EECS225d Audio Signal Processing in Humans and Machines **Spring 1995**

Lecture 31: April 7, 1995

Lecturer: Prof. Nelson Morgan *Scribe: Brett Walker*

Pattern Recognition continued....

31.1 Statistical Discriminant Functions

Today we look at pattern recognition using statistical discriminant functions such as the Bayes Classifier. Since we really don't know or understand all the details of speech we can model it as a random stochastic process and then use statistical tools to analyse it. Instead of assuming the incoming speech is a deterministic quantity we will model it as a random variable with a probability density function. This p.d.f. is normalised such that

$$\int_{-\infty}^{\infty} \rho(x)dx = 1$$

where $\rho(x)$ is the probability density function.

31.2 Bayes Rule

A very important rule which relates post and prior probabilities is Bayes Rule.

$$\rho(a|b) = \frac{\rho(a,b)}{\rho(b)} = \frac{\rho(b|a)\rho(a)}{\rho(b)}$$

If we let W_k be the k^{th} class and let $\vec{x}$ be the current feature vector then;

- $\rho(W_k|\vec{x})$ is the posterior probability. That is, the probability that the correct class is W_k given the current feature vector.
- $\rho(W_k)$ is the prior probability, the probability of class W_k
- $\rho(\vec{x}|W_k)$ is the likelihood of seeing the current input vector if the class is W_k

Using Bayes rule we can express the posterior in terms of the prior and the likelihood as follows

$$\rho(W_k|\vec{x}) = \frac{\rho(\vec{x}|W_k)\rho(W_k)}{\rho(\vec{x})}$$

31.3 MAP Classification

We can now classify the feature vector into a class using the Maximum a Posteriori Classifier. We simply find the class that maximises the posterior probability. Since $\rho(\vec{x})$ is a constant for all classes finding the maximum posterior boils down to

$$\operatorname*{argmax}_k \rho(W_k|\vec{x}) = \rho(\vec{x}|W_k)\rho(W_k)$$

In the two class case the probably of error is

$$\rho(\text{error}|x) = \begin{cases} \rho(W_1|x) & \text{if } x \in W_2 \\ \rho(W_2|x) & \text{if } x \in W_1 \end{cases}$$

Which extends naturally in the case of more than two classes.

The MAP Classifier could also include costs.[DH73]

31.4 Likelihood-based MAP Classification

Of course in the real world we don't know the probability densities exactly and have to estimate them. We need very good estimates near the boundary between two classes but far from the boundary the accuracy of the estimate is not as crucial. From our estimated densities we can compute the likelihood ratio.

$$\frac{\rho(W_i|\vec{x})}{\rho(W_j|\vec{x})} = \frac{\rho(\vec{x}|W_i)\rho(W_i)}{\rho(\vec{x}|W_j)\rho(W_j)}$$

If this ratio is > 1 we pick W_i over W_j and so on.

This becomes simpler if we take the log of the likelihood ratio. Now we pick W_i over W_j if

$$\log \rho(\vec{x}|W_i) + \log \rho(W_i) > \log \rho(\vec{x}|W_j) + \log \rho(W_j)$$

MAP classification based on Likelihoods can then be done using the statistical discriminant function

$$\operatorname*{argmax}_k \log \rho(\vec{x}|W_k) + \log \rho(W_k)$$

In practice the training of posterior estimators and the training of likelihood estimators may lead to different results

31.5 Bayes Classifier

So how do we find the Bayes Classifier? We need to estimate the probability density functions then use the above classification techniques. To do this we

1. Assume a form for the density. (e.g. Gaussian)
2. Quantize features (e.g. Vector Quantization) and count instances.

3. Use automatic interpolation/learning to estimate posteriors.

Assuming a Gaussian form for the density requires estimating the second order statistics and fitting them to the Gaussian form

$$\rho(x|W_i) = \frac{1}{\sqrt{2\pi}\,\sigma_i} \exp[-\frac{1}{2}(\frac{x-\mu_i}{\sigma_i})^2]$$

or in many dimensions

$$\rho(\vec{x}|W_i) = \frac{1}{(\sqrt{2\pi})^d |\Sigma_i|^{\frac{1}{2}}} \exp[-\frac{1}{2}(\vec{x}-\vec{\mu_i})^T \Sigma_i^{-1} (\vec{x}-\vec{\mu_i})]$$

where Σ_i is the covariance matrix of the i^{th} class and d is the dimension of $\vec{x}$.

The product $(\vec{x}-\vec{\mu_i})^T \Sigma_i^{-1}(\vec{x}-\vec{\mu_i})$ is called the Mahalanobis distance.

If we let $g_i(\vec{x}) = \log \rho(\vec{x}|\rho(W_i) + \log \rho(W_i)$ then for the Gaussian form above

$$g_i(\vec{x}) = -\frac{1}{2}(\vec{x}-\vec{\mu_i})^T \Sigma_i^{-1}(\vec{x}-\vec{\mu_i}) - \frac{d}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_i| + \log \rho(W_i)$$

so maximising $g_i(\vec{x})$ over i is like minimising the Mahalanobis distance.

Maximising $g_i(\vec{x})$ is simplified in two cases

1. There is no correlation between the classes. i.e. $\Sigma_i = \sigma^2 \vec{I}$

$$\Rightarrow |\Sigma_i| = \sigma^{2d}$$

$$\Rightarrow g_i(\vec{x}) = -\frac{||\vec{x}-\vec{\mu_i}||^2}{2\sigma^2} + \log \rho(W_i)$$

2. Correlation matrix same for all classes. i.e. $\Sigma_i = \Sigma$

$$\Rightarrow \text{linear discriminant function}$$

The Gaussian approximation does not work in all cases since many p.d.f.'s can have the same second order statistics yet be totally different. A better solution is to use a mix of Gaussians.

References

[DH73] Duda and Hart, *Pattern Classification and Scene Analysis*, Wiley-Interscience, 1973.