Kernels for classification

Kernel regression (last class):

$$\widehat{y} = \phi(x)^T w \implies \widehat{y} = \sum_i \alpha_i K(x, x_i)$$

High-dimensional feature transformation

Kernel: measures similarity between x and x_i .

Kernels for binary classification (today):

Classification, after feature map:

$$\widehat{y} = \operatorname{sign}(\phi(\boldsymbol{x})^T \boldsymbol{w})$$
 (1)
 \boldsymbol{w} depends on $x_1, y_1, x_2, y_2...$

Kernel methods - re-write above as:

$$\widehat{y} = \operatorname{sign}\left(\sum_{i} \alpha_{i} K(\boldsymbol{x}, \boldsymbol{x}_{i})\right) \tag{2}$$

weighted sum of similarities between feature vector and each training point Representer Theorem: (1) and (2) are the same, when

$$egin{aligned} oldsymbol{w}^* &= rg \min_{oldsymbol{w}} ||\Phi oldsymbol{w} - oldsymbol{y}|| + \lambda ||oldsymbol{w}||^2 \ oldsymbol{w}^* &= (oldsymbol{\Phi}^T oldsymbol{\Phi} + \lambda oldsymbol{I})^{-1} oldsymbol{\Phi}^T oldsymbol{y} \end{aligned}$$

$$\alpha = (K + \lambda I)^{-1} y$$

where K has ℓ, m entry $K(x_{\ell}, x_m)$

[Kimeldorf 1970]

Example of kernel classification

How do we predict class of x?

$$m{x} = egin{bmatrix} 0.3 \ 0.1 \end{bmatrix}$$

$$K(oldsymbol{x},oldsymbol{x}_i) = \exp\left(-\left|\left|oldsymbol{x}-oldsymbol{x}_i
ight|
ight|^2
ight)$$

$$\widehat{y} = \operatorname{sign}\left(\sum_{i} \alpha_{i} \exp\left(-\left|\left|\begin{bmatrix} 0.3 \\ 0.1 \end{bmatrix} - \boldsymbol{x}_{i}\right|\right|^{2}\right)\right)$$

