Kernel Trick

Monday, May 8, 2017 10:32 PM

Let $Z_{j\ell} = \varphi_\ell(x_j)$ for any basis then

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

$$(\star)\quad \min_\beta \underbrace{\frac{1}{n}\sum_{i=1}^n (1 - y_i z_i^T \beta)_+}_{\text{function of } Z\beta} + \lambda \|\beta\|_2^2$$

claim $\hat\beta$ solves SVM can be written as $Z^T\alpha$ $\alpha \in \mathbb{R}^n$

Derive kernel trick

$$\beta = \sum_i \alpha_i z_i + \beta^\perp \quad \text{w/} \quad z_i^T \beta^\perp = 0$$

(1) $z_i^T\beta = \sum_j \alpha_j z_i^T z_j + \cancel{z_i^T\beta^\perp} \to 0$

$z_i^T \beta = z_i^T Z^T \alpha$ (I)

$\beta^\perp$ does not impact R_n

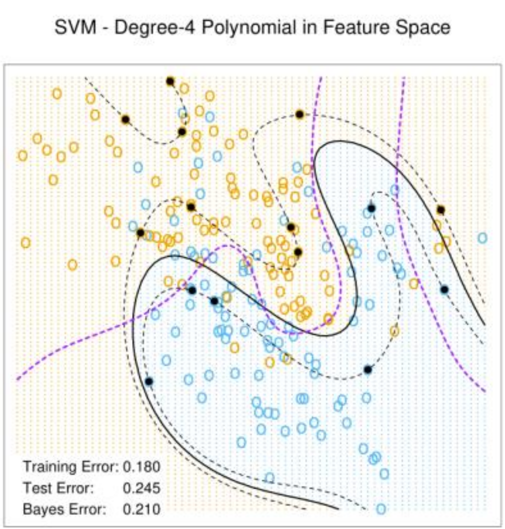
(2) $\|\beta\|_2^2 = \|Z^T\alpha + \beta^\perp\|_2^2 = \|Z^T\alpha\|_2^2 + \underbrace{2\beta^{\perp T} Z^T \alpha}_{(Z\beta^\perp)^T\alpha = 0} + \|\beta^\perp\|_2^2$

$= \|Z^T\alpha\|_2^2 + \|\beta^\perp\|_2^2$ (II)

$(\star)$ is $\min_{\alpha\in\mathbb{R}^n,\ \beta^\perp\in\mathbb{R}^p} \frac{1}{n}\sum_i (1 - y_i \overset{(I)}{z_i^T Z^T \alpha})_+ + \lambda\left(\|Z^T\alpha\|_2^2 + \cancel{\|\beta^\perp\|_2^2}\right) \to 0$

s.t. $\beta^{\perp T} z_i = 0 \ \forall i$

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



Ridge Regression

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

Same for logistic reg. (ridge)

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

$$\text{|||}$$

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

$$\text{|||}$$

$\alpha \in \mathbb{R}^n \quad R_n(y,$

$\parallel\parallel$

$$\min_{\alpha \in \mathbb{R}^n} R_n(y, K\alpha) + \lambda \alpha^T K \alpha \quad \text{where}$$

$$K = ZZ^T$$

$$K_{ij} = z_i^T z_j = \Phi(x_i)^T \Phi(x_j)$$

<u>define</u> kernel $k(x_i, x_j) = \Phi(x_i)^T \Phi(x_j)$

SVM - Radial Kernel in Feature Space

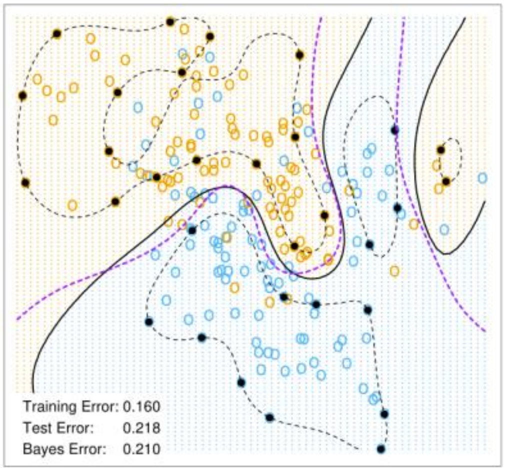


FIGURE 12.3. *Two nonlinear SVMs for the mixture data. The upper plot uses a 4th degree polynomial kernel, the lower a radial basis kernel (with $\gamma = 1$). In each case C was tuned to approximately achieve the best test error performance, and $C = 1$ worked well in both cases. The radial basis kernel 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 function

$k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ 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: $k(x, x') = (1 + x^T x')^d$

$$d = 2: \quad (1 + x_1 x_1' + x_2 x_2')^2 = 1 + 2x_1 x_1' + 2x_2 x_2' + x_1^2 x_1'^2 + x_2^2 x_2'^2 + 2x_1 x_1' x_2 x_2'$$

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

$$\Leftrightarrow \Phi(x) = (1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1 x_2, x_1^2, x_2^2)$$

<u>ex</u> Radial basis function

$$k(x, x') = e^{-\frac{\|x - x'\|_2^2}{\sigma^2}} \qquad \sigma \text{ is bandwidth}$$

<u>thm</u> Every mercer kernel has a Hilbert embedding

Φ (perhaps ∞-dimensional) s.t.

$$k(x, x') = \Phi(x)^T \Phi(x').$$

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

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

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

Prediction new x^*

$$\hat{y} = \begin{cases} 1, & \Phi(x^*)^T\hat{\beta} > 0 \\ 0, & \dots \le 0 \end{cases}$$

$$\hat{\beta} = Z^T\alpha = \sum_i \alpha_i \Phi(x_i)$$

$$\Phi(x^*)^T\hat{\beta} = \sum_i \alpha_i \underbrace{(\Phi(x^*))^T\Phi(x_i)}_{k(x_i,\, x^*)}$$

$$\hat{y} = \begin{cases} 1, & \sum_i \alpha_i k(x_i, x^*) > 0 \\ 0, & \dots \le 0 \end{cases}$$

$\Phi(x) = x \Rightarrow Z = X$

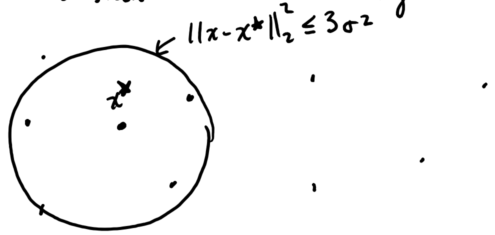
$K = ZZ^T = XX^T \quad (K)_{ij} = x_i^T x_j$

Bandwidth σ: $k(x, x') = e^{-\frac{\|x-x'\|_2^2}{\sigma^2}}$

predict: $\hat{y} = \mathbb{1}\left\{\sum_i \hat{\alpha}_i k(x_i, x^*) > 0\right\}$

σ large: high bias low variance

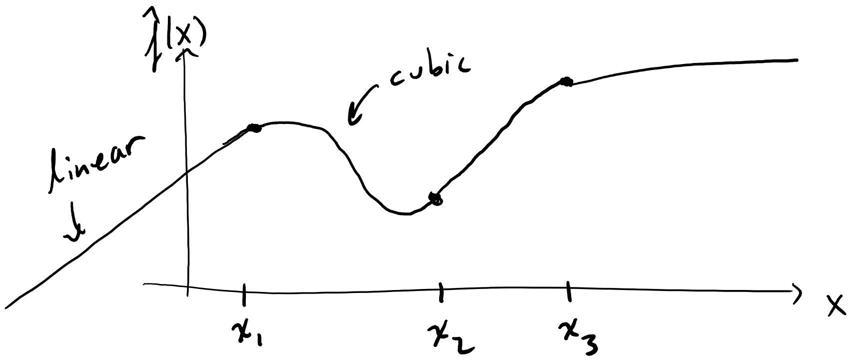
σ small: low bias high variance



Smoothing Splines

Wednesday, May 10, 2017 5:58 PM

$$\min_{f} \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int (f''(t))^2 \, dt$$



Basis is natural cubic splines:

$$N_1(x) = 1 \qquad N_2(x) = x \qquad d_k(x) = \frac{(x - x_k)_+^3 - (x - x_n)_+^3}{x_n - x_k}$$

$$N_{k+2}(x) = d_k(x) - d_{n-1}(x)$$

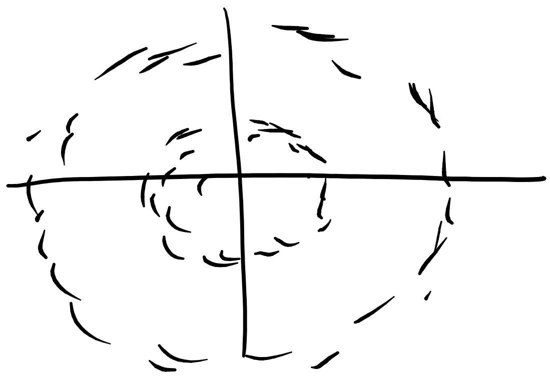
Derive fi-di problem:

Kernel pca

Thursday, May 11, 2017 10:58 AM

SVD of X: $X = U \Sigma V^T$

Apply PCA to $Z = \Phi(X) = U \Sigma V^T$



$K = ZZ^T = U \Sigma V^T V \Sigma^T U^T$

$= U \Sigma \Sigma^T U^T$