Kernel Methods: High Level View

David Rosenberg

New York University

November 1, 2015

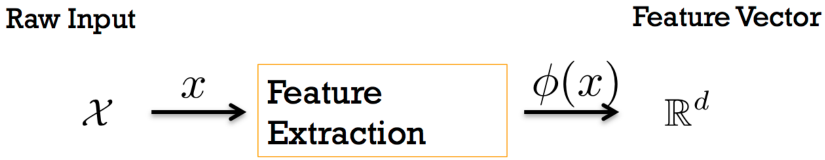
The Input Space $\mathcal{X}$

- Our general learning theory setup: no assumptions about $\mathcal{X}$
- But $\mathcal{X} = \mathbf{R}^d$ for the specific methods we've developed:
    - Ridge regression
    - Lasso regression
    - Linear SVM

Feature Extraction

Definition

Mapping an input from $\mathcal{X}$ to a vector in $\mathbf{R}^d$ is called **feature extraction** or **featurization**.



- e.g. Quadratic feature map: $\mathcal{X} = \mathbf{R}^d$

$$\phi(x) = \left(x_1, \ldots, x_d, x_1^2, \ldots, x_d^2, \sqrt{2}x_1x_2, \ldots, \sqrt{2}x_ix_j, \ldots \sqrt{2}x_{d-1}x_d\right)^T.$$

High-Dimensional Features Good but Expensive

- To get **expressive** hypothesis spaces using linear models,
  - need high-dimensional feature spaces
- But more costly in terms of computation and memory.

Some Methods Can Be "Kernelized"

Definition

A method is **kernelized** if inputs only appear inside inner products: $\langle \phi(x), \phi(y) \rangle$ for $x, y \in \mathcal{X}$.

- The function

  $$k(x, y) = \langle \phi(x), \phi(y) \rangle$$

  is called the **kernel** function.

Kernel Evaluation Can Be Fast

Example

Quadratic feature map

$$\phi(x) = (x_1, \ldots, x_d, x_1^2, \ldots, x_d^2, \sqrt{2}x_1x_2, \ldots, \sqrt{2}x_ix_j, \ldots \sqrt{2}x_{d-1}x_d)^T$$

has dimension $O(d^2)$, but

$$k(w,x) = \langle \phi(w), \phi(x) \rangle = \langle w, x \rangle + \langle w, x \rangle^2$$

- Naively explicit computation of $k(w,x)$: $O(d^2)$
- Implicit computation of $k(w,x)$: $O(d)$

Recap

1. Given a kernelized ML algorithm.
2. Can swap out the inner product for a new kernel function.
3. New kernel may correspond to a high dimensional feature space.
4. Once kernel matrix is computed, computational cost depends on number of data points, rather than the dimension of feature space.