Kernel Regression

Recall

Binary classification: $\widehat{y} = \text{sign}(\boldsymbol{x}^T\boldsymbol{w})$

Linear regression: $\widehat{y} = \boldsymbol{x}^T\boldsymbol{w}$

Kernels (in Machine Learning)

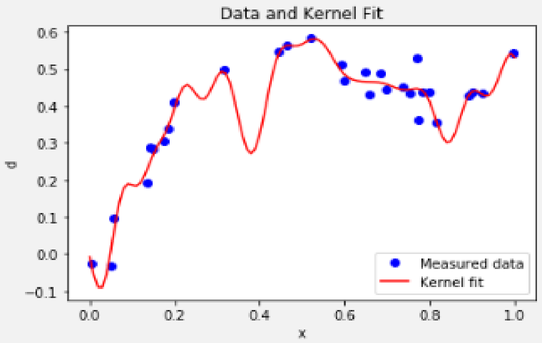
$\boldsymbol{w}$ depends on $\boldsymbol{x}_1, \boldsymbol{y}_1, \boldsymbol{x}_2, \boldsymbol{y}_2 \ldots$

Kernel

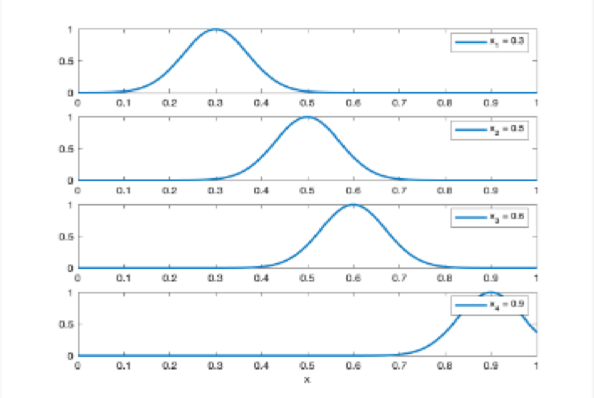
$$\widehat{y} = \phi(\boldsymbol{x})^T\boldsymbol{w} \;\longrightarrow\; \widehat{y} = \sum_i \alpha_i K(\boldsymbol{x}, \boldsymbol{x}_i)$$

High dimensional

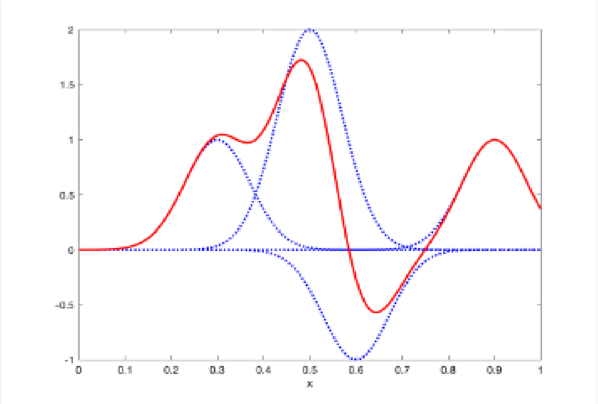
weighted sum of similarities between feature vector and each training point



$$K(x, x_i) = e^{-\frac{(x-x_i)^2}{0.01}}$$



$\widehat{y}(x) = K(x, x_1) + 2K(x, x_2) - K(x, x_3) + K(x, x_4)$



Prediction: $\widehat{y} = \sum_i \alpha_i K(\boldsymbol{x}, \boldsymbol{x}_i)$

How do we find good α_i?

start by finding $\boldsymbol{w}$ using ridge regression

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} ||\boldsymbol{\Phi w} - \boldsymbol{y}|| + \lambda||\boldsymbol{w}||^2$$

$$\boldsymbol{w}^* = (\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\boldsymbol{I})^{-1}\boldsymbol{\Phi}^T\boldsymbol{y}$$

Activity 23, problem 2

$$\boldsymbol{w}^* = \boldsymbol{\Phi}^T(\boldsymbol{\Phi\Phi}^T - \lambda\boldsymbol{I})^{-1}\boldsymbol{y}$$

$$\widehat{y} = \phi(\boldsymbol{x})^T\boldsymbol{w}^* \;\longrightarrow\; \widehat{y} = \sum_i \alpha_i K(\boldsymbol{x}, \boldsymbol{x}_i)$$

$$\boldsymbol{\alpha} = (\boldsymbol{\Phi\Phi}^T + \lambda\boldsymbol{I})^{-1}\boldsymbol{y}$$

where $\boldsymbol{\Phi\Phi}^T$ has i, j entry $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$

No need to compute $\phi(\cdot)$ to compute $K(\cdot,\cdot)$ or $\widehat{y}$!

Kernel can have a few hyper-parameters.

Danger! ➔ Overfitting

Use cross-validation to choose params.