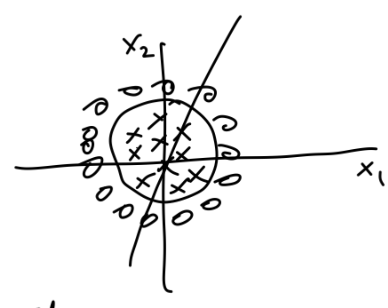
Kernels



add more features

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \to \begin{pmatrix} x_1 \\ x_2 \\ x_1^2 + x_2^2 \end{pmatrix}$$

$$\underset{\in \mathbb{R}^d}{x} \xrightarrow{\phi} \phi(x) \in \mathbb{R}^D$$

$D \gg d$

Adv
- opt. stays convex, well-behaved

Disadv
- ϕ could be very high-dim.!

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} \qquad \phi(x) = \begin{pmatrix} 1 \\ x_1 \\ \vdots \\ x_d \\ x_1 x_2 \\ \vdots \\ x_{d-1} x_d \\ x_1 x_2 x_3 \\ \vdots \\ x_{d-2} x_{d-1} x_d \\ \vdots \\ x_1 x_2 \cdots x_d \end{pmatrix}$$

Quiz: What is dim of ϕ?

Inner Product Computation

$$\phi(x) = \begin{pmatrix} 1 \\ x_1 \\ \vdots \\ x_d \\ x_1 x_2 \\ \vdots \\ x_{d-1} x_d \\ \vdots \\ x_1 x_2 \cdots x_d \end{pmatrix}$$

$$\phi^T(x)\phi(z) = \overbrace{1 \cdot 1 + x_1 z_1 + x_2 z_2 + \cdots + x_1 x_2 z_1 z_2 + \cdots + x_1 \cdots x_d z_1 \cdots z_d}^{2^d} \quad O(2^d)$$

$$= \prod_{k=1}^{d} (1 + x_k z_k) \quad O(d)$$

$$K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$$

$$K_{ij} = K(x_i, x_j)$$

Kernel Trick

Gradient Descent w/ Squared Loss

$$K_{ij} = x_i^T x_j$$

$$\ell(w) = \sum_{i=1}^{n} (w^T x_i - y_i)^2$$

$$w = \sum_{i=1}^{n} \alpha_i x_i$$

$$\boxed{\ell(\alpha) = \sum_{i=1}^{n} \Big(\sum_{j=1}^{n} \alpha_j \underbrace{x_j^T x_i}_{K_{ji}} - y_i \Big)^2}$$

Claim!

Proof

$$w_0 = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \qquad \alpha_1^0 = \cdots = \alpha_n^0 = 0$$

$$w_1 = w_0 - s \sum_{i=1}^{n} \underbrace{2(w_0^T x_i - y_i)}_{\gamma_i^0} x_i$$

$$= \sum_{i=1}^{n} \alpha_i^0 x_i - s \sum_{i=1}^{n} \gamma_i^0 x_i = \sum_{i=1}^{n} \alpha_i^1 x_i \qquad \alpha_i^1 = \alpha_i^0 - s\gamma_i^0$$

$$w_2 = w_1 - s \sum_{i=1}^{n} \underbrace{2(w_1^T x_i - y_i)}_{\gamma_i^1} x_i$$

$$= \sum_{i=1}^{n} \alpha_i^1 x_i - s \sum_{i=1}^{n} \gamma_i^1 x_i = \sum_{i=1}^{n} \alpha_i^2 x_i \qquad \alpha_i^1 - s\gamma_i^1$$

$$w_t = \sum_{i=1}^{n} \alpha_i^{t-1} x_i - s \sum_{i=1}^{n} \gamma_i^{t-1} x_i = \sum_{i=1}^{n} \alpha_i^t x_i \qquad \alpha_i^{t-1} - s\gamma_i^{t-1}$$

$$\alpha_i^t = -s \sum_{r=0}^{t-1} \gamma_i^r$$

1. Predictions

$$w = \sum_{i=1}^{n} \alpha_i \phi(x_i)$$

$$w^T \phi(x') = \sum_{i=1}^{n} \alpha_i \underbrace{\phi(x_i)^T \phi(x')}_{K(x_i, x')}$$

$$\ell(\alpha) = \sum_{i=1}^{n} \Big(\sum_{j=1}^{n} \alpha_j \underbrace{x_j^T x_i}_{K_{ij}} - y_i \Big)^2$$

2. Learning

$$\alpha_i^t = \alpha_i^{t-1} - s\gamma_i^{t-1} \qquad \gamma_i^{t-1} = 2\Big(\sum_{j=1}^{n} \alpha_j K_{ij} - y_i\Big)$$

Kernel Functions

linear $K(x, z) = x^T z$

polynomial $K(x, z) = (1 + x^T z)^d$

Radial Basis Function (RBF) (Gaussian Kernel) $K(x, z) = e^{-\frac{\|x - z\|^2}{\sigma^2}}$

exponential Kernel $K(x, z) = e^{-\frac{\|x - z\|}{2\sigma^2}}$

Laplacian Kernel $K(x, z) = e^{-\frac{|x - z|}{2\sigma}}$

Sigmoid Kernel $K(x, z) = \tanh(x^T z + c)$

$\boxed{\text{Can we use any } K \to \mathbb{R}?}$

No!

Def: A matrix $A \in \mathbb{R}^{n \times n}$ is positive semi-definite iff $\forall q \in \mathbb{R}^n$, $q^T A q \geq 0$

Lemma: $K_{ij} = \phi(x_i)\phi(x_j)$ is p.s.d.

Proof

① $K = \Phi^T \Phi$, $\Phi = [\phi(x_1) \cdots \phi(x_n)]^T$

$$q^T K q = (\Phi^T q)^2 \geq 0$$

② $A = BB \to$ p.s.d