CMSC35000-1 Introduction to Artificial Intelligence **Winter 2005**

Lecture 7: 25/1/05

Lecturer: Partha Niyogi *Scribe: Tien Nguyen*

7.1 We have discussed so far

- Perceptrons
- MLP
- Kernel base method
- Decision tree
- Generalization Error (Model selection)
- Genetic algorithm
- Bayes net

In the 2 class problem

$$f : X \to Y, Y = \{-1, 1\} \text{ or } \{0, 1\}$$

$$f : X \to R \text{ e.g. } [0, 1]$$

the goal is to find a function to

$$\min\ (y - f(x))^2$$

7.2 Multiclass supervised learning

7.2.1 Perceptron and MLP

For Perceptrons and MLP, we can do as follows. If we have a problem with k possible class labels, the label is a vector $Y \in R^k$. Y_i are identified with the corners of a simplex Δ^{k-1}.

$$Y_i = (0, 0, ..., 1, ..., 0)$$

Y_i are binary vectors which have the value of 1 at the index corresponding to the class of the data point x.

The classification function has the form:

$$f : X \to \Delta^{k-1} \in R^k$$

Thus, we can work with this:

$$\min\ ||y - f(x)||^2$$

To classify a data point, use

$$f(x) = \text{argmax}_j f_j(x)$$

7.2.2 Kernel-based methods

For Kernel-based methods, we need k functions

$$F = \{f_i(x_i) : f_i(x) = \sum_{j=1}^{n} \alpha_j * K(x_j, x_i)\},$$

so one will have to estimate $n \times k$ parameters.

We define a function

$$g : X \to \ \{1, 2, ..., k\}$$

and

$$y \in \{1, 2, ..., k\}$$

Then, we define an error function

$$e(y, g(x)) = 1,\ \text{if } y \neq g(x)$$
$$e(y, g(x)) = 0,\ \text{if } y = g(x)$$

We have to compute $E(e(y, g(x)))$ and $\hat{E}(e(y, g(x)))$ which can be combinatorially difficult.

Kernel methods and MLP use the relaxation to

$$f : X \to R$$

and classify a data point by thresholding the output value of the function f:

$$g(x) = 2,\ \text{if } f(x) > 0$$
$$g(x) = 1,\ \text{if } f(x) < 0$$

7.2.3 Decision Trees

For decision tree, we defined the impurity function as follows:

$$g(D, q) = \frac{n_1 I(D_1)}{(n_1 + n_2)} + \frac{n_2 I(D_2)}{(n_1 + n_2)},$$

where

$$I(D) = p\log(1/p) + (1-p)\log(1/(1-p))$$

To generalize to k classes, we modify the formula for $I(D)$ and use the Shannon's entropy formula:

$$I(D) = \sum_{1}^{k} p_i \log(1/p_i)$$

$I(D)$ is large when p is uniform, and small if probability mass is concentrated on one of the classes.

7.3 Unsupervised learning

Dataset: $x_1, x_2, .., x_n$ and no label set

Two things we could do in unsupervised learning are:

- Density estimation P(x)
- Clustering: partition data into classes $C_1, C_2, .., C_k$ such that
  $\bigcup C_i = \{x_1, x_2, ..., x_n\}$
  $C_i \bigcap C_j = \emptyset$

7.3.1 K-means clustering

Pseudo-code

$x_1, x_2, ..., x_n \in R^N$

1. Pick K points in the space, representing initial group centroids (for example, at random. We will discuss the issue of this initialization later)
2. Assign each point to the closest centroid.
3. Recompute K centroids as means of the partitions.
4. Repeat Steps 2 and 3 until the centroids remain unchanged

7.3.2 Questions

Can you find a configuration on which k-means does not converge?

Consider the following cases, in which k-means

1. always converges but the final configuration is not unique
2. cycles on some configurations
3. depends on the initialization

4. what happens for co-centric circles

For k-means, we want to minimize over $\{u_1, u_2, ..., u_k\}$

$$\min \sum_{j=1}^{k} e(x_j; \{u_1, u_2, ..., u_k\})$$

where

$$e(x_i; \{u_1, u_2, ..., u_k\}) = \min_j ||u_j - x_i||^2$$

This function is uniquely defined, will k-means find the optimal configuration? Given a partitioning X_1, X_2,...,X_k, and centroids $\hat{u}_1$, $\hat{u}_2$,...,$\hat{u}_k$ after the *ith* iteration, and consider for this iteration

$$\min \sum_{j=1}^{k} e^i(x_j; \{u_1, u_2, ..., u_k\})$$

is it always the case that

$$\min \sum_{j=1}^{k} e^{i+1}(x_j; \{u_1, u_2, ..., u_k\}) <$$

$$< \min \sum_{j=1}^{k} e^i(x_j; \{u_1, u_2, ..., u_k\})$$

One can replace e with a differentiable function and do gradient descent. Can you formulate the k-means algorithm for this case?