CMSC35000-1 Introduction to Artificial Intelligence **Winter 2007**

Lecture 6: 1/30/2007

Lecturer: Partha Niyogi *Scribe: Sonjia Waxmonsky*

6.1 Supervised learning

- Neural Networks / Perceptrons
- Decision trees
- Nearest Neighbor
- Math Programming

6.1.1 Kernel-based methods

Definition 6.1 *A Kernel K is defined as* $K : (\mathbb{R}^k \times \mathbb{R}^k) \rightarrow \mathbb{R}$.

K is (a) **symmetric**: $\forall_{x,y} K(x,y) = K(y,x)$ and (b) **positive definite**: $\forall_{z_1,\ldots,z_n \in X} K_{i,j} = K(z_i, z_j) > 0$.

A **positive definite matrix** is a Hermitian matrix all of whose eigenvalues are positive.

$\forall x$, define $K(x, \bullet)$ as $K_x : \mathbb{R}^k \rightarrow \mathbb{R}$

The linear combinations of these functions $\sum \alpha_i K\mathbf{x}_i$ gives a linear space of functions. We define an inner product for this space:

Given $g = K_x, h = K_z$,

$$< g, h >=< K_x, K_z >= K(x,z) \tag{6.1}$$

$$< \sum \alpha_x K_x, \sum \beta_z K_z >= \sum \alpha_x \beta_z K(x,z) \tag{6.2}$$

The set of linear functions H_k is the Reproducing Kernel Hilbert Space associated with kernel K:

Examples of Kernels

- Linear kernel: $K(x,y) = x^T y$, $K_x(y) = x^T y$
  $K_x(x) + K_z(y) = x^T y + x^T z = (x+z)^T y$,
  So $||K_w|| =< K_w, K_w >= w^T w = ||w||^2$
- Polynomial kernel: $K(x,z) = (x^T z)^d$
  $H_k = \{\text{polynomials of degree } d\}$
- Gaussian kernel: $K(x,y) = e^{-\frac{\|x-y\|}{\sigma^2}}$

6.2 Unsupervised learning

Given unlabled data $\mathbf{x}_1, \ldots, \mathbf{x}_n$, we intend to discover some natural structure of this data.

Examples:

- Density Estimation (Statistics)
- Clustering / Categorization
  - K-means
  - Hierarchical
  - Spectral

6.2.1 K-means clustering

Algorithm to partition data into k clusters, where k is given in advance.

Pseudocode for k-means:
Input: $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathbb{R}^N$

1. Pick the initial centroid of group: $\mathbf{u}_1, \ldots, \mathbf{u}_k \in \mathbb{R}^N$ (Each $\mathbf{u}_i$ can be selected at random.)
2. Assign each point to the cluster of closest centroid: $\mathbf{x}_i \rightarrow \arg\min_l \parallel \mathbf{x}_i - \mathbf{u}_l \parallel$
3. Recompute k centroids as means of each partitions X_i: $\mathbf{u}_i := \frac{1}{|\mathbf{X}_i|}\sum_{\mathbf{x}\in\mathbf{X}_i}\mathbf{x}$
4. Repeat Steps 2 and 3 until the centroids remain unchanged, or change is minimal.

We want to minimize the following measure of tighness of the partitions, where a small value indicates that the data is tightly clustered:

$$\min_{\substack{\{\mathbf{u}_1,\ldots,\mathbf{u}_k\}\\ \{\mathbf{X}_1,\ldots,\mathbf{X}_k\}}} \sum_{i=1}^{k}\sum_{\mathbf{x}\in\mathbf{X}_i} ||\mathbf{x}-\mathbf{u}_i||^2$$

For each partition $\mathbf{X}_i$ we want to find:

$$\min_{\{\mathbf{u}\}} \sum_{\mathbf{x}\in\mathbf{X}_i} ||\mathbf{x}-\mathbf{u}||^2$$

Note that:

$$\sum_{\mathbf{x}\in\mathbf{X}_i} ||\mathbf{x}-\mathbf{u}||^2 = \sum_{\mathbf{x}\in\mathbf{X}_i} (\mathbf{x}-\mathbf{u})^T(\mathbf{x}-\mathbf{u}) = \sum_{\mathbf{x}\in\mathbf{X}_i} \mathbf{x}^T\mathbf{x} - 2\mathbf{u}^T\mathbf{x} + \mathbf{u}^T\mathbf{u}$$

We take the partial derivative of the above sum:

$$\frac{\delta}{\delta\mathbf{u}} \sum_{\mathbf{x}\in\mathbf{X}_i} \mathbf{x}^T\mathbf{x} - 2\mathbf{u}^T\mathbf{x} + \mathbf{u}^T\mathbf{u} = 0$$

$$(-2\sum_{\mathbf{x}\in\mathbf{X}_i}\mathbf{x}) + (2\mathbf{u}\sum_{\mathbf{x}\in\mathbf{X}_i} 1) = 0$$

$$\mathbf{u} = \sum_{\mathbf{x}\in\mathbf{X}_i} \frac{\mathbf{x}}{|\mathbf{X}|}$$