}
g(\alpha,\lambda) &= \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i\left(1 - y_i\left(w^\top x_i + b\right) - \xi_i\right) \\
&\quad + \sum_{i=1}^{n}\lambda_i(-\xi_i) \\
&= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^\top x_j + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^\top x_j \\
&\quad \underbrace{-b\sum_{i=1}^{n}\alpha_i y_i}_{0} + \sum_{i=1}^{n}\alpha_i - \sum_{i=1}^{n}\alpha_i\xi_i - \sum_{i=1}^{n}(C-\alpha_i)\xi_i \\
&= \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^\top x_j.
\end{aligned}
$$

Dual C-SVM

$$\text{maximize} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^\top x_j,$$

subject to the constraints

$$0 \leq \alpha_i \leq C, \quad \sum_{i=1}^{n} y_i \alpha_i = 0$$

This is a quadratic program. From α, obtain the hyperplane with

$$w = \sum_{i=1}^{n} \alpha_i y_i x_i$$

(follows from complementary slackness in the derivation of the dual). Offset b can be obtained from any of the margin SVs (for which $\alpha_i \in (0, C)$): $1 = y_i \left(w^\top x_i + b \right)$.

Solution depends on data through inner products only

Dual program

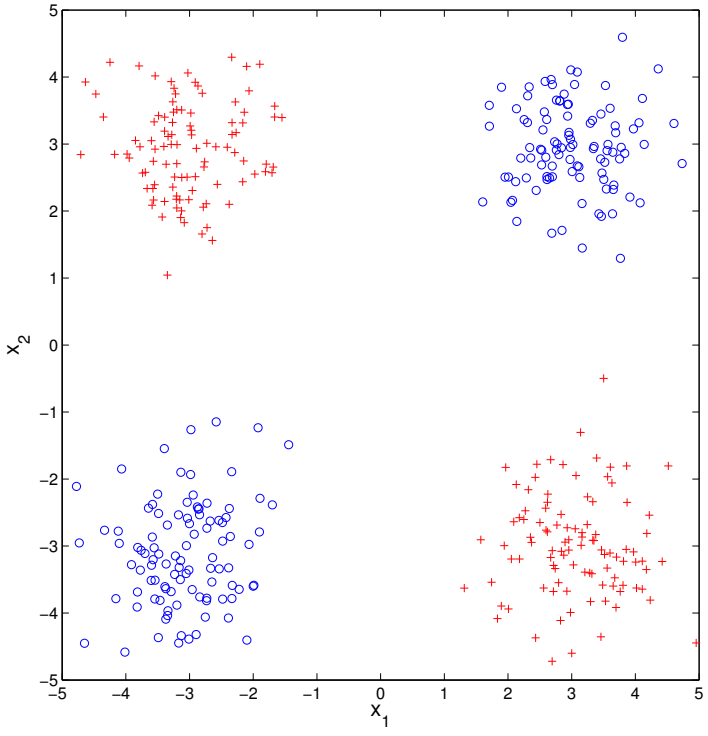
$$\max_{\alpha} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^\top x_j \qquad \text{subject to} \quad \begin{cases} \sum_{i=1}^{n} \alpha_i y_i = 0 \\ 0 \preceq \alpha \preceq C \end{cases}$$

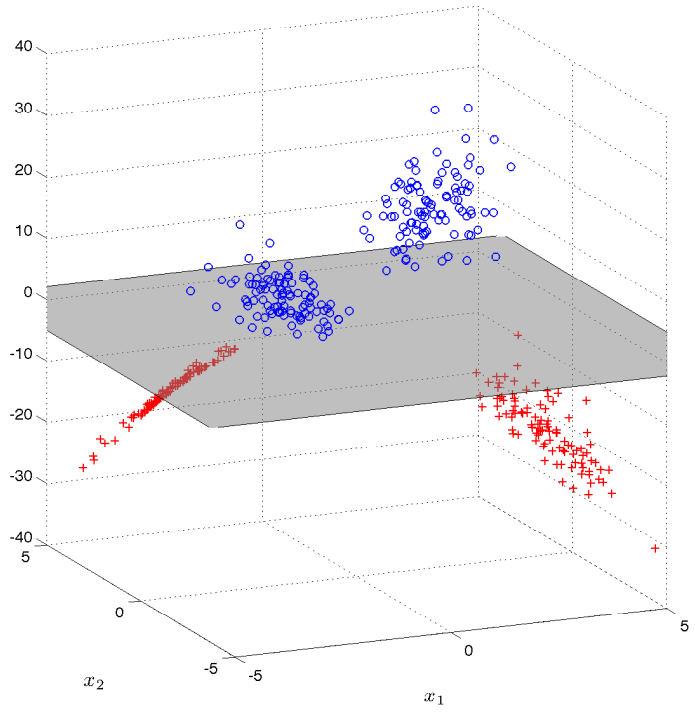
only depends on inputs x_i through their inner products (similarities) with other inputs.
Decision function

$$\hat{y}(x) = \text{sign}(w^\top x + b) = \text{sign}(\sum_{i=1}^{n} \alpha_i y_i x_i^\top x + b)$$

also depends only on the similarity of a test point x to the training points x_i.
Thus, we do not need explicit inputs - just their pairwise similarities.
Key property: even if $p > n$, it is still the case that $w \in \text{span}\{x_i : i = 1, \ldots, n\}$ (normal vector of the hyperplane lives in the subspace spanned by the datapoints).

Beyond Linear Classifiers





- No linear classifier separates red from blue.
- Linear separation after mapping to a **higher dimensional feature space**:

$$\mathbb{R}^2 \ni \begin{pmatrix} x^{(1)} & x^{(2)} \end{pmatrix}^\top = x \mapsto \varphi(x) = \begin{pmatrix} x^{(1)} & x^{(2)} & x^{(1)}x^{(2)} \end{pmatrix}^\top \in \mathbb{R}^3$$

Non-Linear SVM

- Consider the dual C-SVM with explicit non-linear transformation $x \mapsto \varphi(x)$:

$$\max_{\alpha} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \varphi(x_i)^\top \varphi(x_j) \qquad \text{subject to} \quad \begin{cases} \sum_{i=1}^{n} \alpha_i y_i = 0 \\ 0 \preceq \alpha \preceq C \end{cases}$$

- Suppose $p = 2$, and we would like to introduce quadratic non-linearities,

$$\varphi(x) = \left(1, \sqrt{2}x^{(1)}, \sqrt{2}x^{(2)}, \sqrt{2}x^{(1)}x^{(2)}, \left(x^{(1)}\right)^2, \left(x^{(2)}\right)^2\right)^\top.$$

  Then

$$\varphi(x_i)^\top \varphi(x_j) = 1 + 2x_i^{(1)}x_j^{(1)} + 2x_i^{(2)}x_j^{(2)} + 2x_i^{(1)}x_i^{(2)}x_j^{(1)}x_j^{(2)} + \left(x_i^{(1)}\right)^2 \left(x_j^{(1)}\right)^2 + \left(x_i^{(2)}\right)^2 \left(x_j^{(2)}\right)^2 = (1 + x_i^\top x_j)^2$$

- Since only inner products are needed, non-linear transform need not be computed explicitly - inner product between features can be a simple function (**kernel**) of x_i and x_j: $k(x_i, x_j) = \varphi(x_i)^\top \varphi(x_j) = (1 + x_i^\top x_j)^2$
- d-order interactions can be implemented by $k(x_i, x_j) = (1 + x_i^\top x_j)^d$ (**polynomial kernel**). Never need to compute explicit feature expansion of dimension $\binom{p+d}{d}$ where this inner product happens!

Kernel SVM: Kernel trick

- Kernel SVM with $k(x_i, x_j)$. Non-linear transformation $x \mapsto \varphi(x)$ still present, but **implicit** (coordinates of the vector $\varphi(x)$ are never computed).

$$\max_{\alpha} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j k(x_i, x_j) \qquad \text{subject to} \quad \begin{cases} \sum_{i=1}^{n} \alpha_i y_i = 0 \\ 0 \preceq \alpha \preceq C \end{cases}$$

- Prediction? $\hat{y}(x) = \text{sign}\left(w^\top \varphi(x) + b\right)$, where $w = \sum_{i=1}^{n} \alpha_i y_i \varphi(x_i)$ and offset b obtained from a margin support vector x_j with $\alpha_j \in (0, C)$.
  - No need to compute w either! Just need

$$w^\top \varphi(x) = \sum_{i=1}^{n} \alpha_i y_i \varphi(x_i)^\top \varphi(x) = \sum_{i=1}^{n} \alpha_i y_i k(x_i, x).$$

  - Get offset from

$$b = y_j - w^\top \varphi(x_j) = y_j - \sum_{i=1}^{n} \alpha_i y_i k(x_i, x_j)$$

    for any margin support-vector x_j ($\alpha_j \in (0, C)$).
- Fitted a separating hyperplane in a high-dimensional feature space without ever mapping explicitly to that space.