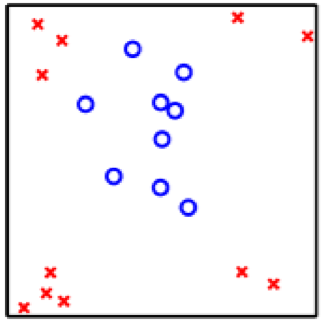
Machine Learning from Data

Lecture 26: Spring 2021

Today’s Lecture

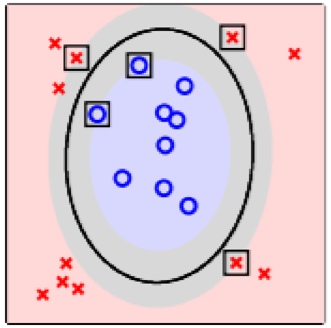
- Kernel Machines
  - Popular Kernels
  - The Kernel Measures Similarity
  - Kernels in Different Applications

RECAP: The Kernel Allows Us to Bypass $\mathcal{Z}$-space



$\mathbf{x}_n \in \mathcal{X}$

$\Big\downarrow \; K(\cdot,\cdot)$



$$g(\mathbf{x}) = \text{sign}\left(\sum_{\alpha_n^*>0} \alpha_n^* y_n K(\mathbf{x}_n, \mathbf{x}) + b^*\right)$$

$$b^* = y_s - \sum_{\alpha_n^*>0} \alpha_n^* y_n K(\mathbf{x}_n, \mathbf{x}_s)$$

(One can compute b^* for several SVs and average)

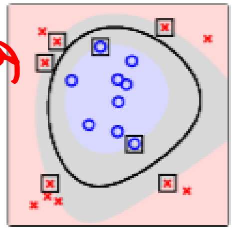
Solve the QP

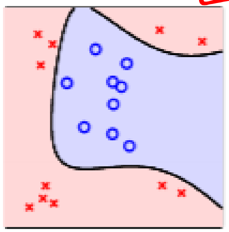
$$\left.\begin{array}{ll} \underset{\boldsymbol{\alpha}}{\text{minimize}} & \frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\mathrm{G}\boldsymbol{\alpha} - \mathbf{1}^{\mathrm{T}}\boldsymbol{\alpha} \\ \text{subject to:} & \mathbf{y}^{\mathrm{T}}\boldsymbol{\alpha} = 0 \\ & \mathbf{C} \geq \boldsymbol{\alpha} \geq \mathbf{0} \end{array}\right\} \longrightarrow \begin{array}{l} \boldsymbol{\alpha}^* \\ \text{index } s : C > \alpha_s^* > 0 \end{array}$$

$\uparrow$ free support vectors

Overfitting

SVM | Pseudo-inverse





Eout

high $\tilde{d} \rightarrow$ complicated separator

few support vectors $\rightarrow$ low effective complexity

Can go to high (infinite) $\tilde{d}$

Computation

Inner products with Kernel

$K(\cdot,\cdot)$

high $\tilde{d} \rightarrow$ expensive or infeasible computation

kernel $\rightarrow$ computationally feasible to go to high $\tilde{d}$

Can go to high (infinite) $\tilde{d}$

Polynomial Kernel

2nd-Order Polynomial Kernel

$$\Phi(\mathbf{x}) = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \\ x_1^2 \\ x_2^2 \\ \vdots \\ x_d^2 \\ \sqrt{2}x_1x_2 \\ \sqrt{2}x_1x_3 \\ \vdots \\ \sqrt{2}x_1x_d \\ \sqrt{2}x_2x_3 \\ \vdots \\ \sqrt{2}x_{d-1}x_d \end{bmatrix}$$

$$\begin{aligned} K(\mathbf{x}, \mathbf{x}') &= \Phi(\mathbf{x})^{\mathrm{T}}\Phi(\mathbf{x}') \\ &= \sum_{i=1}^{d} x_i x_i' + \sum_{i=1}^{d} x_i^2 x_i'^2 + 2\sum_{i<j} x_i x_j x_i' x_j' \qquad \leftarrow O(d^2) \\ &= \boxed{\left(\frac{1}{2} + \mathbf{x}^{\mathrm{T}}\mathbf{x}'\right)^2 - \frac{1}{4}} \end{aligned}$$

$\uparrow$
computed quickly
in $\mathcal{X}$-space, in $O(d)$

Q is large

Q-th order polynomial kernel

$$K(\mathbf{x}, \mathbf{x}') = (r + \mathbf{x}^{\mathrm{T}}\mathbf{x}')^{Q} \qquad \leftarrow \text{inhomogeneous kernel}$$

$$K(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^{\mathrm{T}}\mathbf{x}')^{Q} \qquad \leftarrow \text{homogeneous kernel}$$

RBF-Kernel

Gaussian kernel

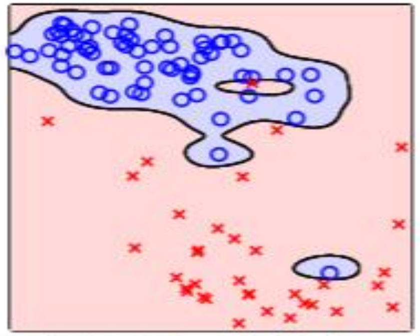
One dimensional RBF-Kernel

$$\Phi(x) = e^{-x^2} \begin{bmatrix} 1 \\ \sqrt{\frac{2^1}{1!}}x \\ \sqrt{\frac{2^2}{2!}}x^2 \\ \sqrt{\frac{2^3}{3!}}x^3 \\ \sqrt{\frac{2^4}{4!}}x^4 \\ \vdots \end{bmatrix}$$

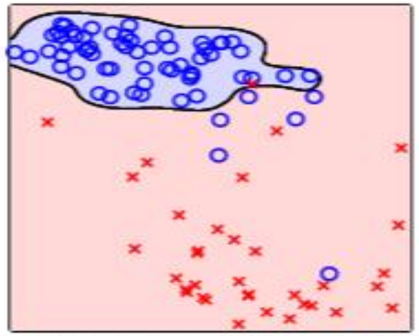
$$\begin{aligned} K(\mathbf{x}, \mathbf{x}') &= \Phi(\mathbf{x})^{\mathrm{T}}\Phi(\mathbf{x}') \\ &= e^{-x^2}e^{-x'^2}\sum_{i=0}^{\infty}\frac{(2xx')^i}{i!} \quad \leftarrow \text{not feasible} \\ &= e^{-x^2}e^{-x'^2}e^{2xx'} \\ &= \underbrace{e^{-(x-x')^2}}_{\substack{\uparrow \\ \text{computed quickly} \\ \text{in } \mathcal{X}\text{-space, in } O(d)}} \end{aligned}$$

d-dimensional RBF-Kernel

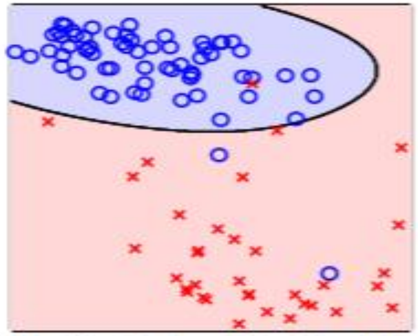
$$\boxed{K(\mathbf{x}, \mathbf{x}') = e^{-\gamma\|\mathbf{x}-\mathbf{x}'\|^2}} \qquad (\gamma > 0)$$



Hard Margin ($\gamma = 2000$, $C = \infty$)



Soft Margin ($\gamma = 2000$, $C = 0.25$)



Soft Margin ($\gamma = 100$, $C = 0.25$)

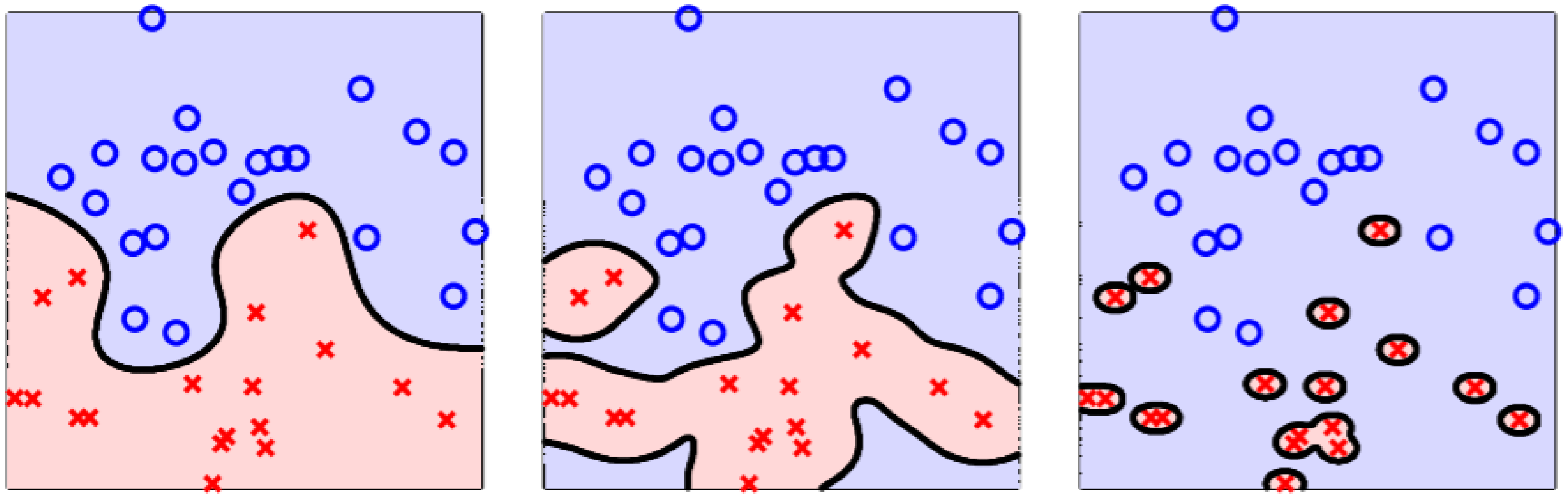
Choosing RBF-Kernel $\boxed{\text{Width } \gamma}$

$$e^{-\gamma \| \mathbf{x} - \mathbf{x}' \|^2}$$

C :

γ

→ cross validation



Small γ | Medium γ | Large γ!

$$K(x, x') = e^{-\|x - x'\|^2} \qquad \gamma = 1$$

$$\underline{\underline{g(x)}} = \text{sign}\left(\sum_{\alpha_n^* > 0} \alpha_n^* y_n K(\underline{\underline{x_n}}, x) + b^*\right) \qquad \alpha \geq 0$$

$$g(x) = \text{sign}\left(\sum_{\alpha_n^* > 0} \underbrace{\alpha_n^* y_n}_{w_j} \underbrace{e^{-\|x_n - x\|^2}}_{K} + \underline{\underline{b}}^*\right) \qquad \alpha_n = 0$$

$$\alpha_n^*$$

RBF Network

$$g(x) = \text{sign}\left(\underline{\underline{w_0}}^{b^*} + \sum_{j=1}^{K} w_j \cdot \underbrace{e^{-\|x - \mu_j\|^2}}_{K}\right)$$

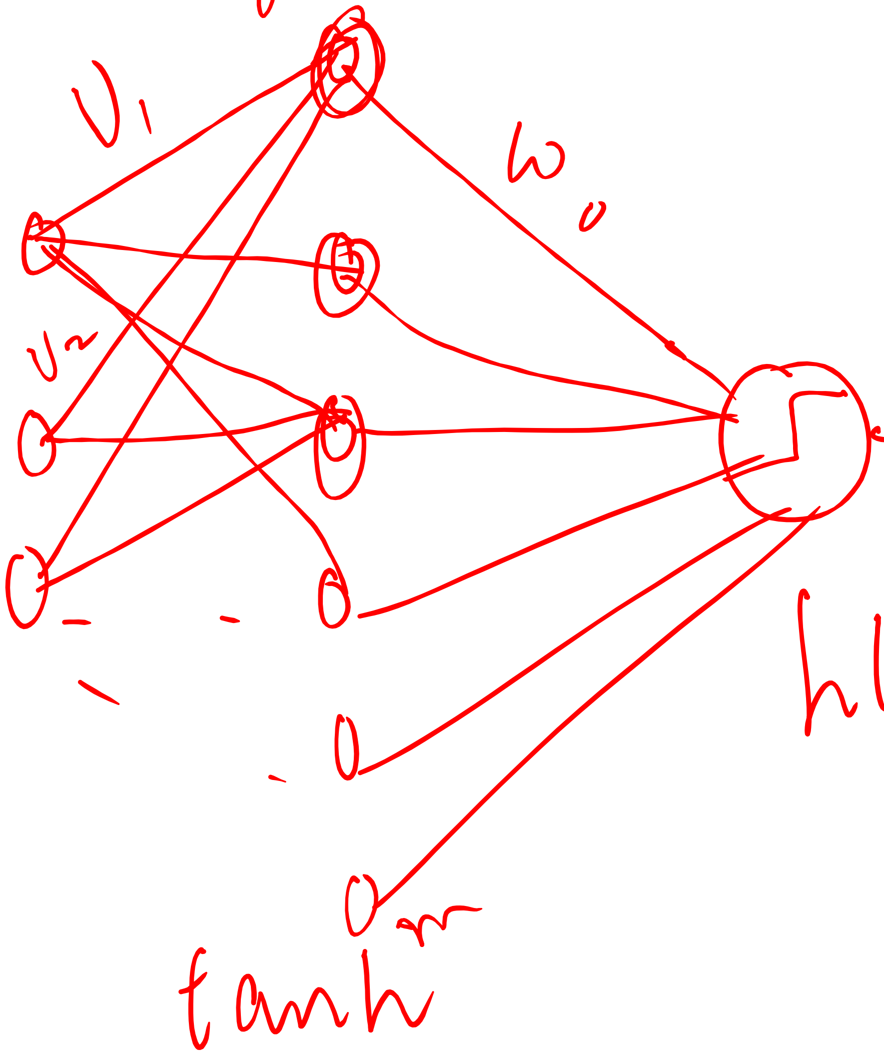
Similarity measure

$$\underbrace{e^{-\|x - x_j'\|^2}}_{\text{fit of linear model.}} \rightarrow \boxed{\text{constructed}}$$

$w_0, w_j \ldots \rightarrow$ fit of linear model

Centers $\Longleftrightarrow$ SVs

2-layer NN



$h(x) = \text{sign}(\quad)$

$h(x) = \text{sign}\left(\underline{w_0} + \sum_{i=1}^{m} w_i \tanh(v_i^T x)\right)$

$\underbrace{(v_i^T x + b)}$

$$K(x, x') = \tanh(c\, x^T x' + \nu)$$

$$(c + x^T x')^Q$$

$$g(x) = \operatorname{sign}\Big(\sum_{\alpha_n^* > 0} \underbrace{\alpha_n^* y_n}_{w_i} \tanh(\underbrace{c\, x_n^T}_{v_i^T} x + \underbrace{\nu}_{\nu_0}) + \underbrace{b^*}_{w_0}\Big)$$

no. of hidden nodes.

+ve def. and symmetric

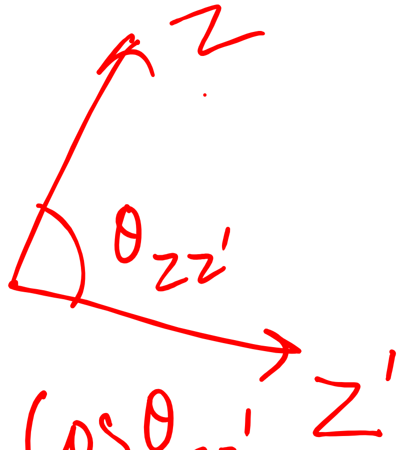
The Inner Product Measures Similarity

$$\rightarrow g(x) = \text{sign}\left(\underbrace{\sum_{\alpha_n^*>0}}\ \underbrace{\alpha_n^* y_n}_{w} K(x_n, x) + b^*\right)$$

$\cos\theta_{zz'} = 1$

$$K(\mathbf{x},\mathbf{x}') = \boxed{\mathbf{z}^{\mathrm{T}}\mathbf{z}' = \|\mathbf{z}\|\cdot\|\mathbf{z}'\|\cdot\cos(\theta_{\mathbf{z},\mathbf{z}'})}$$

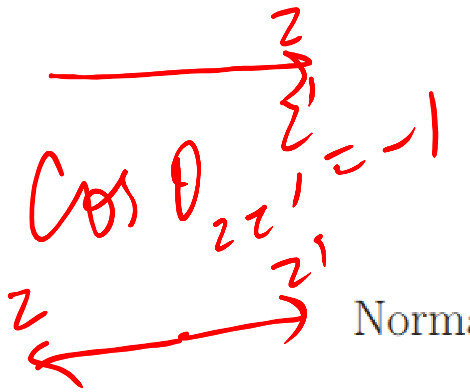
$$= \|\mathbf{z}\|\cdot\|\mathbf{z}'\|\cdot CosSim(\mathbf{z},\mathbf{z}')$$



$x \rightarrow \phi(x) = z$

$x' \rightarrow \phi(x') = z'$

$$\frac{z}{\|z\|}\cdot\frac{z'}{\|z'\|} = \cos\theta_{zz'}$$



Normalizing for size, Kernel measures similarity between input vectors

Similarity

Designing Kernels

+ve def.
symmetric

optimally!

- Construct a similarity measure for the data
- A linear model should be plausible in that transformed space

Using Kernel M/CS

1) What is a good measure of similarity?

2) Will a linear model work in feature transformed space.

Use SV machinery → solve any problem.

String Kernels

Applications: DNA sequences, Text

Business

ACGGTGTCAAACGTGTCAGTGTG

GTCGGGTCAAAACGTGAT

longest subsequence

length.

```
Dear Sir,
With reference to your letter dated 26th
March, I want to confirm the Order No.
34-09-10 placed on 3rd March, 2010.  I
would appreciate if you could send me
the account details where the payment has
to be made.  As per the invoice, we are
entitled to a cash discount of 2%.  Can
you please let us know whether it suits
you if we make a wire transfer instead of
a cheque?
```

```
Dear Jane,
I am terribly sorry to hear the news of
your hip fracture.  I can only imagine
what a terrible time you must be going
through.  I hope you and the family are
coping well.  If there is any help you
need, don't hesitate to let me know.
```

Personal

Similar?

Yes, if classifying spam versus non-spam

No, if classifying business versus personal

To design the kernel ⟶ measure similarity between strings

Bag of words (number of occurences of each atom)

Co-occurrence of substrings or subsequences

Graph Kernels

Performing classification on:

Graph structures (eg. protein networks for function prediction)

Graph nodes within a network (eg. advertise of not to Facebook users)

Similarity between **graphs**:

random walks degree sequences, connectivity properties, mixing properties.

Measuring similarity between **nodes**:

Looking at neighborhoods, $K(v, v') = \dfrac{|N(v) \cap N(v')|}{|N(v) \cup N(v')|}$.

Image similarity

-

Person 1

Person 2

Yes.

SVMs.

K

C

Linear Model

Thanks!