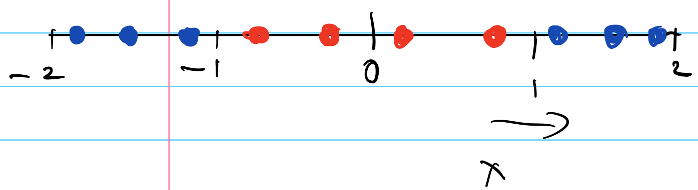
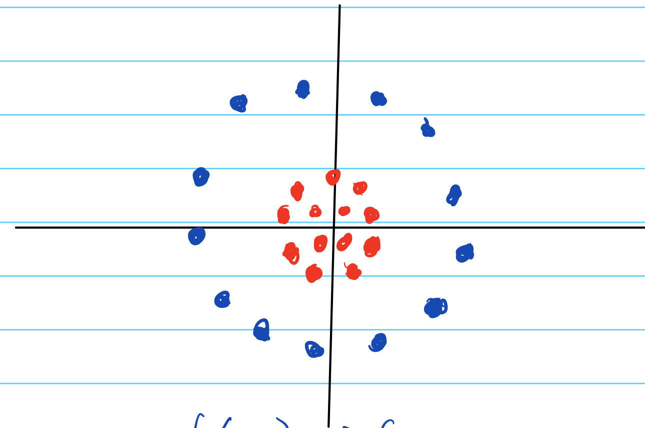
# CS3780/5780
# Kernel Method



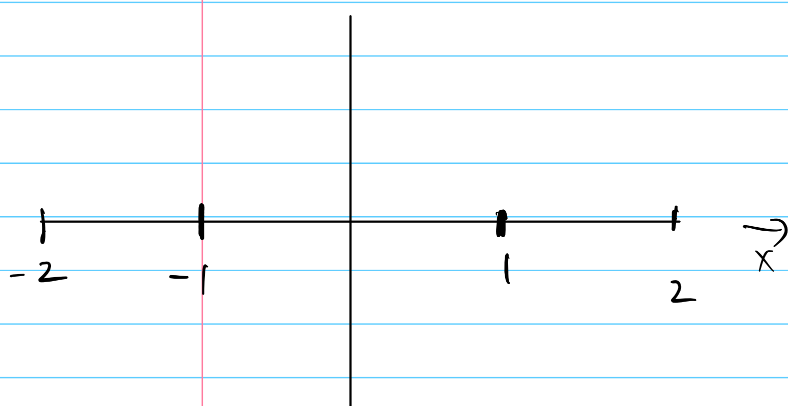


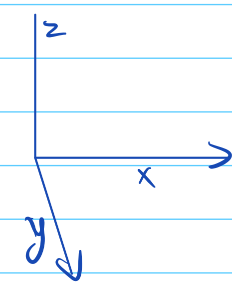
$x \rightarrow (x, x_2, x_3, \ldots)$

$(x_1, x_2) \rightarrow (x_1, x_2, x_3, x_4)$

Can we add features so that data becomes linearly separable??

**Magic 1:**





**Problem in Practice: Not obvious what features to add/construct**

Enumerate as many features as possible.

**Eg 1. Polynomial features**

$$\vec{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \xrightarrow{\text{quadratic features}} \phi(\vec{x}) = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_1^2 \\ x_2^2 \\ x_1 x_2 \end{bmatrix}$$

**Eg 2. all interactions**

$$\vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \xrightarrow{\text{interactions}} \phi(\vec{x}) = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_3 \\ x_1 x_2 \\ x_2 x_3 \\ x_1 x_3 \\ x_1 x_2 x_3 \end{bmatrix}$$

Problem: quickly gets very high dimensional

Quiz: If $\vec{x}$ is d dimensional, then:

1. For example 1, if we want all quadratic functions, then what is the dimensionality of $\phi(\vec{x})$
2. In Example 1, if we want all degree p polynomial functions, then what is the dimensionality of $\phi(\vec{x})$
3. In Example 2, what is the dimensionality of

Magic 2: Implicit feature expansion

Never explicitly compute feature map $\phi(\vec{x})$

Only directly compute inner product $\phi(x_1)^T\phi(x_2)$ given $x_1, x_2$

$$\phi(x_1)^T\phi(x_2) = k(x_1, x_2)$$

Kernel function : $k(x_1, x_2)$

Eg 1 $\quad \phi(x)^T\phi(y) = (1 + x^Ty)^2 = k(x, y)$ $\qquad$ what is $\phi(x)$ ?

Eg 2 $\quad \phi(x)^T\phi(y) = \prod_{\alpha=1}^{d} (1 + x_\alpha y_\alpha) = k(x, y)$

How do we use this kernel trick?

SVM:

Minimize $\sum_{i=1}^{n} \max(0, 1 - y_i w^T \phi(x_i)) + \frac{1}{C} ||w||_2^2$

Logistic Regression:

Minimize $\sum_{i=1}^{n} \log(1 + \exp(-y_i \, w^T \phi(x_i))) + \frac{\lambda}{2} ||w||_2^2$

Linear Regression:

Minimize $\sum_{i=1}^{n} (w^T \phi(x_i) - y_i)^2 + \frac{\lambda}{2} ||w||_2^2$

More generally:

$$L(w) = \sum_{i=1}^{n} \text{loss}(w^T \phi(x_i), y_i) + \frac{\lambda}{2} ||w||_2^2$$

Claim: w that minimizes L(w) admits form

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

For some $\alpha_1 \cdots \alpha_n \in \mathbb{R}$

ie $w$ in span of $\phi(x_1) \cdots \phi(x_n)$

Say $w = w_D + w_\perp$

$w_D$ in span of data

$w_\perp$ $\perp$ to subspace containing data

$$\forall i, \quad w^T \phi(x_i) = w_D^T \phi(x_i)$$

$$\|w\|_2^2 = \|w_D\|_2^2 + \|w_\perp\|_2^2$$

Hence

$$L_D(w) = L_D(w_D) + \frac{\lambda}{2}\|w_\perp\|_2^2 \geq L_D(w_D)$$

Hence minimizer of $L(w)$ will be in span of Data

What does this buy us? w is still very high dim (even $\infty$)

But, for a new point x,

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

$$= \sum_{i=1}^{n} \alpha_i \, k(x_i, x)$$

Hence is we had the $\alpha$'s, then we can compute prediction for any new x using only kernel function

$$L(\alpha) = \underset{\alpha_1 \cdots \alpha_n}{\text{Minimize}} \sum_{i=1}^{n} loss\Big(\underbrace{\sum_{j=1}^{n} \alpha_j K_{j,i}}_{\alpha^T K_{\cdot,i}}, y_i\Big) + \underbrace{\frac{\lambda}{2} \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j K_{ij}}_{\frac{\lambda}{2}\alpha^T K \alpha}$$

K : Kernel Matrix $\quad K_{ij} = k(x_i, x_j)$

Next lecture:

1. how do we find alpha's
2. What functions are kernels? How do we make kernels?