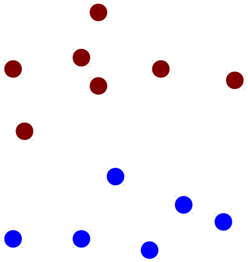
Some intuition to the Kernel trick for Support Vector Machines

Algorithms & Data Challenges Berlin
Talks n' Beer Meetup

Alexander Weiß

March 19, 2013

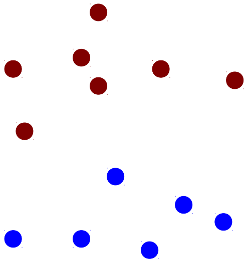
The problem: binary classification



sample set $x_1, \ldots, x_N$, each $x_n \in \mathbb{R}^m$

each x_n **labelled** by $y_n \in \{-1, +1\}$

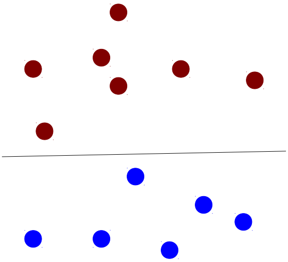
The problem: binary classification



sample set $x_1, \ldots, x_N$, each $x_n \in \mathbb{R}^m$
each x_n **labelled** by $y_n \in \{-1, +1\}$

Task: Find *rule* to decide if a data point $x \in \mathbb{R}^m$ belongs to group -1 or $+1$

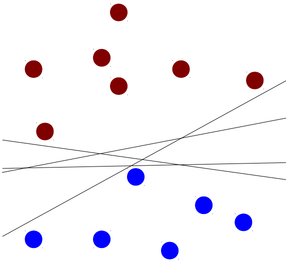
A solution: a hyperplane



sample set $x_1, \dots, x_N$, each $x_n \in \mathbb{R}^m$
each x_n **labelled** by $y_n \in \{-1, +1\}$

Task: Find *rule* to decide if a data point $x \in \mathbb{R}^m$ belongs to group -1 or $+1$

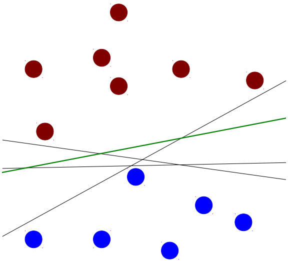
Some solutions: some hyperplanes



sample set $x_1, \ldots, x_N$, each $x_n \in \mathbb{R}^m$
each x_n **labelled** by $y_n \in \{-1, +1\}$

Task: Find *rule* to decide if a data point $x \in \mathbb{R}^m$ belongs to group -1 or $+1$

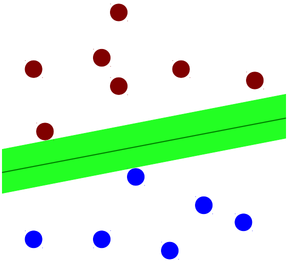
The SVM solution: the distance maximizing hyperplane



sample set $x_1, \ldots, x_N$, each $x_n \in \mathbb{R}^m$
each x_n **labelled** by $y_n \in \{-1, +1\}$

Task: Find *rule* to decide if a data point $x \in \mathbb{R}^m$ belongs to group -1 or $+1$

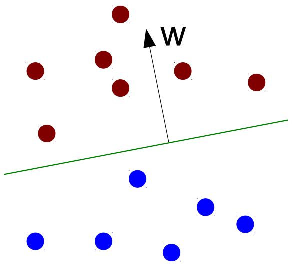
The SVM solution: the distance maximizing hyperplane



sample set $x_1, \ldots, x_N$, each $x_n \in \mathbb{R}^m$
each x_n **labelled** by $y_n \in \{-1, +1\}$

Task: Find *rule* to decide if a data point $x \in \mathbb{R}^m$ belongs to group -1 or $+1$

The SVM solution: how to find it?



hyperplane is set of all vectors z such that $w^\intercal z + b = 0$
vector w is orthogonal to hyperplane, b is shift away from origin

Task: Find w and b for optimal hyperplane

The SVM solution: the optimization problem

$$\max_{w \in \mathbb{R}^m} \frac{1}{||w||}$$

$$\text{s. t.} \quad \min_{n} \left| w^\intercal x_n + b \right| = 1$$

The SVM solution: the optimization problem

$$\max_{w \in \mathbb{R}^m} \frac{1}{||w||}$$

$$\text{s. t.} \quad \min_n \left| w^\intercal x_n + b \right| = 1$$

is equivalent to

$$\min_{w \in \mathbb{R}^m} \frac{1}{2} w^\intercal w$$

$$\text{s. t.} \quad y_n(w^\intercal x_n + b) \geq 1 \text{ for } n = 1, \ldots, N$$

The SVM solution: Applying KKT conditions

$$\min_{\alpha \in \mathbb{R}^N} \frac{1}{2}\alpha^\mathsf{T} \begin{pmatrix} y_1 y_1 x_1^\mathsf{T} x_1 & \dots & y_1 y_N x_1^\mathsf{T} x_N \\ \dots & \dots & \dots \\ y_N y_1 x_N^\mathsf{T} x_1 & \dots & y_N y_N x_N^\mathsf{T} x_N \end{pmatrix} \alpha + (-1^\mathsf{T})\alpha$$

$$\text{s. t.} \qquad y^\mathsf{T}\alpha = 0$$

$$0 \leq \alpha \leq \infty$$

The SVM solution: Applying KKT conditions

$$\min_{\alpha\in\mathbb{R}^N} \frac{1}{2}\alpha^\intercal \begin{pmatrix} y_1 y_1 x_1^\intercal x_1 & \dots & y_1 y_N x_1^\intercal x_N \\ \dots & \dots & \dots \\ y_N y_1 x_N^\intercal x_1 & \dots & y_N y_N x_N^\intercal x_N \end{pmatrix} \alpha + (-1^\intercal)\alpha$$

$$\text{s. t.} \qquad y^\intercal \alpha = 0$$

$$0 \leq \alpha \leq \infty$$

Quadratic Program $\Rightarrow$ can be solved numerically

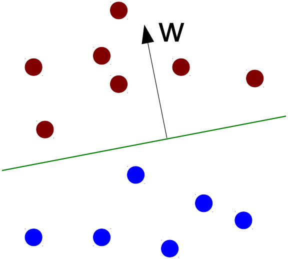
$$w = \sum_n \alpha_n y_n x_n$$

Find b by solving

$$y_n(w^\intercal x_n + b) = 1$$

for any n with $\alpha_n \neq 0$.

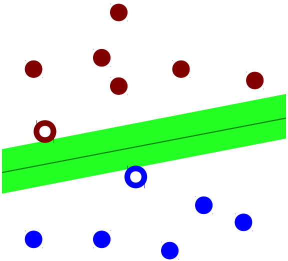
The SVM solution: support vectors



$$w = \sum_n \alpha_n y_n x_n$$

x_n with $\alpha_n \neq 0$ are called **support vectors**

The SVM solution: support vectors



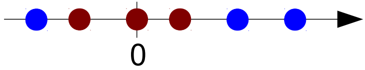
$$w = \sum_n \alpha_n y_n x_n$$

x_n with $\alpha_n \neq 0$ are called **support vectors**

x_n is support vector $\Leftrightarrow$ x_n is closest to hyperplane

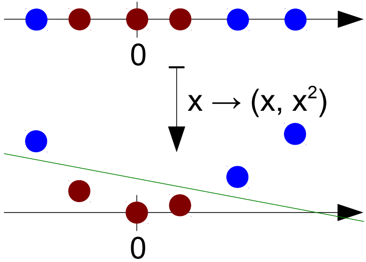
Support Vector Machine: the non-linear case

An *unsolvable* example for a linear SVM:



Support Vector Machine: the non-linear case

Well, solvable in a higher dimensional space:



Support Vector Machine: the non-linear case

General idea:

$$\mathbb{R}^m \to \mathbb{R}^{m+k}, k > 0$$

In practice, one often takes polynomial approach, e.g.

$$(x_1, x_2, x_3) \to (1, x_1, x_2, x_3, x_1^2, x_2^2, x_3^2, x_1x_2, x_1x_3, x_2x_3)$$

General notation:

$$\begin{aligned} X &\to Z \\ x &\mapsto z \end{aligned}$$

Applying KKT conditions to transformed space

$$\min_{\alpha \in \mathbb{R}^N} \frac{1}{2}\alpha^\intercal \begin{pmatrix} y_1 y_1 z_1^\intercal z_1 & \dots & y_1 y_N z_1^\intercal z_N \\ \dots & \dots & \dots \\ y_N y_1 z_N^\intercal z_1 & \dots & y_N y_N z_N^\intercal z_N \end{pmatrix} \alpha + (-1^\intercal)\alpha$$

$$\text{s. t.} \qquad y^\intercal \alpha = 0$$

$$0 \leq \alpha \leq \infty$$

$$w = \sum_n \alpha_n y_n z_n$$

Find b by solving

$$y_n(w^\intercal z_n + b) = 1$$

for any n with $\alpha_n \neq 0$.

Applying KKT conditions to transformed space

$$\min_{\alpha \in \mathbb{R}^N} \frac{1}{2}\alpha^\intercal \begin{pmatrix} y_1 y_1 z_1^\intercal z_1 & \dots & y_1 y_N z_1^\intercal z_N \\ \dots & \dots & \dots \\ y_N y_1 z_N^\intercal z_1 & \dots & y_N y_N z_N^\intercal z_N \end{pmatrix} \alpha + (-1^\intercal)\alpha$$

$$\text{s. t.} \qquad y^\intercal \alpha = 0$$

$$0 \leq \alpha \leq \infty$$

$$w = \sum_n \alpha_n y_n z_n$$

Find b by solving

$$y_n(w^\intercal z_n + b) = 1 = y_n \left(\sum_m \alpha_m y_m z_m^\intercal z_n + b \right)$$

for any n with $\alpha_n \neq 0$.

Kernel function

Lesson: We only need inner product in transformed space!

Kernel function

Lesson: We only need inner product in transformed space!

Define **kernel function**:

$$K(x, \bar{x}) := z^{\intercal} \bar{z}$$

where z and $\bar{z}$ are the transformations of x and $\bar{x}$

Attention: Not every function is a kernel function.

Kernel function - two examples

Example 1: $x, \bar{x} \in \mathbb{R}^2$

$$
\begin{aligned}
K(x, \bar{x}) &:= (1 + x^\intercal \bar{x})^2 \\
&= 1 + x_1^2 \bar{x}_1^2 + x_2^2 \bar{x}_2^2 + 2x_1\bar{x}_1 + 2x_2\bar{x}_2 + 2x_1\bar{x}_1 x_2\bar{x}_2
\end{aligned}
$$

K is inner product of transformation

$$
(x_1, x_2) \mapsto (1, x_1^2, x_2^2, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1 x_2)
$$

Kernel function - two examples

Example 1: $x, \bar{x} \in \mathbb{R}^2$

$$\begin{aligned} K(x,\bar{x}) &:= (1 + x^\intercal \bar{x})^2 \\ &= 1 + x_1^2\bar{x}_1^2 + x_2^2\bar{x}_2^2 + 2x_1\bar{x}_1 + 2x_2\bar{x}_2 + 2x_1\bar{x}_1x_2\bar{x}_2 \end{aligned}$$

K is inner product of transformation

$$(x_1, x_2) \mapsto (1, x_1^2, x_2^2, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1x_2)$$

In general:

$$K(x,\bar{x}) := (1 + x^\intercal \bar{x})^Q$$

is called the **polynomial kernel**

Kernel function - two examples

Example 2: $x, \bar{x} \in \mathbb{R}$

$$
\begin{aligned}
K(x,\bar{x}) &:= \exp\left(-(x-\bar{x})^2\right) \\
&= \exp(-x^2)\exp(-\bar{x}^2)\underbrace{\sum_{k=0}^{\infty}\frac{2^k x^k \bar{x}^k}{k!}}_{exp(2x\bar{x})} \\
&= \sum_{k=0}^{\infty}\left[\left(\sqrt{\frac{2^k}{k!}}\exp(-x^2)x^k\right)\left(\sqrt{\frac{2^k}{k!}}\exp(-\bar{x}^2)\bar{x}^k\right)\right]
\end{aligned}
$$

Kernel function - two examples

Example 2: $x, \bar{x} \in \mathbb{R}$

$$\begin{aligned} K(x,\bar{x}) &:= \exp\left(-(x-\bar{x})^2\right) \\ &= \exp(-x^2)\exp(-\bar{x}^2)\underbrace{\sum_{k=0}^{\infty}\frac{2^k x^k \bar{x}^k}{k!}}_{exp(2x\bar{x})} \\ &= \sum_{k=0}^{\infty}\left[\left(\sqrt{\frac{2^k}{k!}}\exp(-x^2)x^k\right)\left(\sqrt{\frac{2^k}{k!}}\exp(-\bar{x}^2)\bar{x}^k\right)\right] \end{aligned}$$

In general:

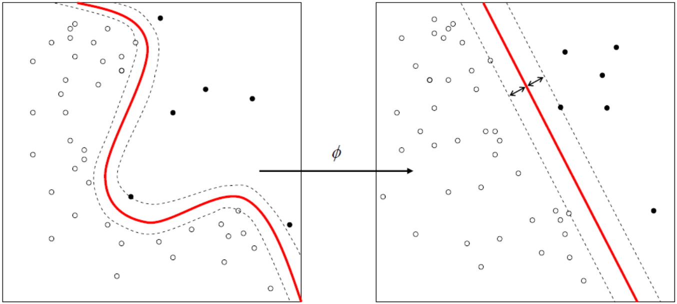
$$K(x,\bar{x}) := \exp\left(-\gamma||x-\bar{x}||^2\right)$$

is called the **radial basis function (RBF) kernel**

Kernel function - two examples

Z is infinite dimensional in this case

The hyperplane is still defined by a few support vectors.



Source of illustration: Wikipedia

Thanks.

(Examples inspired by lectures of Y. S. Abu-Mostafa @ Caltech)

By the way: Trademob is hiring!