Kernels

Lyle Ungar

Penn
University of Pennsylvania

What is a kernel?

- **k(x,y)**
  - Measures the *similarity* between a pair of points ***x*** and ***y***
- **Required properties**
  - Symmetric and positive semi-definite (PSD)
  - Often tested using a *Kernel Matrix,*

    —a matrix K with elements $K_{ij} = k(x_i,x_j)$

    where the points are all pairs of rows of a matrix *X* of predictors

    A *PSD matrix* has only non-negative singular values
- **Uses**
  - Anywhere you want to replace inner products $x^T y$ with inner products of $\phi(x)^T \phi(y)$

Properties of a Kernel

Definition: *A finitely positive semi-definite function* $k : x \times y \rightarrow R$ *is a symmetric function of its arguments for which matrices formed by restriction on any finite subset of points is positive semi-definite.*

$$\alpha^T K \alpha \geq 0 \quad \forall \alpha$$

Theorem: *A function* $k : x \times y \rightarrow R$ *can be written as* $k(x, y) = <\Phi(x), \Phi(y)>$ *where* $\Phi(x)$ *is a feature map* $x \rightarrow \Phi(x) \in F$ *iff k(x,y) satisfies the semi-definiteness property.*

Relevance: We can now check if k(x,y) is a proper kernel using only properties of k(x,y) itself, i.e. without the need to know the feature map!

Credit: Max Welling

Where are kernels used?

- **Nearest neighbors**
  - Measure similarity in the kernel space
- **Linear regression**
  - Map points to new, transformed feature space
- **Logistic regression**
  - Map points to new, transformed feature space
- **SVMs and Perceptrons**
- **PCA**
  - SVD[X^TX]
- **CCA**
  - SVD[$(X^TX)^{-1/2}(X^TY)\,(Y^TY)^{-1/2}$]

What is the most common kernel method for linear regression?

What are we seeking to accomplish with kernels for classification?

What is the main benefit for PCA?

When are kernels most necessary for CCA?

Kernels form a dual representation

- Start with an ***n*p*** matrix ***X*** of predictors
- Generate an ***n*n*** *kernel matrix* ***K***
  - with elements $K_{ij} = k(x_i, x_j)$

When did we work in the dual? Why is it not bad to generate a potentially much larger feature space

The "kernel trick" avoids computing $\phi(x)$

- $k(x,y) = \phi(x)^T \phi(y)$
- So we can compute *k(x,y)* and never compute the expanded features $\phi(x)$

How are kernels selected?

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

True or false

- Kernels in effect transform observations **x** to a higher dimension space **ϕ(x)**
- Since kernels measure similarity,

  k(x,y) > k(x,x) for **x != y**
- If there exists a pair of points **x** and **y** such that **k(x,y) < 0**, then **k()** is not a kernel
- A quadratic kernel, when used in linear regression, gives results very similar to including quadratic interaction terms in the regression
- Any function **ϕ(x)** can be used to generate a kernel using **k(x,y) = ϕ(x)^Tϕ (y)**
- Any distance metric **d(x,y)** can be used to generate a kernel using **k(x,y) = exp(-d(x,y))**