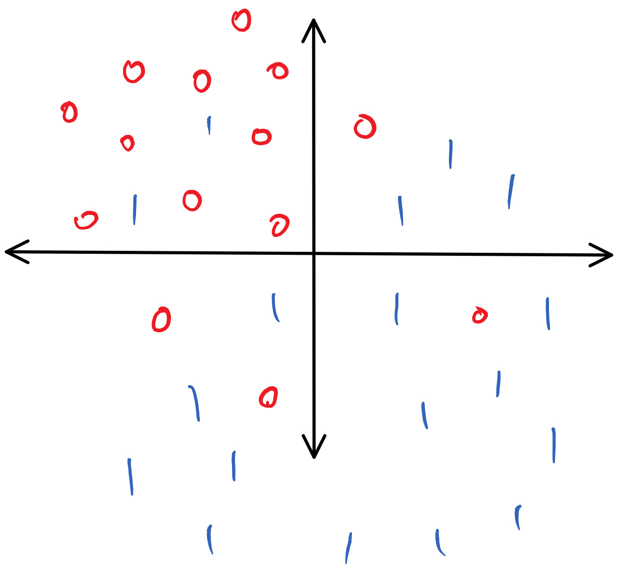
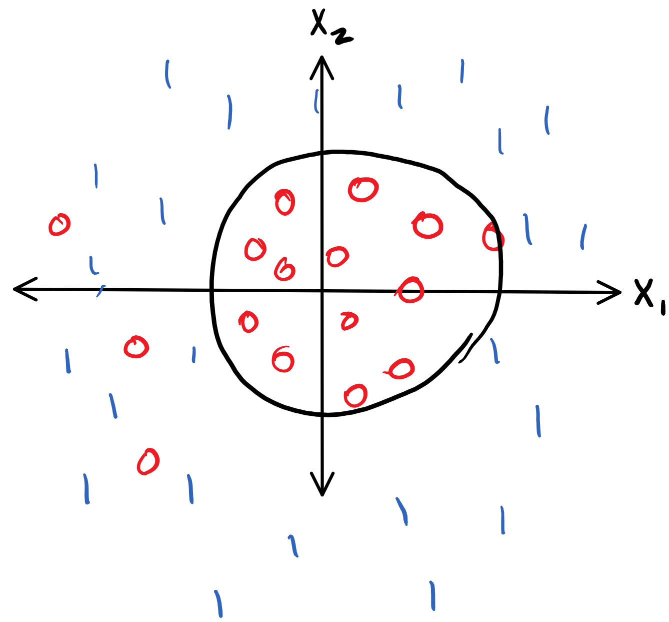
HiDi Embedding

Monday, May 8, 2017 8:11 PM



Linear decision boundary

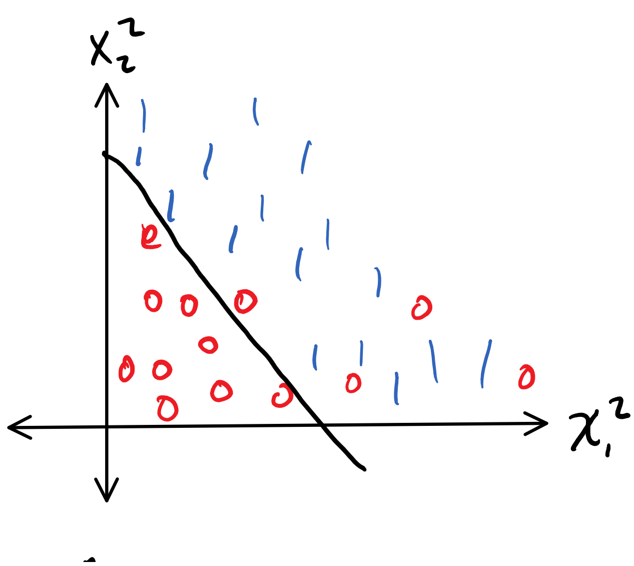


Non-linear decision boundary

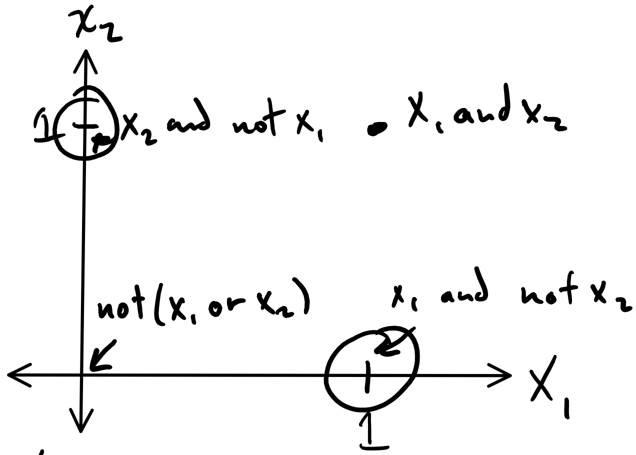
<u>define</u> higher dim embedding $\Phi : \mathbb{R}^p \rightarrow \mathbb{R}^D$

$\Phi(x) \in \mathbb{R}^D$

<u>ex</u> $\Phi(x_1, x_2) = (1, x_1, x_2, x_1^2, x_2^2)$



Φ make linear methods into non-linear method ex Logic: $x_1, \ldots, x_p$ are propositions encoded as $\{0,1\}$.

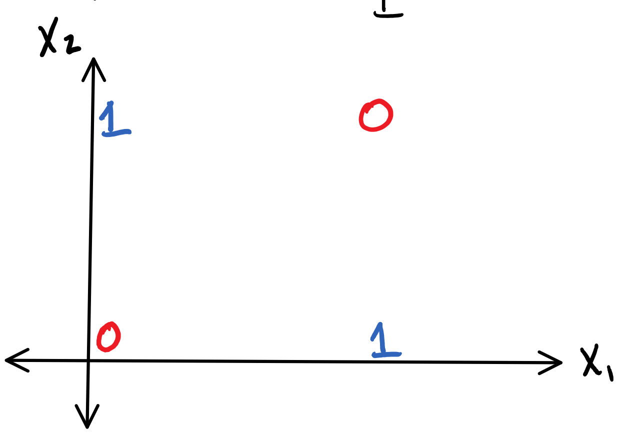


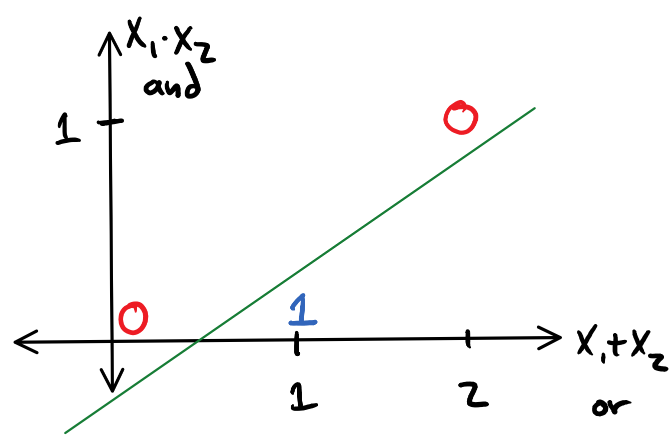
x_1 xor x_2

(x_1 and not x_2) or

(x_2 and not x_1)

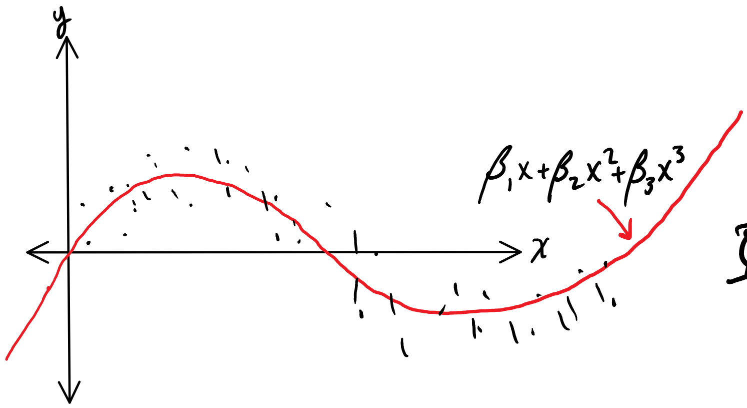
$$\Phi(x_1, x_2) = (x_1, x_2, x_1 \cdot x_2)$$





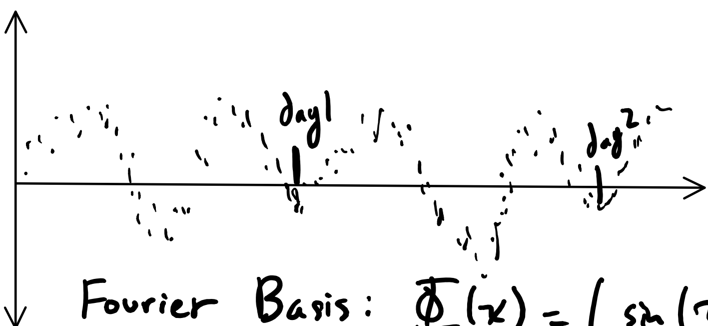
Basis Expansion

Monday, May 8, 2017 8:42 PM



Polynomial Basis

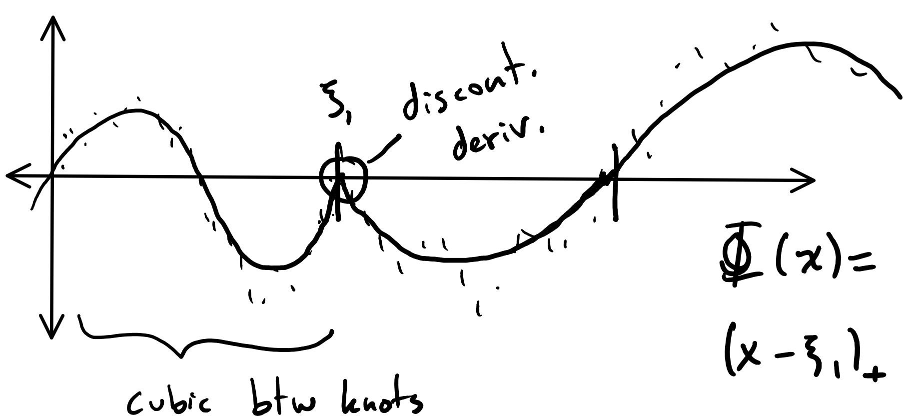
$$\Phi(x) = (1, x, x^2, \ldots, x^d)$$



Periodic (1 day period)

$\sin(4\pi x)$

Fourier Basis: $\Phi(x) = (\sin(2\pi x), \cos(2\pi x), \sin(4\pi x), \cos(4\pi x), \ldots)$



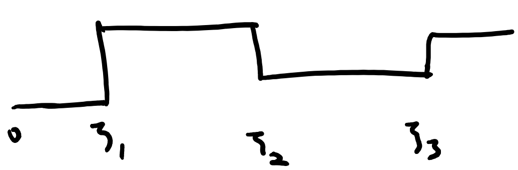
$$\Phi(x) = (1, x, x^2, x^3, (x-\xi_1)_+, (x-\xi_1)_+^2, (x-\xi_1)_+^3, \ldots)$$

Derivatives & Constraints

0th order: $\phi_1(x) = \mathbb{1}\{0 \le x < \xi_1\}$

$\varphi_2(x) = \mathbb{1}\{\xi_1 \le x < \xi_2\}$

$\varphi_3(x) = \mathbb{1}\{\xi_2 \le x < \xi_3\} \equiv \mathbb{1}\{x \ge \xi_j\}$



k^{th} order poly: $\mathbb{1}\{x \geq \xi_j\}$, $\cancel{(x-\xi_j)_+}$, $(x-\xi_j)_+^2, \ldots,$
$(x-\xi_j)_+^k$

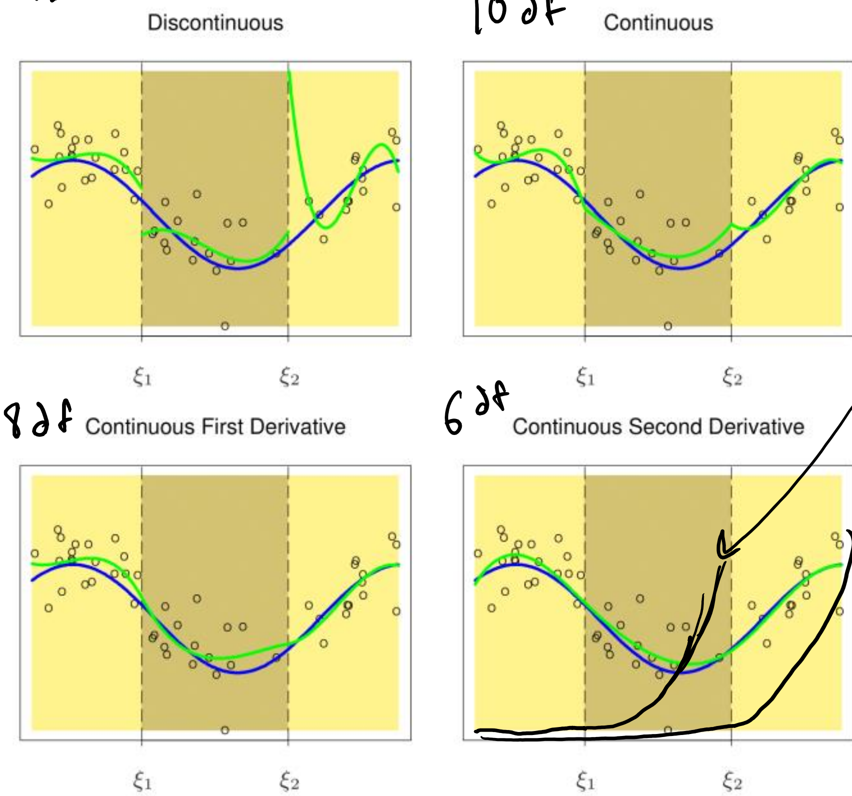
k-1 cont deriv: then remove

12 df

10 df

8 df

6 df



$\underline{1}, \underline{x}, \underline{x^2}, \underline{x^3},$
$\underline{(x-\xi_1)_+^3}, \underline{(x-\xi_2)_+^3}$

FIGURE 5.2. *A series of piecewise-cubic polynomials, with increasing orders of continuity.*

ESL 5.2

Kernel Trick

let $Z_{ik} = \varphi_k(x_i) \quad i=1,\dots,n \quad k=1,\dots,D$

SVM for $y_i \in \{-1,1\}$

$$\min_\beta \frac{1}{n}\underbrace{\sum_{i=1}^n (1-y_i z_i^T\beta)_+}_{(\mathrm{I})} + \underbrace{\lambda\|\beta\|_2^2}_{(\mathrm{II})}$$

$Z\beta \leftrightarrow$ loss

claim $\hat\beta$ solves SVM can be written as $Z^T\alpha$

$\alpha \in \mathbb{R}^n$ i.e. $\hat\beta_j = \underbrace{Z_j^T}_{\text{column}}\alpha = \sum_i \alpha_i z_{ij}$

proof $\beta = \sum_i \alpha_i z_i + \beta^\perp \qquad z_i^T\beta^\perp = 0$

$$z_i^T\beta = \sum_j \alpha_j z_i^T z_j + \cancel{z_i^T\beta^\perp}_{0} \qquad (\mathrm{I})$$

$$= z_i^T Z^T\alpha$$

$\Rightarrow \beta^\perp$ does not impact R_n

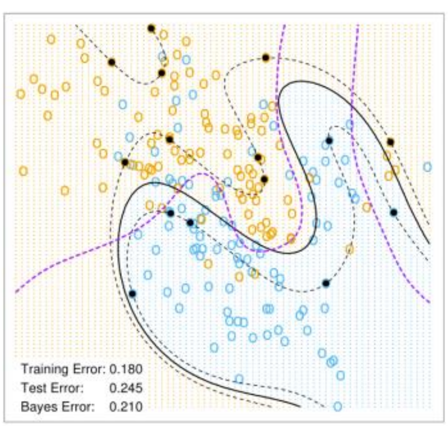
$$\|\beta\|_2^2 = \|Z^T\alpha + \beta^\perp\|_2^2 = \|Z^T\alpha\|_2^2 + 2\underbrace{\beta^{\perp T} Z^T\alpha}_{=} + \|\beta^\perp\|_2^2$$

$$= \|Z^T\alpha\|_2^2 + \|\beta^\perp\|_2^2 \qquad (\mathrm{II}) \quad (Z\beta^\perp)^T\alpha = 0$$

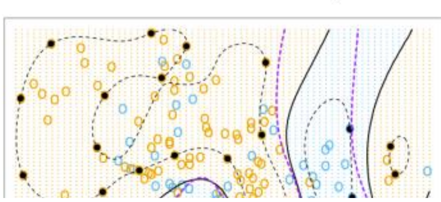
$$\text{SVM: } \min_{\substack{\alpha\in\mathbb{R}^n,\ \beta^\perp\in\mathbb{R}^D \\ \text{s.t. } \beta^{\perp T} z_i = 0\ \forall i}} \frac{1}{n}\sum_i (1-y_i z_i^T Z^T\alpha)_+ + \lambda\left(\|Z^T\alpha\|_2^2 + \|\beta^\perp\|_2^2\right)$$

min'ed when $\beta^\perp = 0$

SVM - Degree-4 Polynomial in Feature Space



SVM - Radial Kernel in Feature Space



Ridge Regression

$$\min_\beta \sum_i (y_i - z_i^T\beta)^2 + \lambda\|\beta\|_2^2$$

same story!

General

$$\min_{\beta\in\mathbb{R}^D} R_n(y, Z\beta) + \lambda\|\beta\|_2^2$$

$$|||$$

$$\min_{\alpha\in\mathbb{R}^n} R_n(y, ZZ^T\alpha) + \lambda \underbrace{\|Z^T\alpha\|_2^2}_{\alpha^T ZZ^T\alpha}$$

define $K = ZZ^T \quad \left(K_{ij} = z_i^T z_j = \Phi(x_i)^T\Phi(x_j)\right)$

$$\min_\alpha R_n(y, K\alpha) + \lambda\alpha^T K\alpha$$

Method 1 define transformation

Φ compute either $Z = \Phi(X)$

or compute $K = \Phi(X)^T\Phi(X)$

and solve SVM

Method 2 define kernel function

$k(x_i, x_j) := \Phi(x_i)^T\Phi(x_j)$

the compute K and solve SVM

Do we need Φ to use a kernel k?

ex $k(x,x') = e^{-\|x-x'\|_2^2/\sigma^2}$

σ bandwidth parameter.

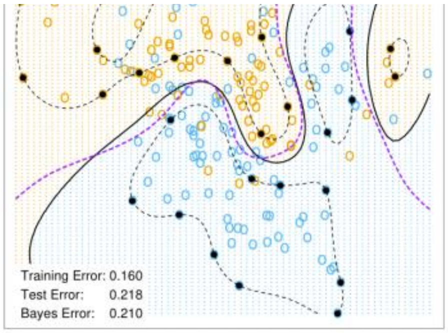


FIGURE 12.3. *Two nonlinear SVMs for the mixture data. The upper plot uses a 4th degree polynomial kernel, the lower a radial basis kernel (with $\gamma = 1$). In each case C was tuned to approximately achieve the best test error performance, and $C = 1$ worked well in both cases. The radial basis kernel performs the best (close to Bayes optimal), as might be expected given the data arise from mixtures of Gaussians. The broken purple curve in the background is the Bayes decision boundary.*

ESL 12.3

<u>def</u> Mercer kernel is a function

$k : \mathbb{R}^p \times \mathbb{R}^p \longrightarrow \mathbb{R}_+$ that is PSD

(for any $\{x_i\} \subseteq \mathbb{R}^d$ $\;(k(x_i, x_j))_{ij}$ is PSD)

<u>ex</u> d^{th} degree poly. $\;k(x, x') = (1 + x^T x')^d$

$d=2$, $p=2$:
$$(1 + x_1 x_1' + x_2 x_2')^2 = 1 + 2x_1x_1' + 2x_2x_2' + x_1^2 x_1'^2 + x_2^2 x_2'^2 + 2x_1x_1'x_2x_2'$$

$$= (1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1x_2, x_1^2, x_2^2)^T \; (\cdots\; x' \;\cdots)$$

$$\hookrightarrow \; \underline{\Phi}(x) = (1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1x_2, x_1^2, x_2^2)$$

<u>thm</u> Every Mercer kernel has a Hi-di embedding

$\underline{\Phi}$ s.t. $\;k(x, x') = \underline{\Phi}(x)^T \underline{\Phi}(x')$

($\underline{\Phi}$ may be infinite dim)

<u>ex</u> RBF in 1d

$$\underline{\Phi}(x) = e^{-x^2/2\sigma^2}\left[1, \sqrt{\tfrac{1}{1!\sigma^2}}\, x, \sqrt{\tfrac{1}{2!\sigma^4}}\, x^2, \ldots\right]$$

Predict new x^*

$$\Phi(x^*)^T\hat{\beta} = \Phi(x^*)^T \underbrace{Z^T}_{\uparrow\text{training}}\hat{\alpha} = \sum_i \hat{\alpha}_i \left(\Phi(x^*)^T\Phi(x_i)\right)$$

$$= \sum_i \hat{\alpha}_i \, k(x^*, x_i)$$

- now predict scales w/ n!