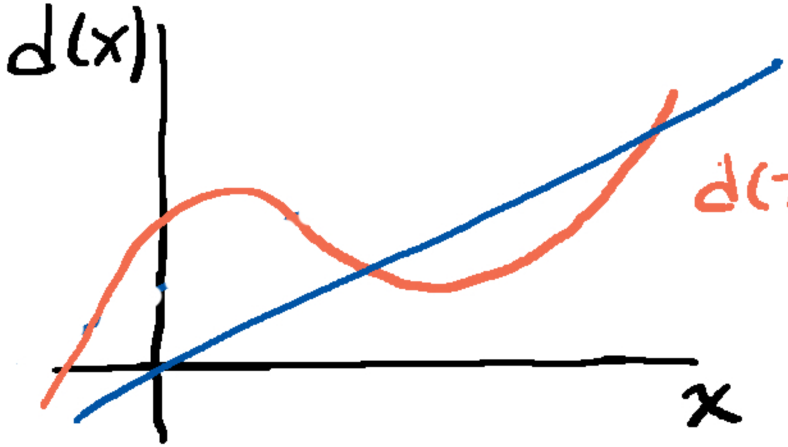
# Kernel Regression

# Objectives

- Why use higher-dimensional feature spaces
- Reformulate regression in terms of kernels
- Popular kernels
- Cautions and considerations

# Higher dimensional feature spaces extend regression






$d(x) = w_1 x$

$d(x) = w_3 x^3 + w_2 x^2 + w_1 x + w_0$

Let $\underline{x} = [x_1 \ x_2 \ \cdots \ x_M]^T \in \mathbb{R}^M$

Consider $d(\underline{x}) = \underline{\phi}^T(\underline{x})\,\underline{w}$, $\phi(\underline{x}) \in \mathbb{R}^P$, $P > M$

Example: $\underline{x} = [x_1 \ x_2]^T$, $\underline{\phi}^T(\underline{x}) = [x_1^2 \ x_2^2 \ \sqrt{2}x_1x_2 \ x_1 \ x_2 \ 1]$

Finding $\underline{w}$: "training" data $\underline{x}^i, d^i, i = 1, 2, \cdots N$

$$\min_{\underline{w}} \sum_{i=1}^{N} (d^i - \underline{\phi}^T(\underline{x}^i)\,\underline{w})^2 + \lambda \|\underline{w}\|_2^2 \quad \text{(Ridge)}$$

$\underline{d} = [d^1 \ d^2 \ \cdots \ d^N]^T$

$$\min_{\underline{w}} \|\underline{d} - \underline{\Phi}\,\underline{w}\|_2^2 + \lambda \|\underline{w}\|_2^2$$

$\underline{\Phi} = [\phi(\underline{x}^1) \ \phi(\underline{x}^2) \cdots \phi(\underline{x}^N)]^T$ $(N \times P)$ $\Rightarrow$ $\underline{w} = (\underline{\Phi}^T\underline{\Phi} + \lambda \underline{I})^{-1}\underline{\Phi}^T\underline{d}$

# Regression is a weighted sum of "kernels" 3

$$d(\underline{x}) = \underline{\phi}^T(\underline{x})\,\underline{w} = \underline{\phi}^T(\underline{x})\underbrace{(\underline{\Phi}^T\underline{\Phi}+\lambda\underline{I})}_{P\times P}{}^{-1}\underline{\Phi}^T\underline{d}$$

Matrix identity: $(\underline{\Phi}^T\underline{\Phi}+\lambda\underline{I})^{-1}\underline{\Phi}^T = \underline{\Phi}^T(\underline{\Phi}\,\underline{\Phi}^T+\lambda\underline{I})^{-1}$ (activity)

Thus $d(\underline{x}) = \underline{\phi}^T(\underline{x})\,\underline{\Phi}^T\underbrace{(\underline{\Phi}\,\underline{\Phi}^T+\lambda\underline{I})}_{N\times N}{}^{-1}\underline{d}$

Note: $[\underline{\Phi}\,\underline{\Phi}^T]_{i,j} = \underline{\phi}^T(\underline{x}^i)\,\underline{\phi}(\underline{x}^j)$

$[\underline{\phi}^T(\underline{x})\,\underline{\Phi}^T]_j = \underline{\phi}^T(\underline{x})\,\underline{\phi}(\underline{x}^j)$

Define "Kernel"
$K(\underline{u},\underline{v}) = \underline{\phi}^T(\underline{u})\,\underline{\phi}(\underline{v})$

Let $\underline{\alpha} = [\alpha_1 \;\cdots\; \alpha_N]^T$
$= (\underline{\Phi}\,\underline{\Phi}^T+\lambda\underline{I})^{-1}\underline{d}$

$$d(\underline{x}) = \sum_{i=1}^{N}\alpha_i\,\underline{\phi}^T(\underline{x})\,\underline{\phi}(\underline{x}^i) = \underline{\sum_{i=1}^{N}\alpha_i K(\underline{x},\underline{x}^i)}$$

# Kernel methods find $d(\underline{x})$ without computing $\phi(\underline{x})$ 

$$d(\underline{x}) = \sum_{i=1}^{N} \alpha_i K(\underline{x}, \underline{x}^i) \qquad \underline{\alpha} = (\underline{\Phi}\,\underline{\Phi}^T + \lambda \underline{I})^{-1} \underline{d} \rightarrow \underline{K} = \underline{\Phi}\,\underline{\Phi}^T$$

$$= (\underline{K} + \lambda \underline{I})^{-1} \underline{d}$$

$$[\underline{K}]_{ij} = \underline{\phi}^T(\underline{x}^i)\underline{\phi}(\underline{x}^j) = K(\underline{x}^i, \underline{x}^j)$$

$\underline{K}$ can be computed efficiently!

Ex: Monomials of degree $g$ $\quad \underline{\phi}(\underline{x}) \rightarrow \underbrace{x_1^g,\ x_1^{g-1}x_2,\ \ldots x_3^{g-5}x_6^2x_8^3 \ldots}$

$$K(\underline{u}, \underline{v}) = \underbrace{\underline{\phi}^T(\underline{u})\,\underline{\phi}(\underline{v})}_{O(P)} = (\underbrace{\underline{u}^T\underline{v}}_{O(M)})^g \quad \text{(activity)} \qquad P = \frac{(g+M-1)!}{g!\,(M-1)!} \text{ terms}$$

Suppose $M = 10,\ g = 5 \rightarrow P \sim 2000$ $\qquad$ computing $O(P)$ vs $O(M)$

$M = 100,\ g = 5 \rightarrow P \sim 10^8$ $\qquad$ memory $O(NP)$ vs $O(N^2)$

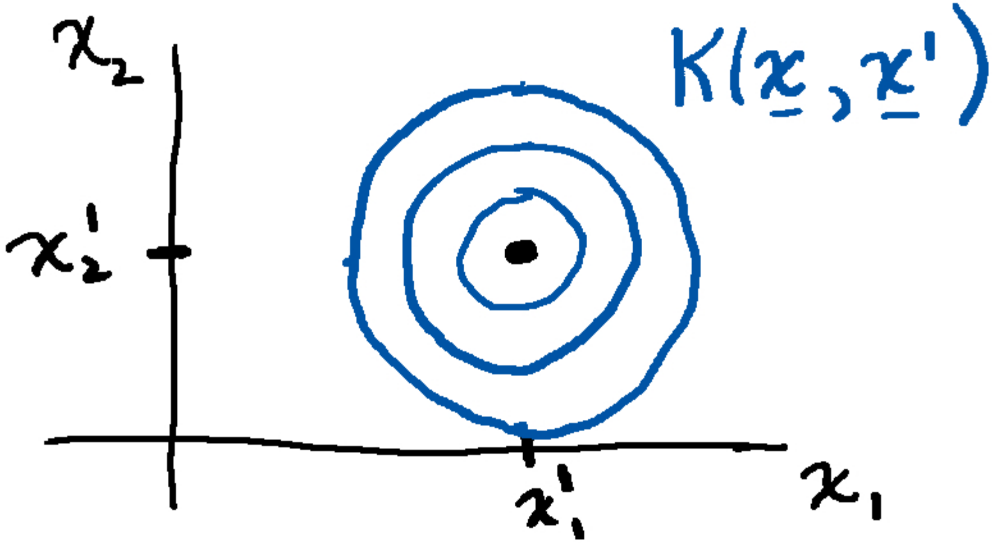
# Popular kernels depend on similarity of $\underline{u}, \underline{v}$ 

$$\underline{u}^T \underline{v} = \|\underline{u}\|_2 \|\underline{v}\|_2 \cos\theta$$

Monomials of degree $q$: $K(\underline{u},\underline{v}) = (\underline{u}^T\underline{v})^q$

Polynomials up to degree $q$: $K(\underline{u},\underline{v}) = (\underline{u}^T\underline{v}+1)^q$

Gaussian/radial kernel: $K(\underline{u},\underline{v}) = \exp\left\{-\frac{\|\underline{u}-\underline{v}\|_2^2}{2\sigma^2}\right\}$

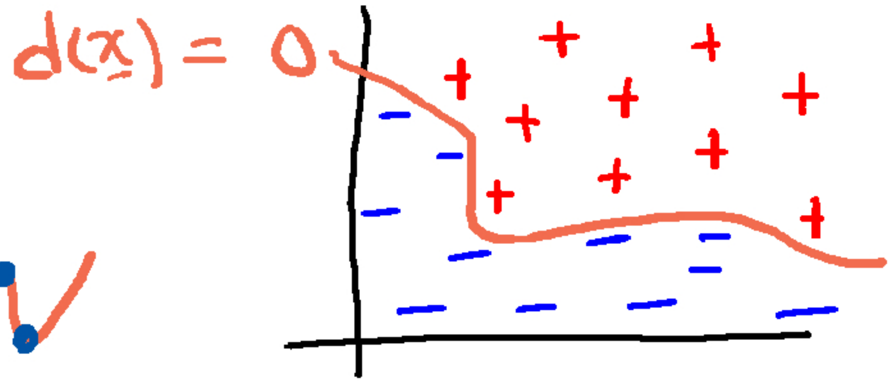



- No explicit $\underline{\phi}(\underline{x})$
- All polynomial orders
- smoothness controlled by $\sigma$

# Kernel regression considerations

$$d(\underline{x}) = \underline{\phi}^T(\underline{x})\,\underline{w} \quad \text{vs} \quad d(\underline{x}) = \sum_{i=1}^{N} \alpha_i \, k(\underline{x}, \underline{x}^i)$$

- Store and compute $\underline{\alpha}$ $(N \times 1)$ vs $\underline{w}$ $(P \times 1)$
- Binary classification $\text{sign}\{d(\underline{x})\}$




- Avoid "overfitting" with high-D feature spaces

(cross-validation)