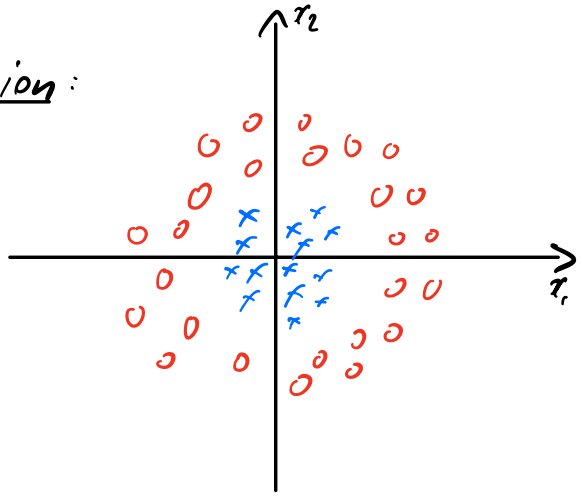
Kernels

Explicit Feature Expansion:



Quiz: Add new feature(s) such that this data set becomes linearly separable.

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \rightarrow \begin{pmatrix} x_1 \\ x_2 \\ ? \end{pmatrix}$$

$\vec{x}$ $\quad$ $\phi(\vec{x})$

Adding features can reduce Bias (but increase Variance).
Allows us to capture non-linear interactions between original features.

Problem in Practice: It is not obvious what new features to construct.

Case I:

Polynomial Features:

$\vec{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ ⟹ quadratic expansion ⟹ $\phi(\vec{x}) = \begin{pmatrix} 1 \\ x_1 \\ x_2 \\ x_1^2 \\ x_2^2 \\ x_1 x_2 \end{pmatrix}$ ← constant; } linear features; } quadratic features; } cross product

$$f(x) = \theta^T \hat{x} = w^T x + x^T M x + b$$

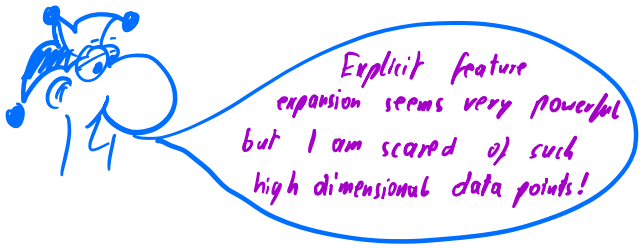
linear; quadratic & cross products; constant ⟸ quadratic function

Case II:

All possible interactions:

$\vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \rightarrow \phi(\vec{x}) = \begin{pmatrix} 1 \\ x_1 \\ x_2 \\ x_3 \\ x_1 x_2 \\ x_1 x_3 \\ x_2 x_3 \\ x_1 x_2 x_3 \end{pmatrix}$

Quiz: If $x \in \mathbb{R}^d$, what is the dimensionality of $\phi(\vec{x})$?



Solution: Implicit feature expansion (Aka Kernels)

Assume, all you want to do is compute inner products.

$\vec{x}^T \vec{z} = \sum_{\alpha=1}^{d} x_\alpha z_\alpha$ how about $\phi(\vec{x})^T \phi(\vec{z})$?

Case I:

$$\phi(\vec{x})^T \phi(\vec{z}) = (1 + \vec{x}^T \vec{z})^2$$

Case II:

$$\phi(\vec{x})^T \phi(\vec{z}) = \prod_{\alpha=1}^{d} (1 + x_\alpha z_\alpha)$$

⟹ Never compute $\vec{x} \rightarrow \phi(\vec{x})$.

Only compute $\vec{x}^T \vec{z} \rightarrow \phi(\vec{x})^T \phi(\vec{z}) = k(\vec{x}, \vec{z})$ ← kernel function

This is just the inner-product function in the high dimensional feature space.

ERM* can be written entirely in terms of inner products?

* At least the losses that we discussed.

Classifier: $h(\vec{x}) = w^T \phi(\vec{x}) + b$ Loss: $L = \sum_{i=1}^{n} \ell(w^T \phi(\vec{x}_i), y_i)$

Claim: $\vec{w} = \sum_{i=1}^{n} \beta_i \phi(\vec{x}_i)$ $\Leftarrow$ $\vec{w}$ lies in the span of $\phi(\vec{x}_1), \dots, \phi(\vec{x}_n)$

$\Rightarrow$ at test time: $h(\vec{z}) = w^T \phi(\vec{z}) + b = \sum_{i=1}^{n} \beta_i \phi(\vec{x}_i)^T \phi(\vec{z}) + b$

$\vec{z}$: test point

only inner products

Proof by induction: We learn with gradient descent. Induction over update t.

Base case $t=0$: We initialize $\vec{w}_0 = \vec{0} = \sum_{i=1}^{n} 0 \cdot \phi(\vec{x}_i)$

Assume $\vec{w}^t = \sum_{i=1}^{n} \beta_i^t \phi(\vec{x}_i)$

Gradient update: $\vec{w}_{t+1} \leftarrow \vec{w}_t - \alpha \frac{\partial L}{\partial \vec{w}_t}$

$\frac{\partial L}{\partial \vec{w}} = \sum_{i=1}^{n} \frac{\partial \ell(\vec{w}^T \phi(\vec{x}_i), y_i)}{\partial \vec{w}} = \sum_{i=1}^{n} \ell'(\vec{w}^T \phi(\vec{x}_i), y_i) \cdot \phi(\vec{x}_i)$ (chain rule; γ_i^{t+1}: scalar)

$= \sum_{i=1}^{n} \beta_i^t \phi(\vec{x}_i) - \alpha \sum_{i=1}^{n} \gamma_i^{t+1} \phi(\vec{x}_i) = \sum_{i=1}^{n} (\beta_i^t - \alpha \gamma_i^{t+1}) \phi(\vec{x}_i)$

β_i^{t+1}

$\Rightarrow \beta_i^{t+1} \leftarrow \beta_i^t - \alpha \gamma_i^{t+1}$

□

$\Rightarrow h(\vec{z}) = \sum_{i=1}^{n} \beta_i k(\vec{x}_i, \vec{z}) + b$ ← learning of b is not affected.

we now learn $\vec{\beta} \in \mathbb{R}^n$ instead of $\vec{w} \in \mathbb{R}^d$

Popular Kernels:

Linear: $k(\vec{x}, \vec{z}) = \vec{x}^T \vec{z}$

Polynomial: $k(\vec{x}, \vec{z}) = (1 + \vec{x}^T \vec{z})^p$

Radial Basis Function (RBF): $k(\vec{x}, \vec{z}) = \exp(-\lambda(\vec{x} - \vec{z})^2)$

Sigmoid: $k(\vec{x}, \vec{z}) = \tanh(a x^T + c)$

$K \in \mathbb{R}^{m \times m}$

For any finite data set $D = \{\vec{x}_1, \ldots, \vec{x}_m\}$ define $K_{ij} := k(\vec{x}_i, \vec{x}_j)$.

K must be p.s.d.

A symmetric matrix $K \in \mathbb{R}^{n \times n}$ is p.s.d. if and only if

1. $\exists A \in \mathbb{R}^{n \times n}$ s.t. $K = A^T A$
2. All eigenvalues of K are non-negative ⇕
3. $\forall q \in \mathbb{R}^n \quad q^T K q \geq 0$ ⇕