

Introduction to Machine Learning

Kernel Methods

Mingchen Gao

Computer Science & Engineering
State University of New York at Buffalo
Buffalo, NY, USA
mgao8@buffalo.edu
Slides adapted from Varun Chandola



University at Buffalo
Department of Computer Science
and Engineering
School of Engineering and Applied Sciences



Kernel Methods

- Extension to Non-Vector Data Examples
- Kernel Regression

Kernel Trick

- Choosing Kernel Functions
- Constructing New Kernels Using Building Blocks

Kernels

- RBF Kernel
- Probabilistic Kernel Functions
- Kernels for Other Types of Data

More About Kernels

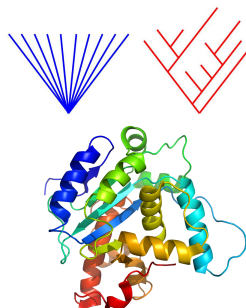
- Motivation
- Gaussian Kernel

Kernel Machines

- Generalizing RBF

Regression for Non-Vector Data Examples

- ▶ What if $\mathbf{x} \notin \mathbb{R}^D$?
- ▶ Does $\mathbf{w}^\top \mathbf{x}$ make sense?
- ▶ How to adapt?
 1. Extract features from \mathbf{x}
 2. Is not always possible

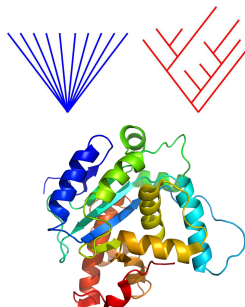


When, in the course of human events, it becomes necessary for one people to dissolve the political bands which have connected them with another, and to assume among the powers of the earth, the separate and equal station to which the laws of nature and of nature's God entitle them, a decent respect to the opinions of mankind requires that they should declare the causes which impel them to the separation.

We hold these truths to be self-evident, that all men are created equal, that they are endowed by their Creator with certain unalienable rights, that among these are life, liberty and the pursuit of happiness.

Regression for Non-Vector Data Examples

- ▶ What if $\mathbf{x} \notin \mathbb{R}^D$?
- ▶ Does $\mathbf{w}^\top \mathbf{x}$ make sense?
- ▶ How to adapt?
 1. Extract features from \mathbf{x}
 2. Is not always possible
- ▶ Sometimes it is easier/natural to compare two objects.
 - ▶ A similarity function or **kernel**
 - ▶ Nonparametric method



When, in the course of human events, it becomes necessary for one people to dissolve the political bands which have connected them with another, and to assume among the powers of the earth, the separate and equal station to which the laws of nature and of nature's God entitle them, a decent respect to the opinions of mankind requires that they should declare the causes which impel them to the separation.

We hold these truths to be self-evident, that all men are created equal, that they are endowed by their Creator with certain unalienable rights, that among these are life, liberty and the pursuit of happiness.

A Similarity Kernel

- ▶ Domain-defined measure of similarity

Example

Strings: Length of longest common subsequence, inverse of edit distance

Example

Multi-attribute Categorical Vectors: Number of matching values

Can Regression be Adapted to Use a Kernel?

- ▶ Ridge regression estimate:

$$\mathbf{w} = (\lambda I_D + \mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

- ▶ Prediction at \mathbf{x}^* :

$$y^* = \mathbf{w}^\top \mathbf{x}^* = ((\lambda I_D + \mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y})^\top \mathbf{x}^*$$

- ▶ Still needs training and test examples as D length vectors
- ▶ Rearranging above (Sherman-Morrison-Woodbury formula or *Matrix Inversion Lemma* [See Murphy p120, Matrix Cookbook])

$$y^* = \mathbf{y}^\top (\lambda I_N + \mathbf{X} \mathbf{X}^\top)^{-1} \mathbf{X} \mathbf{x}^*$$

Using the Dot Product

$$\mathbf{y}^* = \mathbf{y}^\top (\lambda \mathbf{I}_N + \mathbf{X}\mathbf{X}^\top)^{-1} \mathbf{X}\mathbf{x}^*$$

$\mathbf{X}\mathbf{X}^\top$?

$$\mathbf{X}\mathbf{X}^\top = \begin{pmatrix} \langle \mathbf{x}_1, \mathbf{x}_1 \rangle & \langle \mathbf{x}_1, \mathbf{x}_2 \rangle & \cdots & \langle \mathbf{x}_1, \mathbf{x}_N \rangle \\ \langle \mathbf{x}_2, \mathbf{x}_1 \rangle & \langle \mathbf{x}_2, \mathbf{x}_2 \rangle & \cdots & \langle \mathbf{x}_2, \mathbf{x}_N \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \mathbf{x}_N, \mathbf{x}_1 \rangle & \langle \mathbf{x}_N, \mathbf{x}_2 \rangle & \cdots & \langle \mathbf{x}_N, \mathbf{x}_N \rangle \end{pmatrix}$$

Using the Dot Product

$$\mathbf{y}^* = \mathbf{y}^\top (\lambda \mathbf{I}_N + \mathbf{X}\mathbf{X}^\top)^{-1} \mathbf{X}\mathbf{x}^*$$

$\mathbf{X}\mathbf{X}^\top$?

$$\mathbf{X}\mathbf{X}^\top = \begin{pmatrix} \langle \mathbf{x}_1, \mathbf{x}_1 \rangle & \langle \mathbf{x}_1, \mathbf{x}_2 \rangle & \cdots & \langle \mathbf{x}_1, \mathbf{x}_N \rangle \\ \langle \mathbf{x}_2, \mathbf{x}_1 \rangle & \langle \mathbf{x}_2, \mathbf{x}_2 \rangle & \cdots & \langle \mathbf{x}_2, \mathbf{x}_N \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \mathbf{x}_N, \mathbf{x}_1 \rangle & \langle \mathbf{x}_N, \mathbf{x}_2 \rangle & \cdots & \langle \mathbf{x}_N, \mathbf{x}_N \rangle \end{pmatrix}$$

$\mathbf{X}\mathbf{x}^*$?

$$\mathbf{X}\mathbf{x}^* = \begin{pmatrix} \langle \mathbf{x}_1, \mathbf{x}^* \rangle \\ \langle \mathbf{x}_2, \mathbf{x}^* \rangle \\ \vdots \\ \langle \mathbf{x}_N, \mathbf{x}^* \rangle \end{pmatrix}$$

Generalizing to Non-linear Regression

- Consider a set of P functions that can be applied on input example \mathbf{x}

$$\phi = \{\phi_1, \phi_2, \dots, \phi_P\}$$

$$\Phi = \begin{pmatrix} \phi_1(\mathbf{x}_1) & \phi_2(\mathbf{x}_1) & \cdots & \phi_P(\mathbf{x}_1) \\ \phi_1(\mathbf{x}_2) & \phi_2(\mathbf{x}_2) & \cdots & \phi_P(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(\mathbf{x}_N) & \phi_2(\mathbf{x}_N) & \cdots & \phi_P(\mathbf{x}_N) \end{pmatrix}$$

- Prediction:

$$y^* = \mathbf{y}^\top (\lambda \mathbf{I}_N + \Phi \Phi^\top)^{-1} \Phi \phi(\mathbf{x}^*)$$

- Each entry in $\Phi \Phi^\top$ is $\langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle$

The Great Kernel Trick

- ▶ Replace dot product $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$ with a function $k(\mathbf{x}_i, \mathbf{x}_j)$
- ▶ Replace $\mathbf{X}\mathbf{X}^\top$ with \mathbf{K}

$$K[i][j] = k(\mathbf{x}_i, \mathbf{x}_j)$$

- ▶ \mathbf{K} - *Gram Matrix*
- ▶ k - kernel function
 - ▶ Similarity between two data objects

Kernel Regression

$$y^* = \mathbf{y}^\top (\lambda \mathbf{I}_N + \mathbf{K})^{-1} k(\mathbf{X}, \mathbf{x}^*)$$

How to Construct a Kernel?

- ▶ Already know the simplest kernel function:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^\top \mathbf{x}_j$$

Approach 1: Start with basis functions

$$k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j)$$

Approach 2: Direct design (good for non-vector inputs)

- ▶ Measure **similarity** between \mathbf{x}_i and \mathbf{x}_j
- ▶ **Gram matrix must be positive semi-definite**
- ▶ k should be *symmetric*

Using Building Blocks

$$k(\mathbf{x}_i, \mathbf{x}_j) = ck_1(\mathbf{x}_i, \mathbf{x}_j)$$

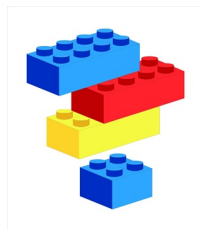
$$k(\mathbf{x}_i, \mathbf{x}_j) = f(\mathbf{x})k_1(\mathbf{x}_i, \mathbf{x}_j)f(\mathbf{x}_j)$$

$$k(\mathbf{x}_i, \mathbf{x}_j) = q(k_1(\mathbf{x}_i, \mathbf{x}_j)) \text{ } q \text{ is a polynomial}$$

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp(k_1(\mathbf{x}_i, \mathbf{x}_j))$$

$$k(\mathbf{x}_i, \mathbf{x}_j) = k_1(\mathbf{x}_i, \mathbf{x}_j) + k_2(\mathbf{x}_i, \mathbf{x}_j)$$

$$k(\mathbf{x}_i, \mathbf{x}_j) = k_1(\mathbf{x}_i, \mathbf{x}_j)k_2(\mathbf{x}_i, \mathbf{x}_j)$$



- ▶ If \mathbf{K} is positive definite - **Mercer Kernel**
- ▶ **Radial Basis Function** or **Gaussian Kernel**

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{x}_i - \mathbf{x}_j\|^2\right)$$

- ▶ **Cosine Similarity**

$$k(\mathbf{x}_i, \mathbf{x}_j) = \frac{\mathbf{x}_i^\top \mathbf{x}_j}{\|\mathbf{x}_i\| \|\mathbf{x}_j\|}$$

The RBF Kernel

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp \left(-\frac{1}{2\sigma^2} \|\mathbf{x}_i - \mathbf{x}_j\|^2 \right)$$

- Mapping inputs to an infinite dimensional space

Probabilistic Kernel Functions

- ▶ Allows using generative distributions in discriminative settings
- ▶ Uses class-independent probability distribution for input \mathbf{x}

$$k(\mathbf{x}_i, \mathbf{x}_j) = p(\mathbf{x}_i|\boldsymbol{\theta})p(\mathbf{x}_j|\boldsymbol{\theta})$$

- ▶ Two inputs are more similar if both have high probabilities

Bayesian Kernel

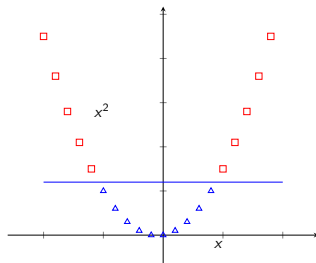
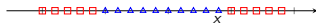
$$k(\mathbf{x}_i, \mathbf{x}_j) = \int p(\mathbf{x}_i|\boldsymbol{\theta})p(\mathbf{x}_j|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$$

Kernels for Non-vector Data

- ▶ **String Kernel**
- ▶ **Pyramid Kernels**

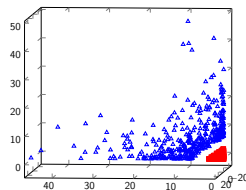
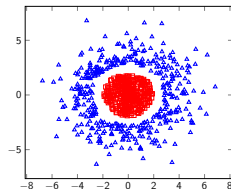
Why Use Kernels?

- ▶ $x \in \mathbb{R}$
- ▶ No linear separator
- ▶ Map $x \rightarrow \{x, x^2\}$
- ▶ Separable in 2D space



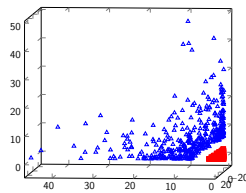
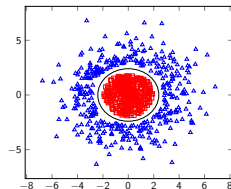
Another Example

- ▶ $\mathbf{x} \in \mathbb{R}^2$
- ▶ No linear separator
- ▶ Map $\mathbf{x} \rightarrow \{x_1^2, \sqrt{2}x_1x_2, x_2^2\}$
- ▶ A circle as the decision boundary



Another Example

- ▶ $\mathbf{x} \in \mathbb{R}^2$
- ▶ No linear separator
- ▶ Map $\mathbf{x} \rightarrow \{x_1^2, \sqrt{2}x_1x_2, x_2^2\}$
- ▶ A circle as the decision boundary



The Gaussian Kernel

- ▶ The *squared dot product* kernel ($\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^2$):

$$k(\mathbf{x}_i, \mathbf{x}_j) \triangleq \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j)$$

$$\phi(\mathbf{x}_i) = \{x_{i1}^2, \sqrt{2}x_{i1}x_{i2}, x_{i2}^2\}$$

- ▶ What about the Gaussian kernel (radial basis function)?

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{x}_i - \mathbf{x}_j\|^2\right)$$

Why is the Gaussian Kernel Mapping to Infinite Dimensions

- Assume $\sigma = 1$ and $\mathbf{x} \in \Re$ (denoted as x)

$$\begin{aligned}k(x_i, x_j) &= \exp(-x_i^2) \exp(-x_j^2) \exp(2x_i x_j) \\&= \exp(-x_i^2) \exp(-x_j^2) \sum_{k=0}^{\infty} \frac{2^k x_i^k x_j^k}{k!} \\&= \sum_{k=0}^{\infty} \left(\frac{2^{k/2}}{\sqrt{k!}} x_i^k \exp(-x_i^2) \right) \left(\frac{2^{k/2}}{\sqrt{k!}} x_j^k \exp(-x_j^2) \right)\end{aligned}$$

- Using Taylor Series Expansion

$$k(x_i, x_j) = \begin{pmatrix} 1 \\ 2^{1/2} x_i^1 \exp(-x_i^2) \\ \frac{2^{2/2}}{2} x_i^2 \exp(-x_i^2) \\ \vdots \end{pmatrix}^{\top} \times \begin{pmatrix} 1 \\ 2^{1/2} x_j^1 \exp(-x_j^2) \\ \frac{2^{2/2}}{2} x_j^2 \exp(-x_j^2) \\ \vdots \end{pmatrix}$$

- ▶ We can use kernel function to *generate* new features
- ▶ Evaluate kernel function for each input and a set of K centroids

$$\phi(\mathbf{x}) = [k(\mathbf{x}, \mu_1), k(\mathbf{x}, \mu_2), \dots, k(\mathbf{x}, \mu_K)]$$

$$y = \mathbf{w}^\top \phi(\mathbf{x}), \quad y \sim \text{Ber}(\mathbf{w}^\top \phi(\mathbf{x}))$$

- ▶ If k is a Gaussian kernel \Rightarrow Radial Basis Function Network (RBF)
- ▶ How to choose μ_i ?
 - ▶ Clustering
 - ▶ Random selection

- ▶ Another option: Use every input example as a “centroid”

$$\phi(\mathbf{x}) = [k(\mathbf{x}, \mathbf{x}_1), k(\mathbf{x}, \mathbf{x}_2), \dots, k(\mathbf{x}, \mathbf{x}_N)]$$

- ▶ Murphy book Chapter 17.1