CSCE 970: Pattern Recognition	Spring Semester, 1999

Topic 5: Feature Selection

Lecturer: Stephen Scott *Scribe: Jamil Saquer*

1 Introduction

In previous lectures we talked about classifiers design where we assumed that the feature vectors were available and we didn't worry about how these features were generated or selected. In this lecture we talk about feature selection which is one of the important stages involved in the design of a classification system. From Figure 1 we notice that feature selection is an intermediate stage between feature generation and classifier design. The goal of feature selection is to choose the best features from all the features that we get as a result of the feature generation stage, which could be in the hundreds, to use in classifier design.

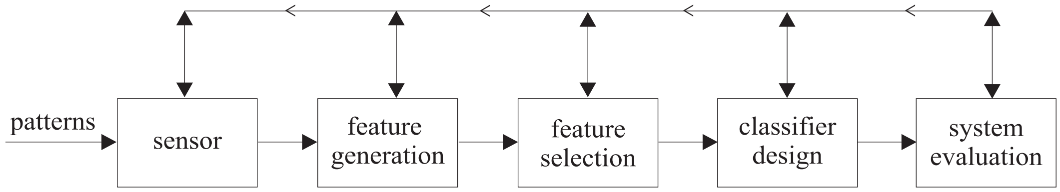

Figure 1: The basic stages involved in the design of a classification system.

Feature selections is important for many reasons. First, the data we get from the feature generation stage is defined by sensor data which may be defined for convenience rather than classification discriminating power [1]. Second, to reduce computational complexity in training and classification. Third, classifiers that use less features tend to generalize better. This is because the higher the ratio of the number of training patterns N to the number of classifier parameters l (which are affected by the number of features), the better the generalization properties of the classifier. Fourth, classifier performance evaluation can be assessed more accurately when less features are involved.

In the feature selection process it is important that the selected features take distant values in the different classes and close values in the same class. This means that we should get features that provide large between-class distance and small withen-class distance. Furthermore, we should drop features that do not help in the classification process.

There are three approaches to estimate the discriminating capabilities of the features. The first is to examine the features individually and discard those with little discriminatory capabilities. The second is to examine features in combinations. The third approach is to apply linear or nonlinear transformations to a feature vector which may lead to new feature vectors with better discriminatory properties.

2 Preprocessing

Before selecting the features, we need to preprocess the data and have it ready for feature selection. This includes removing outliers, normalizing the data, and handling missing data.

2.1 Outlier Removal

Outliers are points that differ greatly from the mean of a random variable. They can be the result of noisy measurements and may produce large errors during training. A threshold value, usually a number of times of the standard deviation, can be used to discard outliers if their number is small. However, if number of outliers is not small and they are the result of a distribution with long tails, then the designer may have to use cost functions that are not very sensitive to outliers [2].

2.2 Data Normalization

Features with values in different ranges can have different influence on the cost function during training. For example, a feature with range [0,1000] can affect training a neural network more than a feature with range [0,10] even though both are equally significant. This problem is overcome by normalizing the features so that their values lie withen similar ranges. There are three approaches for normalization.

• *Approach 1:* Linearly scale each feature x_{ik} $(k = 1, \ldots, l)$ to an interval such as [0,1] by dividing through by $\max\{x_{ik}\}$, where $1 \leq i \leq N$. A problem will result when using this approach if the value we divide by is not really the maximum. This might happen because training data is only a subset of the whole date.

• *Approach 2:* Transform each feature x_{ik} so that it has mean 0 and variance 1. This is done by using the following linear transformation:

$$\overline{x}_k = \tfrac{1}{N}\textstyle\sum_{i=1}^{N} x_{ik}, \quad k = 1, 2, \ldots l$$

$$\sigma^2 = \tfrac{1}{N-1}\textstyle\sum_{i=1}^{N}(x_{ik} - \overline{x}_k)^2$$

$$\hat{x}_{ik} = \tfrac{x_{ik} - \overline{x}_k}{\sigma_k}$$

• *Approach 3:* Use softmax scaling to map data to the interval [0,1]. Here each value x_{ik} will be transformed to a value y_{ik} which in turn is transformed to $\hat{x}_{ik} \in [0, 1]$ using the following nonlinear transformations:

$$y_{ik} = \frac{x_{ik} - \overline{x}_k}{r\sigma_k}, \quad \hat{x}_{ik} = \frac{1}{1 + \exp(-y)}$$

This is a squashing transformation that maps data values near the mean approximately linearly to the range [0,1] and squashes values far from the mean to 0 or 1.

2.3 Missing Data

In practice, it may happen that not all values of the training features are available. For example, some portions of a patient's medical history may be lost. In this case we may discard some of these features especially if number of training data is high enough. If number of training data is small, we may not afford to drop these features. However, we can replace the missing values with the mean or more sophisticated statistical information. Another approach that is used in data mining is to replace the missing data by all possible values but this can be done only if the range of the missing values is discrete and small.

3 Individual Feature Selection Based on Statistical Hypothesis Testing

A first step in feature selection is to look at features individually and test their discrimination power for the different classes. This helps eliminate obviously poor features before proceeding to more sophisticated feature selection methods. The main idea is that if a given feature x_k takes close values on the different classes, there is a strong probability that x_k has little ability to discriminate between the different classes. We will consider standard statistical hypotheses testing techniques to determine with high confidence if values of x_k different significantly for the different classes. For a review of statistical hypothesis testing, the reader may refer to [3].

3.1 Background

For a given feature x and class w_i, where $i = 1, 2$, let

$$
\begin{array}{ll}
\mu_i = \text{true mean} & \overline{x}_i = \text{empirical mean} \\
\sigma_i^2 = \text{true variance} & s_i^2 = \text{empirical variance} \\
n_i = \text{number of features in } w_i & N = n_1 + n_2
\end{array}
$$

Let the *null hypothesis* H_0 and the *alternative hypotheses* H_1 be as follows:

$$H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 \neq \mu_2$$

If the empirical data fail to reject the null hypothesis with high confidence (say 95%), then we conclude that the means are statistically similar. This means that there is a high probability that the feature under consideration does not distinguish very well between w_1 and w_2. Therefore, we can discard that feature provided that the variances σ_1 and σ_2 are similar. On the other hand, if the empirical data rejects the null hypothesis, there is a good possibility that that feature is good to keep.

3.2 Differences of Means

If we think of $\overline{x}_1$ and $\overline{x}_2$ above as two random samples with sizes n_1 and n_2 drawn from w_1 and w_2, then they can be considered as random variables. Therefore, their difference $(\overline{x}_1 - \overline{x}_2)$ is also a random variable. It is easy to show that $(\overline{x}_1 - \overline{x}_2)$ is an unbiased estimator of $(\mu_1 - \mu_2)$, that is $\mathrm{E}[\overline{x}_1 - \overline{x}_2] = \mu_1 - \mu_2$.

By the central limit theorem, if n_i is sufficiently large, then $\overline{x}_i$ is approximately a normal distribution $\mathcal{N}(\mu_i, \sigma_i^2/n_i)$ regardless of the distribution of the original data. We notice that as the sample size n_i increases, the ratio of σ_i^2/n_i decreases. Since both $\overline{x}_1$ and $\overline{x}_2$ are approximately normally distributed, the random variable $(\overline{x}_1 - \overline{x}_2)$ is also approximately normally distributed with mean $(\mu_1 - \mu_2)$ and variance

$$\sigma^2_{(\overline{x}_1 - \overline{x}_2)} = \sigma^2_{\overline{x}_1} + \sigma^2_{\overline{x}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

3.3 The z Statistic

If both μ_i and σ_i are known for $i = 1, 2$, we can compute their z statistic which is given by

$$z = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

The z statistic is a measure of the distance from $(\overline{x}_1 - \overline{x}_2)$ to its mean given in number of standard deviation units. If we assume the null hypotheses H_0 is true, that is $\mu_1 = \mu_2$, then

$$z = \frac{(\overline{x}_1 - \overline{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

In this case z follows the normal distribution $\mathcal{N}(0, 1)$.

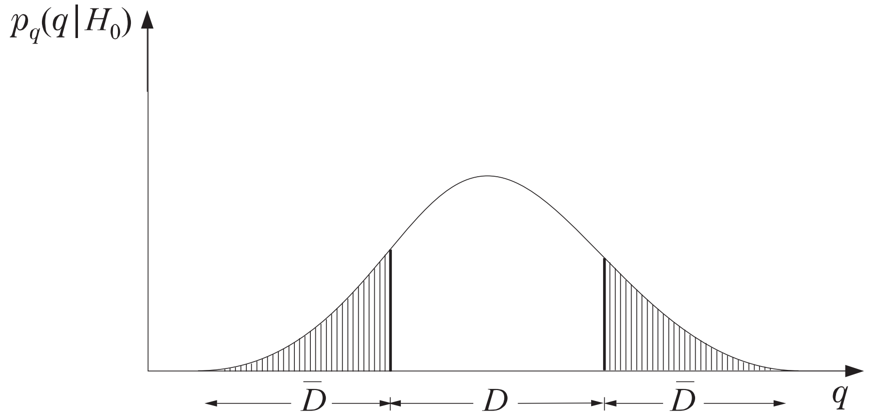

Figure 2: The basic stages involved in the design of a classification system.

In Figure 2, D represents the interval in which there is high probability that z lies under the null hypotheses H_0 and $\overline{D}$ represents its complement, that is, the interval of low probability under H_0. If the value of z that results from the samples lies in D, we accept H_0; otherwise, we reject H_0. D is called the *acceptance interval* and $\overline{D}$ is called the *critical interval.* If H_0 is true, then the probability of error in making a decision is equal to $P(z \in \overline{D} \mid H_0)$ and is denoted by ρ. Clearly ρ is equal to the area of the shaded region in Figure 2.

The value of ρ is known as the *significance level.* In practice a significance level of 0.05 or 0.01 is customary [3]. If a 0.05 or 5% significance level is chosen in designing a test of hypothesis, then we are 95% confident that we have made the right decision. In such case we say that the hypothesis has been rejected at a 0.05 level of significance, which means that we could be wrong with probability 0.05. For significance level ρ, the acceptance interval $D = [-x_\rho, x_\rho]$ is chosen as the interval in which the random variable z lies with probability $1 - \rho$. Values of ρ which represent the probability of being in $\overline{D}$ for different distributions are available in tables for us to use.

The decision on the hypothesis testing can be reached by the following steps.

- Given N experimental samples of x, compute $\overline{x}$ and z.

• Choose a significance level ρ.
• Compute from the corresponding tables of $\mathcal{N}(0,1)$ the acceptance interval $D = [-x_\rho, x_\rho]$ corresponding to probability $1-\rho$.
• If $z \in D$ accept H_0; otherwise, reject H_0 and accept H_1.

Example 1 From table 1 which is based on $\mathcal{N}(0,1)$, the probability $z \notin [-x_\rho, x_\rho]$ is ρ. For example if $z <$ -1.967 or $z >$ 1.967, we are 95% confidant H_0 is false. In this case we say that the means differ significantly at the 0.05 level.

ρ	0.2	0.15	0.1	0.05	0.02	0.01	0.002
$1-\rho$	0.8	0.85	0.9	0.95	0.98	0.99	0.998
x_ρ	1.282	1.440	1.645	1.967	2.326	2.576	3.090

Table 1: Acceptance intervals $[-x_\rho, x_\rho]$ for $\mathcal{N}(0,1)$.

3.4 Testing the Variances

Many times it is the case that sample spaces have the same value for the mean and different values for the covariance as is shown in Figure 3. Therefore, if the means of a feature are equal on two classes, we still can not discard that feature unless the variances are nearly equal. We can use statistical hypotheses testing to test for the variances. In this case the null and alternative hypothesis are be given by

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{and} \quad H_1 : \sigma_1^2 \neq \sigma_2^2$$

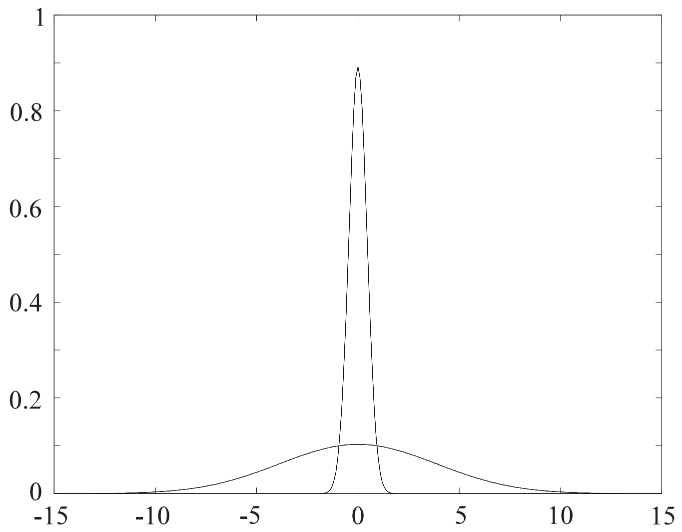

Figure 3: Gaussian pdf's with same mean and different variances.

Now, to check whether the two randomly drawn samples from the classes w_1 and w_2 have the same variance (i.e., to test H_0), we use the F statistic

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$$

If we assume H_0, then

$$F = \frac{s_1^2}{s_2^2}$$

F in the above formula has the F distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom and hence tables such as Table 5.2 in the text book need to be used for testing the hypothesis. Furthermore, the F test is reliable only if the original data comes from approximate normal distributions. Also one of the properties of the F distribution distribution is that it approaches the normal distribution when n_1 and n_2 are large.

4 Individual Feature Selection Based on The Receiver Operating Characteristic Curve

Statistical hypothesis testing is used to discard features with similar values. This doesn't guarantee that the remaining features have good discrimination capabilities among the different classes. For example, we could have a scenario similar to what is shown in Figure 4.a where the means of a given feature are different. However, because of the high values of the variances, there is a large overlap between the pdf's that describe the distribution of a feature in two classes. In this section we talk about a measure of class discrimination capabilities between the different classes for the given features. This method is called the receiver operating characteristic curve (ROC curve) and it can be classified as a feature selection method where features are considered individually.

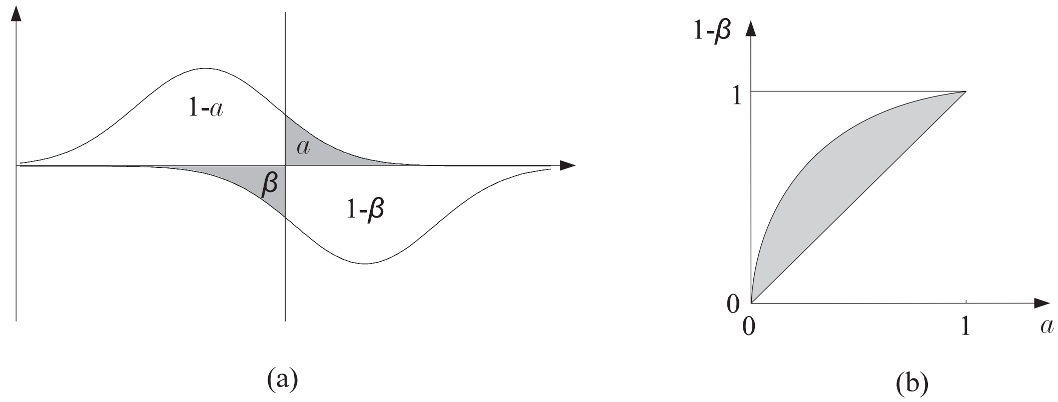

Figure 4: Example of (a) overlapping pdf's of the same feature in two classes and (b) the resulting ROC curve.

The pdf's and a threshold line are drawn in Figure 4.a where the pdf of class w_2 has been inverted for illustration purposes. Let a denotes the probability of error in predicting class w_1 and let $1 - \beta$ denotes the probability of correctness when predicting class w_2. By plotting

$1-\beta$ versus a and sweeping the threshold line over all possible values, we get the diagram shown in Figure 4.b. When the two pdf's completely overlap, $a = 1-\beta$ which corresponds to the straight diagonal line in Figure 4.b. When there is no overlap, value of a is zero and $1-\beta = 1$ which corresponds to the horizontal line in Figure 4.b. As the two pdf's move apart, values of $1-\beta$ increase while values of a decrease. Therefore, the less the overlap between the two classes, the larger the area between the curve and the straight line. The area is zero for complete overlap and 1/2 for no overlap. Thus this area gives a measure of class discrimination capabilities for the feature under consideration. The OCR curve can be constructed using the training data by sweeping the threshold line and computing the percentages of wrong and correct classifications.

References

[1] Andrews, Harry C., Introduction to Mathematical Techniques in Pattern Recognition. John Wiley, Inc., 1972.

[2] Huber P.J. Robust Statistics, Joh Wiley, 1981.

[3] Spiegel, Murray R., Schaum's Outline Series: Probability and Statistics, McGraw-Hill, 1975.