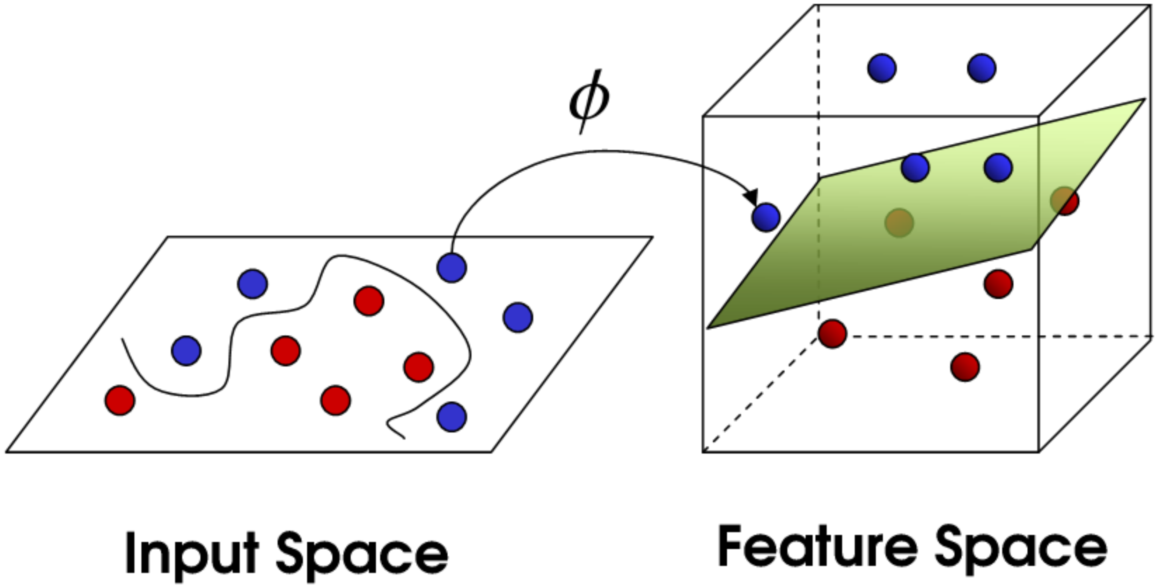
Kernel SVMs

Ali Gooya

Feature transform

- What if the input data points are non-linearly separable?



SVMs with transformed features

- Suppose instead of using the using $\mathbf{x} \in \mathcal{R}^m$, we use a map $\phi(\mathbf{x}) \in \mathcal{R}^M, m << M$
- Then if the data is linearly serapeable in the $\mathcal{R}^M$, we can solve for the same dual problem, replacing $\mathbf{x}_i^\top \mathbf{x}_j$ with $\phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j)$

$$g(\boldsymbol{\alpha}) = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} t_i t_j \alpha_i \alpha_j \boldsymbol{\phi}(\mathbf{x}_i)^\top \boldsymbol{\phi}(\mathbf{x}_j)$$

- This will solve for $\boldsymbol{\alpha}^*$ (hence prediction) in terms of $\phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j)$.

- The problem is that we do not know which mapping should be used.

Kernel Trick

- Note that $\phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j)$ only measures the similarity of the two data points $\mathbf{x}_i$ and $\mathbf{x}_j$ in the higher dimension.

- Can we replace this inner product by a kernel function

$$k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j)$$

Not to be mistaken with corn kernels!

that measures a similarity without knowing the exact form of $\phi(\mathbf{x})$?

- Answer: if we can factorizable it as inner product, we do not need $\phi(\mathbf{x})$

Valid kernels are factorizable.

Example

e.g. Let $m = 2$ and define $k(\mathbf{x}, \mathbf{x}') := (\mathbf{x}^\top \mathbf{x}')^2$. Easy to check that $k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^\top \phi(\mathbf{x}')$ where

$$\phi(\mathbf{x}) := \left(x_1^2,\, \sqrt{2}x_1 x_2,\, x_2^2\right).$$

But calculating $k(\mathbf{x}, \mathbf{x}')$ requires $O(m)$ $(= \dim(\mathbf{x}))$ work whereas calculating $\phi(\mathbf{x})^\top \phi(\mathbf{x}')$ requires $O(M)$ work.

How to construct valid kernels?

We assume:

- $k_1(\mathbf{x}, \mathbf{x}')$ and $k_2(\mathbf{x}, \mathbf{x}')$ are valid kernels.
- $c > 0$ is a constant.
- $f(\cdot)$ is any function.
- $q(\cdot)$ is a polynomial with nonnegative coefficients.
- $\boldsymbol{\phi}(\mathbf{x})$ is a function from $\mathbf{x}$ to $\mathbb{R}^M$.
- $k_3(\cdot, \cdot)$ is a valid kernel in $\mathbb{R}^M$.
- $\mathbf{A}$ is a symmetric positive semi-definite matrix.
- $\mathbf{x}_a$ and $\mathbf{x}_b$ are variables (not necessarily disjoint) with $\mathbf{x} = (\mathbf{x}_a, \mathbf{x}_b)$.
- k_a and k_b are valid kernel functions over their respective spaces.

Rules for valid kernels

Then the following are all valid kernels:

$$
\begin{aligned}
k(x, x') &= ck_1(\mathbf{x}, \mathbf{x}') \\
k(x, x') &= f(\mathbf{x})k_1(\mathbf{x}, \mathbf{x}')f(\mathbf{x}') && * \\
k(x, x') &= q(k_1(\mathbf{x}, \mathbf{x}')) \\
k(x, x') &= \exp(k_1(\mathbf{x}, \mathbf{x}')) && ** \\
k(x, x') &= k_1(\mathbf{x}, \mathbf{x}') + k_2(\mathbf{x}, \mathbf{x}') \\
k(x, x') &= k_1(\mathbf{x}, \mathbf{x}')k_2(\mathbf{x}, \mathbf{x}') \\
k(x, x') &= k_3(\boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{x}')) \\
k(x, x') &= \mathbf{x}^\top \mathbf{A}\mathbf{x}' \\
k(x, x') &= k_a(\mathbf{x}_a, \mathbf{x}'_a) + k_b(\mathbf{x}_b, \mathbf{x}'_b) \\
k(x, x') &= k_a(\mathbf{x}_a, \mathbf{x}'_a)k_b(\mathbf{x}_b, \mathbf{x}'_b)
\end{aligned}
$$

Example: The Gaussian kernel

The Gaussian kernel is given by:

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{||\mathbf{x} - \mathbf{x}'||^2}{2\sigma^2}\right)$$

It is a valid kernel because

$$\begin{aligned}
\exp\left(\frac{-||\mathbf{x} - \mathbf{x}'||^2}{2\sigma^2}\right) &= \exp\left(\frac{-\mathbf{x}^\top \mathbf{x}}{2\sigma^2}\right) \exp\left(\frac{\mathbf{x}^\top \mathbf{x}'}{\sigma^2}\right) \exp\left(\frac{-\mathbf{x}'^\top \mathbf{x}'}{2\sigma^2}\right) \\
&= f(\mathbf{x}) \exp\left(\frac{\mathbf{x}^\top \mathbf{x}'}{\sigma^2}\right) f(\mathbf{x}')
\end{aligned}$$

- Then we apply (*) and (**) to infer that the kernel is valid.

Kernel – Separated Dual SVMs

Returning to SVMs, when the data is kernel-separated our dual problem becomes:

$$\begin{aligned} \max_{\alpha \geq \mathbf{0}} \quad & \textstyle\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j t_i t_j k(\mathbf{x}_i, \mathbf{x}_j) \\ \text{subject to} \quad & \textstyle\sum_{i=1}^{n} \alpha_i t_i = 0. \end{aligned}$$

Given a solution $\boldsymbol{\alpha}^*$ to the dual, can obtain corresponding optimal b^* via

$$b^* = t_j - \sum_{i=1}^{n} \alpha_i^* t_i k(\mathbf{x}_i, \mathbf{x}_j) \quad \text{for any } \alpha_j^* > 0$$

and, for a new data-point $\mathbf{x}$, the prediction

$$\text{sign}\left(\mathbf{w}^{*\top} \boldsymbol{\phi}(\mathbf{x}) + b^*\right) = \text{sign} \quad \sum_{i=1}^{n} \alpha_i^* t_i k(\mathbf{x}_i, \mathbf{x}) + b^*$$

Example of kernel SVMs

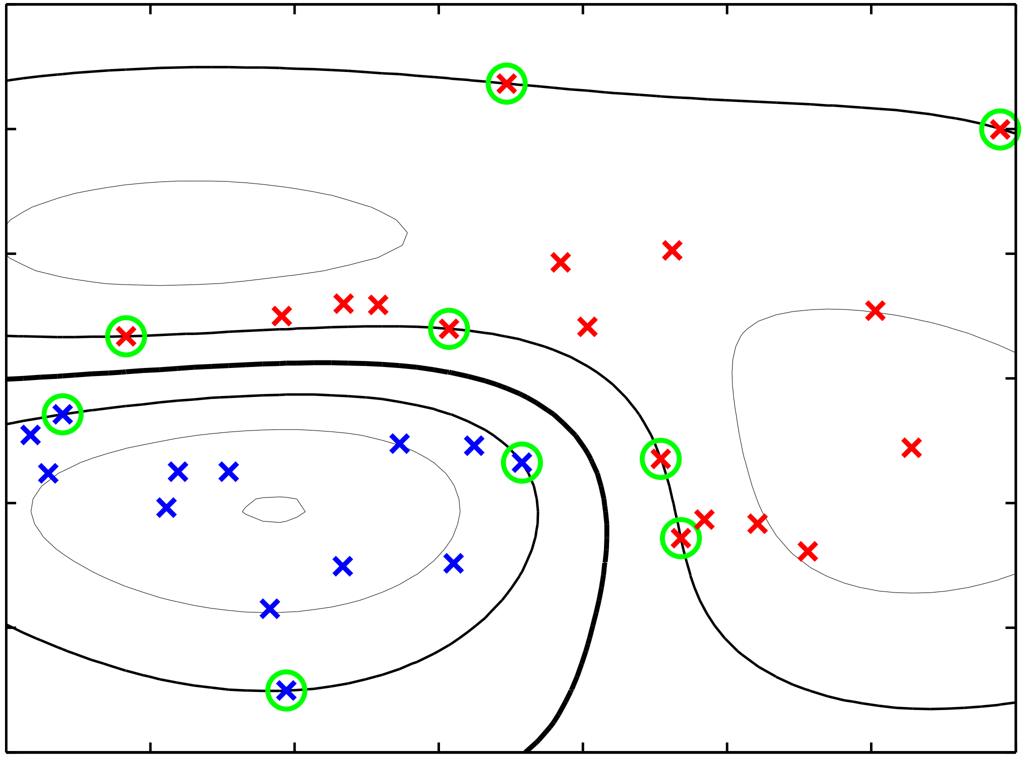


Figure 7.2 from Bishop: Example of synthetic data from two classes in two dimensions showing contours of constant $y(x)$ obtained from a support vector machine having a Gaussian kernel function. Also shown are the decision boundary, the margin boundaries, and the support vectors.

- Note that the data is linearly separable in the Gaussian-kernel space but not in the original space.