Lecture 2: Bayesian–Based Classifiers

Instructor: Stephen D. Scott Student: *Ling Zhong*

January 26, 1999

1 Introduction

Pattern recognition is the discipline of classfication objects into a number of classes by classifier. It generalize the problem of classifier design in a pattern recognition system. Our idea here is to classify an unknown pattern in the most probable classes. Then here comes the task of defining what "most probable" means.

Given M classes, $\omega_1, \omega_2, \ldots, \omega_M$, and an unknown pattern represented by a feature vector x, find the conditional probabilities: $P(\omega_i|x)$, $\forall i = 1, \ldots, M$
this is also called *posteriori probabilities.*, each of them represents the probability that the unknown pattern belongs to the respective class ω_i given it takes the value x. The unknown pattern is then assigned to the class corresponding to the maximun value of $P(\omega_i|x)$.

We start by reviewing the Bayes rule, then we raise the issue of defining the Bayesian classifier based on Bayes decision theory, finally we focus on the technique to estimate the density function *pdf* that yields $P(\omega_i|x)$ based on the available training data and classify to ω_i that maximizes it.

2 Review of the Bayesian Rule

Recall from *total probability* and the *Bayes Rule*, we know that:

Let $P(i)$, $i = 1, \ldots, M$ be M events such that $\sum_i P(\omega_i) = 1$, $i = 1, \ldots, M$.
Then the probability of an arbitary event x is given by :

$$p(x) = \sum_i p(x|\omega_i)P(\omega_i), \quad i = 1, \ldots, M$$

Here the *pdf* $p(x|\omega_i)$ is called the likelihood function of ω_i with respect to x.

The well-known *Bayes rule is given as following:*

$$P(\omega_i|x) = \frac{p(x|\omega_i)P(\omega_i)}{p(x)}$$

Our Bayes Decision theory for classify a random unknown feature is based on this Rule.

3 Bayesian–Based Classifiers

3.1 Bayesian Decision Theory

Thanks to the Bayesian Rule,now we know how to compute the conditional probability, we need to compute $P(\omega_i)$, pdf and $p(x)$. Let's focus on the two-class case. Let ω_1,ω_2 be the two classes in *which our patterns belong. The priori probabilities $P(\omega_1)$ and $P(\omega_2)$ can be estimated by training data:*

$$P(\omega_1) \approx \frac{N_1}{N}, \quad P(\omega_2) \approx \frac{N_2}{N}$$

N is the total number of available training patterns,
N_1 = Number of patterns in ω_1, N_2 = Number of patterns in ω_2,
Notice here we assume N is sufficiently large. The estimate of pdf will be discussed later. Also notice from the review section, $p(x)$ is the same for all ω_i.

Now the Bayes classification can be stated as:

If $P(\omega_1|x) > P(\omega_2|x)$, x is classified to ω_1; if $P(\omega_1|x) < P(\omega_2|x)$, x is classified to ω_2

Notice in the case of equality, the pattern can be assigned to either class since they have same possibility to belong to ω_1 and ω_2, these are the points in the decision surface which we will discuss later.

Furthermore, we assume the priority probabilities *are equal, which is $P(\omega_1) = P(\omega_2) = 1/2$, then the decision is simplified to find out*

$$p(x|\omega_1) > P(x|\omega_2)$$

or

$$p(x|\omega_2) > P(x|\omega_1)$$

3.2 Probability of Error

Figure 1 presents an example of two equiprobable classes where l=1, it shows the variations of $p(x|\omega_i)$, $i = 1, 2$ as the functions of x. According the Bayes rule, we divide the whole space into two regions R_1 and R_2. All the values of x in R_1 are consided to be in class ω_1 whereas the values of x in R_2 are consided to be in class ω_2. Now it's easy to see from the figure that there's possibility of error which is represent by the shaded area.

- *Minimizing the Classfication Probability*

  Initially we assign the same importance to all errors. Then it comes out that Bayesian classifier is optimal,the proof is as following:

  Proof: *the total error probability:*

$$\begin{aligned} P_e &= P(x \in R_2, \omega_1) + P(x \in R_1, \omega_2) \\ &= P(x \in R_2|\omega_1)P(\omega_1) + P(x \in R_1|\omega_2)P(\omega_2) \\ &= P(\omega_1)\int_{R_2} p(x|\omega_1)\,dx + P(\omega_2)\int_{R_1} p(x|\omega_2)\,dx \\ &= \int_{R_2} p(\omega_1|x)p(x)\,dx + \int_{R_1} p(\omega_2|x)p(x)\,dx \qquad (3.1) \end{aligned}$$

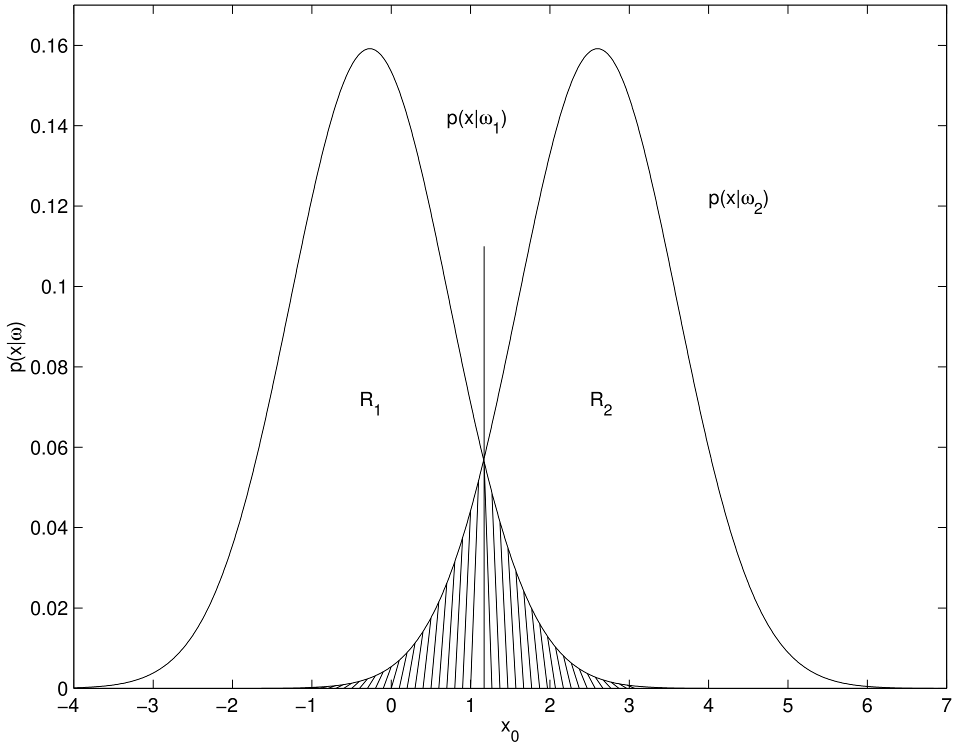

Figure 1: Example of the two region R_1 and R_2 formed by the Bayesian classifier for the case of two equiprobable classes.

Since the region of R_1 and R_2 cover the whole space, from the definition of pdf, we get:

$$\int_{R_1} p(\omega_1|x)p(x)\,dx + \int_{R_2} p(\omega_1|x)p(x)\,dx = P(\omega_1) \qquad (3.2)$$

Combining equations (3.1) and (3.2), we get :

$$P_e = P(\omega_1) - \int_{R_1} (p(\omega_1|x) - p(\omega_2|x))p(x)\,dx \qquad (3.3)$$

Obviously, P_e is minimized if :

$$R_1 = \{x \in \Re^l : P(\omega_1|x) > P(\omega_2|x)\},$$

The similar conclusion holds for R_2. This is what Bayesian classifier does!

Also, we can easily apply the conclusion to multiply classes. For example, in the case of M classes and an unknown pattern x, assign x to ω_i such that:

$$P(\omega_i|x) > P(\omega_j|x), \quad \forall j \neq i \qquad (3.4)$$

- *Minimizing Risk*

  The above solution is based on the assumption that all the errors have the same weight. Unfortunately, in real life, things are probably that different errors may have different effect. Think of the instance of cancer diagnosis, what about if a healthy person is told to be a cancer patient, on the contrary, what about a patient's illness is overlooked and become worse. The result is absolutely different. Because that different error may have different penalties, our reserch now start with assign different weight to errors.

  Again we will consider a M class problem, define λ_{ki} to be loss(penalty) if we predict ω_i whereas the it actually belongs to ω_k, a loss matrix *is formed in such a way. The risk or loss associated with ω_k is definced as:*

$$r_k = \sum_{i=1}^{M} \lambda_{ki} \int_{R_1} p(x|\omega_k)\, dx \qquad (3.5)$$

  It's the overall probabilty of a feature vertor x from class ω_k being classified to ω_i, our goal is to minimize the average risk *which is :*

$$\begin{aligned} r &= \sum_{k=1}^{M} r_k P(\omega_k) \\ &= \sum_{k=1}^{M} \int_{R_1} (\sum_{i=1}^{M} \lambda_{ki} p(x|\omega_k) P(\omega_k))\, dx \qquad (3.6) \end{aligned}$$

  This can be achieved by minimizing each integral, this means to find the region ω_i , such that:

$$R_i = \{x \in \Re^l : \sum_{k=1}^{M} \lambda_{ki} p(x|\omega_k) P(\omega_k) < \sum_{k=1}^{M} \lambda_{kj} p(x|\omega_k) P(\omega_k), \forall j \neq i\} \qquad (3.7)$$

  Notice that here if we assign 1 to λ_{ki} for $k \neq i$, assign 0 to λ_{ki} for $k = i$,(this means we treat each error equally), the above conclusion is exactly the problem to minimzing the classification error probability which we discussed above.

  Now let me look at an two-class case ($l = 2$), assume that

$$P(\omega_1) = P(\omega_2) = 1/2, \quad L = \begin{pmatrix} 0 & \lambda_{12} \\ \lambda_{21} & 0 \end{pmatrix}$$

  and $\lambda_{21} > \lambda_{12}$ (this means misclassification of patterns that comes from ω_2 have more serious consequence).

  Then it comes:

$$\begin{aligned} R_2 &= \{x \in \Re^2 : \lambda_{21} p(x|\omega_2) > \lambda_{12} p(x|\omega_1)\} \\ &= \{x \in \Re^2 : p(x|\omega_2) > p(x|\omega_1) \frac{\lambda_{12}}{\lambda_{21}}\}, \qquad (3.8) \end{aligned}$$

$$\begin{aligned} R_1 &= \{x \in \Re^2 : \lambda_{12} p(x|\omega_1) > \lambda_{21} p(x|\omega_2)\} \\ &= \{x \in \Re^2 : p(x|\omega_1) > p(x|\omega_2) \frac{\lambda_{21}}{\lambda_{12}}\}, \quad (3.9) \end{aligned}$$

Compare Eqs. (3.8) and (3.9), we find we should increase the area of R_2 *and decrease the area of* R_1 *in this case since* $\lambda_{21} > \lambda_{12}$*. In another word, to move the threshold to the left of* x_0*. The opposite holds for the case* $\lambda_{12} > \lambda_{21}$*.*

3.3 Discriminant Functions

- *Decision surface*

  It is clear now that for a feature space with M *regions, if regions* R_i*,* R_j *are contiguous, they are separated by a decision surface in the multidimensional feature space. It is described by the equation:*

  $$P(\omega_i|x) - P(\omega_j|x) = 0$$

  From one side of the surface the difference is positive and from the other it is negative.

- *Discriminant Functions*

  Sometimes instead of using probability function directly, we use a equivalent function of it since it's easier to work with, this function is the so called Discriminant Function.

  $$e.g. \quad g_i(X) \equiv f(P(\omega_i|x))$$

  f(.) is monotonically increasing function, $g_i(x)$ *is the Discriminant Function.*
  Now the decision test becomes:

  $$classify x in \omega_i \quad if g_i(x) > g_j(x), \quad \forall j \neq i$$

  and the decision surface(separating contiguous regions) is :

  $$g_{ij}(x) \equiv g_i(x) - g_j(x) = 0. \quad i, j = 1, 2, \ldots, M \quad i \forall i \neq j$$

  Common choice of $f(.)$ *is natural logarithm where multiplication become sums.*

  This approach can relieve our load, but still it need a good estimate of pdf, there is some cases that the estimation of pdf is really tough, so that we will compute decision surface directly by means of alternative costs, we will touch that issue in the future.

3.4 Normal Distribution

- *Normal Distribution*

  Gaussian(normal) density function is the most commonly encounted pdf in practice, Now we will look at the case when our pdf follows Normal Distribution.

$$p(x|\omega_i) = \frac{1}{(2\pi)^{\frac{l}{2}}|\sum_i|^{\frac{1}{2}}} exp(-\frac{1}{2}(x-\mu_i)^T \sum_i{}^{-1}(x-\mu_i)), \quad i = 1, \ldots, M$$

. $\mu_i = E[x] = mean\ value\ of\ \omega_i\ class$
. $\sum_i = E[(x-\mu_i)(x-\mu_i)^T]$
$|\sum_i|$ *is the determinant of* $\sum_i$, ω_i*'s* *covariance matrix*.

Assuming we already know μ_i *and* $\sum_i, \forall i$

- *Discriminant Functions*

  The Discriminant Functions we use here is as following:

$$g_i(x) = ln(p(x|\omega_i)P(\omega_i))$$

  For normal distribution we get:

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T \sum_i{}^{-1}(x-\mu_i) + ln(P(\omega_i)) - \frac{l}{2}ln(2\pi) - \frac{1}{2}ln|\sum_i|$$

- *Minimum Distance Classifiers*

  For simplification, we consider the case when all the $P(\omega_i)$ *is equal, so do all the* $\sum_i$*, then*

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T \sum{}^{-1}(x-\mu_i)$$

  (Notice here we neglect all the constant.)

  Furthermore, if we assure all the features are statistically independent with the same variance, which means $\sum = \sigma^2 I$, g_i *is even simplified which is*

$$g_i(x) = -\frac{1}{2}\sum_{j=1}^{l}(x_j - \mu_{ij})^2$$

  Now maximum $g_i(x)$ *becomes to minimum:*

$$Euclidean distance : d_\epsilon = ||x - \mu_i||$$

  Thus feature vectors are assighed to class according to their Euclidean distance to the mean points.

  What is some features are dependent on each other, then $\sum$ *is not diagonal, in this case maximum* $g_i(x)$ *becomes minimum:*

$$Mahalanobis distance : d_m = \sqrt[2]{(x-\mu_i)^T \sum{}^{-1}(x-\mu_i)}$$

  Now the constant distance $d_m = c$ *becomes ellipses.*