Kernels and Kernel Regression

Lyle Ungar

Learning objectives
Kernel definition and examples
RBF algorithm (again)
Kernel regression

What is a kernel?

- k(**x**,**y**)
  - Measures the *similarity* between a pair of points ***x*** and ***y***
  - Symmetric and positive definite
- **Example: Gaussian kernel**
  - $k(\mathbf{x},\mathbf{y}) = \exp(-||\mathbf{x} - \mathbf{y}||^2/\sigma^2) = \exp(-d(\mathbf{x}, \mathbf{y})^2/\sigma^2)$
- **Uses of kernels**
  - RBF
  - Kernel regression
  - SVMs

Kernel definition

A symmetric function $k\,(\mathbf{x}_i, \mathbf{x}_j)$: $\boldsymbol{X} \times \boldsymbol{X} \rightarrow \mathbb{R}$

is a positive definite kernel on $\boldsymbol{X}$ if

$$\Sigma_{i,j}\, c_i\, c_j\, k(\mathbf{x}_i, \mathbf{x}_j) \geq 0$$

for all c_i c_j $\mathbf{x}_i$, $\mathbf{x}_j$

summed over any set of i,j pairs

We won't actually use this

What is a kernel?

- k($\mathbf{x}$,$\mathbf{y}$)
  - Measures the *similarity* between a pair of points $\boldsymbol{x}$ and $\boldsymbol{y}$
  - Symmetric and positive semi-definite (PSD)
  - Often tested using a *Kernel Matrix,*
    - a PSD matrix **K** with elements $K_{ij} = k(\mathbf{x}_i,\mathbf{x}_j)$ from all pairs of rows of a matrix $\boldsymbol{X}$
    - A *PSD matrix* has only non-negative eigenvalues

Kernel examples

- **Linear kernel**
  - $k(\mathbf{x},\mathbf{y}) = \mathbf{x}^T\mathbf{y}$
- **Gaussian kernel**
  - $k(\mathbf{x},\mathbf{y}) = \exp(-||\mathbf{x} - \mathbf{y}||^2/\sigma^2)$
- **Quadratic kernel**
  - $k(\mathbf{x},\mathbf{y}) = (\mathbf{x}^T\mathbf{y})^2$ or $(\mathbf{x}^T\mathbf{y} + 1)^2$
- **Combinations and transformations of kernels**

Radial Basis Functions (RBFs)

1) Pick k basis function centers μ_j ***using k-means clustering***

2) Let $h(\mathbf{x}) = w_1\, \phi_1(\mathbf{x}) + w_2\, \phi_2(\mathbf{x}) + \ldots\, w_k\, \phi_k(\mathbf{x})$

where

$\phi_j(\mathbf{x}) = k(\mathbf{x}, \boldsymbol{\mu}_j) = exp(-||\mathbf{x} - \mu_j||_2^2 / C)$

3) Estimate w using linear regression

RBFs can do …

- **Use k < p basis vectors**
  - Dimensionality reduction
  - Good for high dimensional feature spaces
- **Use k > p basis vectors**
  - Increases the dimensionality
  - Can make a formerly nonlinear problem linear
- **Use k=n basis vectors**
  - Switches to a *dual* representation

Kernel Regression

$$\hat{y}(\mathbf{x}) = \frac{\sum_{i=1}^{n} K(\mathbf{x},\mathbf{x}_i)y_i}{\sum_{i=1}^{n} K(\mathbf{x},\mathbf{x}_i)}$$

https://alliance.seas.upenn.edu/~cis520/wiki/index.php?n=Lectures.KernelRegression

Kernel classification

$$\hat{y}(\mathbf{x}) = sign(\sum_{i=1}^{n} K(\mathbf{x}, \mathbf{x}_i)y_i)$$

y_i = -1,1

KNN vs Kernel regression

- When is k-NN better than kernel regression?
- When is kernel regression better than k-NN

A kernel *k(x,y)*

- **Measures the *similarity* between a pair of points *x* and *y***
- **Symmetric and positive semi-definite**
- **Often tested using a *Kernel Matrix,***
  - a PSD matrix **K** with elements $K_{ij} = k(\mathbf{x}_i,\mathbf{x}_j)$ from all pairs of rows of a matrix *X* of predictors
  - A *PSD matrix* has only non-negative eigenvalues

Kernel matrix example

- Pick a matrix X

$$\begin{vmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{vmatrix}$$

- Compute $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$
- Test the eigenvalues

- What is K for X using the linear kernel?

How was my speed

- A Slow
- B Good
- C Fast