CMSC 35000 Introduction to Artificial Intelligence **Winter 2005**

Lecture 2: January 11, 2007

Lecturer: Partha Niyogi *Scribe: Katherine Riley*

1 Additional Methods for Finding $w_\star$

In Lecture 1, we discussed the Perceptron Learning Algorithm (PLA) as a method to find $w_\star$ and proved it's convergence as a limit with the number of mistakes. In Lecture 2 we briefly discussed two other methods of finding $w_\star$: the least squares methods and linear programming. While these show alternative methods of finding $w_\star$, we still do not have a justification for merits of $w_\star$.

1.1 Least Squares

As before, we have $i = 1 \ldots n$ (x_i, y_i) pairs where $y \in \{-1, 1\}$ and $x \in \mathbb{R}^D$. The PLA will not stop if there is no separating hyperplane. However, if we find $\min_w \sum (y_i - w \cdot x_i)^2$. We can solve this linear system of equations for to find w. We are not assured of finding a separating hyperplane, but we are assured the algorithm will stop.

1.2 Math Programming/Linear Programming

We want w such that $y_i(w \cdot x_i) \geq 1$. For each data point, we have n linear constraints and we want to minimize $w \cdot w$ such that $y_i(w \cdot x_i) \geq 1$. It follows directly that the distance to the i_{th} points is $\frac{y_i(w \cdot x_i)}{||w||}$

We can always find such an w if we are strictly greater than 1. Consider w such that $y_i(w \cdot x_i) > 1, \forall i$. If $w' = \alpha w$ then we can choose an α such that $\alpha(y_i(w \cdot x_i)) \geq 1$.

We want to find a $w_\star$ such that $\exists i$ such that $y_i(w_\star \cdot x_i) = 1$, then

$$\frac{1}{||w_\star||} = \min_i \frac{y_i(w_\star \cdot x_i)}{||w_\star||} = \text{dist. to nearest point, the margin}$$

2 Formal Definition of Classifier

In pattern recognition problems, we are given a set of objects chosen at random from a pattern space X, and we have to classify each object into a class from a set of classes Y (we will consider $Y = \{1, -1\}$ for now).

Example 2.1 *We can take the pattern space to be the set of all pictures, and the set we want to recognize the set of all pictures which correspond to my grandmother. If we say X is the set of all $n \times n$ pixel pictures, we can represent X by assigning a real number value to each pixel. In this case $X = \mathbb{R}^{n^2}$.*

We can also do some pre-processing on the input, called **feature extraction**, in which we pick out from the raw data the relevant features

Example 2.2 *We can look at speech recognition. Here the raw data is a function $x(t)$ (the wave function). As our pattern space X we can choose the set of all such wave functions, but this might be hard to work with. Instead we can look at some features such as power$= \frac{1}{T}\sum_{t=1}^{T} x(t)^2$, periodicity (we assign a real number, close to 1 if the wave is periodic, and closer to 0 if the wave is more noisy), and possibly other features. Then we can say that our pattern space X is the space of all these features.*

We also assume that there exists a probability distribution P over X, according to which elements of the pattern space are drawn:

$$P(x|y = 1)$$

is the probability that a specific pattern $x \in X$ (a picture in our example) is the next pattern we perceive, knowing that it will be one of the elements we're trying to identify (a picture of my grandmother in our example). Similarly, we define

$$P(x|y = -1)$$

to be the probability distribution over patterns that are not what we're trying to identify. We also must assume that our probability distribution also specifies $P(y = 1)$, and $P(y = -1)$.

Let $Z = X \times Y$, we can define P on an element of Z by Bayes's formula:

$$P(x, y) = P(y)P(x|y).$$

Definition 2.3 *A **classifier** is a function*

$$f : X \longrightarrow Y.$$

3 Error rate of a classifier and optimal classifiers

Now that we laid down the formal definitions, we can try to answer the question *What is an optimal classifier?*. In order to answer this question we need to have a way of comparing classifiers. For this purpose we define the error rate. To do this we first define a couple useful sets.

Definition 2.4 *If f is a classifier, let*

$$X_1^{(f)} = f^{-1}(1) = \{x | f(x) = 1\}$$

$$X_{-1}^{(f)} = f^{-1}(-1) = \{x | f(x) = -1\}$$

Definition 2.5 *We can define the error rate in one of several ways. First of all, we can define the error rate to be*

$$Err(f) = \int_{X_1^{(f)}} P(x)P(y = -1|x)dx + \int_{X_{-1}^{(f)}} P(x)P(y = 1|x)dx\,.$$

An alternative definition for error rate requires us to first define

$$e(f, z) = \begin{cases} 1 \text{ if } y \neq f(x) \\ 0 \text{ if } y = f(x) \end{cases} \tag{1}$$

Using this, we can define the error rate as

$$Err(f) = E[e(f, z)] = \sum_{y=\pm 1} \int_X P(x, y)e(f, z)dx = \int_X P(x)e(f, z)dz = P(f \text{ makes a mistake})\,.$$

The optimal classifier will be $\arg\min_f Err(f)$. We can now define the question of finding the best classifier within a specific class of functions $\mathcal{H}$ by $\arg\min_{f\in\mathcal{H}} Err(f)$

Definition 2.6

$$\text{We define } f_* = \begin{cases} 1 \text{ if } P(y = 1|x) > 1/2 - 1 \text{ if } P(y = -1|x) > 1/2 \end{cases} \tag{2}$$

If $P(y = 1|x) = 1/2$ it doesn't matter what we do. We call f_* the Bayes classifier, or discriminant function.

Proposition 2.7 *It seems intuitive that the classifier f_* we just defined is the optimal classifier. Formally*

$$Err(f) \geq Err(f_*)\,\forall f$$

Proof

$$\text{Let } X_1 = X_1^{(f_*)}, \text{ and } X_{-1} = X_{-1}^{(f_*)}\,.$$

Now we can write the error rate of f as

$$Err(f) = \int_{X_1^{(f)}\bigcap X_1} P(x)P(y=-1|x)dx \,+ \int_{X_1^{(f)}\bigcap X_{-1}} P(x)P(y=-1|x)dx$$

$$+ \int_{X_{-1}^{(f)}\bigcap X_1} P(x)P(y=1|x)dx \,+ \int_{X_{-1}^{(f)}\bigcap X_{-1}} P(x)P(y=1|x)dx$$

Noticing that

$$P(y=1|x) = 1 - P(y=-1|x) = 1 - 2P(y=-1|x) + P(y=-1|x)\,,$$

we can rewrite the sum of the first and the third term as

$$S_1 = \int_{X_1^{(f)}\bigcap X_1} P(x)P(y=-1|x)dx \,+ \int_{X_{-1}^{(f)}\bigcap X_1} P(x)P(y=-1|x)dx$$

$$+ \int_{X_{-1}^{(f)}\bigcap X_1} P(x)(1-2P(y=-1|x))dx\,.$$

But, by definition of f_*, $P(y=-1|x) \leq 1/2$ on X_1, therefore

$$+ \int_{X_{-1}^{(f)}\bigcap X_1} P(x)(1-2P(y=-1|x))dx \,\geq 0\,.$$

Moreover,

$$\int_{X_1^{(f)}\bigcap X_1} P(x)P(y=-1|x)dx \,+ \int_{X_{-1}^{(f)}\bigcap X_1} P(x)P(y=-1|x)dx$$

$$= \int_{X_1} P(x)P(y=-1|x)dx\,.$$

So we have shown that

$$S_1 \geq \int_{X_1} P(x)P(y=-1|x)dx\,.$$

Similarly, we can show that the sum of the second and the fourth terms gives us

$$S_2 = \int_{X_1^{(f)} \bigcap X_{-1}} P(x)P(y=1|x)dx + \int_{X_{-1}^{(f)} \bigcap X_{-1}} P(x)P(y=1|x)dx$$

$$+ \int_{X_{-1}^{(f)} \bigcap X_{-1}} P(x)(1-2P(y=1|x))dx \geq \int_{X_{-1}} P(x)P(y=1|x)dx\,.$$

So,

$$Err(f) \geq Err(f_x*).$$

□

Let us note that since f_* depends only on the probability distribution, then so does $Err(f_*)$. Since $Err(f_*) = \min_f Err(f)$, $Err(f_*)$ is unique. Moreover, if we assume that $P(y=1|x) \neq 1/2\,\forall x$, then f_* is unique.

4 Inductive inference

In general, we assume that there exists a probability distribution P on $Z = X \times Y$, but we don't know what this probability distribution is. Instead, we draw a set $S = \{z_1, \ldots, z_n\}$ according to P, then we try to come up with a function based on these examples.

Definition 2.8 *We define the training error as*

$$\widehat{Err}(f) = 1/n \sum_{i=1}^{n} e(f, z_i)\,.$$

Definition 2.9 *We also define*

$$\widehat{f}_n = \arg\min_{f \in \mathcal{H}} \widehat{Err}(f)\,,$$

where $\mathcal{H}$ is a class of functions chosen a priori.

Definition 2.10 *We define the best in class classification as*

$$f_{\mathcal{H}} = \arg\min_{f \in \mathcal{H}} Err(f)\,.$$

Our hope is that the best classifier based on the training sets are close to the best in class classifier, as our training set gets large. Formally, we hope that

$$\widehat{f_n} \longrightarrow f_{\mathcal{H}} \text{ as } n \longrightarrow \infty\,.$$

The gap between $Err(\widehat{f_n})$ and $Err(f_(H))$ is called *estimation error*, or *sampling error*.

Even if we minimize the estimation error, there can still be an insurmountable gap due to our choice of $\mathcal{H}$. In fact there will always be a difference between $Err(f_{\mathcal{H}})$ and $Err(f_*)$. However, we have a tension between these two gaps: to reduce this gap, we would want to make $\mathcal{H}$ as large as possible, to reduce the estimation error, we would want to make $\mathcal{H}$ as large as possible. So it is often a matter of compromise to choose the best possible $\mathcal{H}$.