Machine Learning: Tools, Techniques, Applications
(2013-14-I)
1

Aug. 2013

Chapter 2

Bayes decision theory (Ver:0.8)

We observed earlier that the dependence of the predicted variable on the feature vector is stochastic. Consequently, the expected or average error made by any predictor on $\mathbf{X}$ will be non-zero. One question that naturally arises is: what is the least expected error that the best (that is most accurate) predictor will make. This will act as a lower bound on the expected error of any predictor.

Let us understand this through an example. Assume we wish to predict the geographic origin of a person from India as *North* if the person's place of origin is north of the parallel close to Nagpur otherwise as *South*. Let *North* correspond to class ω_1 and *South* correspond class ω_2. We will abuse notation a bit and let ω_1 and ω_2 denote both the label and the set of objects with that label — it will be apparent from the context which is meant. Now suppose we ask the predictor to label a person picked at random from India. How will the predictor do this if no information is available about the individual? One possibility that immediately suggests itself is to use the population ratios of the north and south. So, if f_N is the fraction of the total population that is from the north and f_S the fraction from the south then a simple rule is:

if $(f_N > f_S)$ then ω_1 else ω_2.

If we let f_N be identified with $P(\omega_1)$ and f_S with $P(\omega_2)$ the apriori probabilities for classes ω_1 and ω_2 respectively then we can write the above rule in terms of the apriori probabilities:

if $(P(\omega_1) > P(\omega_2))$ then ω_1 else ω_2.

Now assume, in addition, that we have some information related to skin colour of people in ω_1 and ω_2. These distributions, assumed as similar normal distributions for both classes except for a shifted mean, are shown in figure 2.1 for both classes. Now when we get the feature vector $\mathbf{x}$ of an unlabelled individual whose skin colour feature value is x we can modify our labelling rule to (note $\mathbf{x}$ is a one-dimensional vector):

if $(p(\omega_1|\mathbf{x}) > p(\omega_2|\mathbf{x}))$ then ω_1 else ω_2.

That is given x if the conditional probability for ω_1 is greater then we give the label ω_1 otherwise ω_2. If the class conditional distributions in figure 2.1 and the apriori probabilties are known we can use Bayes rules to get $p(\omega_1|\mathbf{x})$ and $p(\omega_2|\mathbf{x})$ by:

$$p(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{P(\mathbf{x})}\ i \in 1..2$$

If we assume that $\mathbf{X}$ is a d dimensional space then the various class conditional distributions will carve the space into regions or volumes where for each region one of the conditional probabilites $p(\omega_i|\mathbf{x})$ is highest — that is larger than $p(\omega_j|\mathbf{x})\ j \neq i$. If $\mathbf{x}$ is in that region then we give it the corresponding label ω_i. For example in figure 2.1 if $\mathbf{x}$ is in region $\mathbf{R_1}$ then it will be labelled ω_1, if it is in $\mathbf{R_2}$ it will be labelled ω_2.

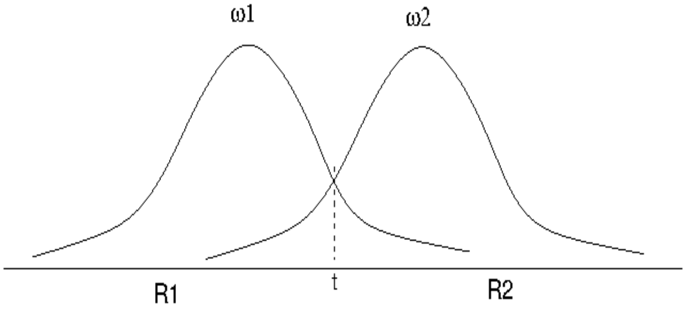

Figure 2.1: Class conditional distribution $p(x|\omega_1)$, $p(x|\omega_2)$ of skin colour for ω_1, ω_2. $\mathbf{R_1}$ is region to left of $\mathbf{t}$ and $\mathbf{R_2}$ the region to its right.

At the point where $\mathbf{R_1}, \mathbf{R_2}$ intersect either label can be give. This is the general Bayes decision rule. Note that in $\mathbf{R_1}$ there is a finite probability of finding objects of class ω_2 and similarly in $\mathbf{R_2}$ there is a finite probability of finding objects belonging to ω_1. The Bayes decision rule will misclassify such objects. The total error probability is clearly the area under the curve — shown shaded in figure 2.2. Intuitively, we can see that if we move the threshold point $\mathbf{t}$ in figure 2.2 either to the right or left the shaded area under the curve increases. Thus, the minimum error is obtained as per the Bayes decision rule. This can be proved in a more rigorous manner but we shall not do that here. More generally we can write:

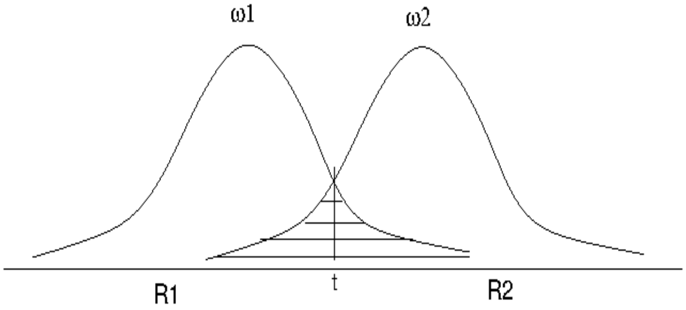

Figure 2.2: Class conditional distribution $p(x|\omega_1)$, $p(x|\omega_2)$ of skin colour for ω_1, ω_2.

$$p(error) = \sum_{i=1}^{C} p(error|\omega_i)P(\omega_i)$$

where $p(error|\omega_i)$ is the probability of misclassifying objects from class ω_i. This is given by:

$$p(error|\omega_i) = \int_{\mathbf{R_i^c}} p(x|\omega_i) d\mathbf{x}$$

where $\mathbf{R_i^c}$ is the complement of region $\mathbf{R_i}$ in the feature space $\mathbf{X}$. So,

$$\begin{aligned} p(error) &= \sum_{i=1}^{C} P(\omega_i) \int_{\mathbf{R_i^c}} p(x|\omega_i) d\mathbf{x} \\ &= \sum_{i=1}^{C} P(\omega_i)[1 - \int_{\mathbf{R_i}} p(\mathbf{x}|\omega_i) d\mathbf{x}] \\ &= 1 - \overbrace{[\sum P(\omega_i) \int_{\mathbf{R_i}} p(\mathbf{x}|\omega_i) d\mathbf{x}]}^{\mathbf{A}} \qquad (2.1) \end{aligned}$$

If we want to minimize $p(error)$ then we have to maximize the value of the expression $\mathbf{A}$ in equation 2.1. So, we must choose regions $\mathbf{R_i}$ such that $\int_{\mathbf{R_i}} P(\omega_i)p(\mathbf{x}|\omega_i)d\mathbf{x}$ is a maximum. The integral will be a maximum when we choose $\mathbf{R_i}$ as the region over which $P(\omega_i)p(\mathbf{x}|\omega_i)$ is the largest over all classes — that is $\int_{\mathbf{X}} \underset{i}{max}\ [P(\omega_i)p(\mathbf{x}|\omega_i)]d\mathbf{x}$, over the whole feature space $\mathbf{X}$. This means the probability of a correct label is:

$$p(correct) = \int_{\mathbf{X}} \underset{i}{max}\ [P(\omega_i)p(\mathbf{x}|\omega_i)]d\mathbf{x}$$

and probability of error is $p(error) = 1 - p(correct)$. But notice that $p(correct)$ is essentially Bayes rule. Therefore, Bayes rule indeed does minimize the probability of error.

When different classification errors are made they are not all equally costly. For example, if tumours are being classified as malignant or benign. A classification error that classifies a malignant tumour as benign is much more costly than one where a benign tumour is misclassified as malignant. In the first case it could mean the death of an individual while in the second further tests that are usually done before treatment is started should be able to catch the mistake. Therefore, we would like to give different weights to different errors. Let λ_{ij} be the weight when true label ω_i is misclassified as ω_j. This weighted probability is often called *risk* or *loss*. For the case where there are only two labels the risk r is:

$$r = \lambda_{12} \int_{R_2} P(\omega_1)p(\mathbf{x}|\omega_1)d\mathbf{x} + \lambda_{21} \int_{R_1} P(\omega_2)p(\mathbf{x}|\omega_2)d\mathbf{x}$$

The integral $r_1 = \int_{R_2} p(\mathbf{x}|\omega_1)d\mathbf{x}$ is called the risk or loss associated with giving a wrong label ω_2 to an object whose real label is ω_1. More generally for C class labels the risk or loss for label ω_k is:

$$r_k = \sum_{i=1}^{C} \lambda_{ki} \int_{R_i} p(\mathbf{x}|\omega_k)d\mathbf{x}$$

The goal is to choose regions $\mathbf{R_i}$ such that expected or average risk is minimized. The expected risk is:

$$\begin{aligned} r &= \sum_{k=1}^{C} P(\omega_k) r_k \\ &= \sum_{k=1}^{C} P(\omega_k) \sum_{i=1}^{C} \lambda_{ki} \int_{\mathbf{R_i}} p(\mathbf{x}|\omega_k)d\mathbf{x} \\ &= \sum_{k=1}^{C} P(\omega_k) \int_{\mathbf{R_i}} \left(\sum_{i=1}^{C} \lambda_{ki} p(\mathbf{x}|\omega_k) \right) d\mathbf{x} \qquad (2.2) \end{aligned}$$

We would like to minimize the risk 2.2 by choosing the regions $\mathbf{R_i}$ suitably.

From analogy with the previous discussion we see that it is really a weighted version of the Bayes decision rule with λ_{ki} as weights. The loss/risk can be specified by a loss/risk matrix:

$$\begin{bmatrix} \lambda_{11} & \dots & \lambda_{1C} \\ \vdots & \ddots & \vdots \\ \lambda_{C1} & \dots & \lambda_{CC} \end{bmatrix}$$

Usually λ_{ii} is 0 — that is there is no classification error. When we have $0 - 1$ loss defined by:

$$\lambda_{ki} = \begin{cases} 0 & k = i \\ 1 & k \neq i \end{cases}$$

then we get the minimum error classification error probability.

A common and useful special case is when we have only two class labels.

$$r_1 = \lambda_{11} p(\mathbf{x}|\omega_1) P(\omega_1) + \lambda_{21} p(\mathbf{x}|\omega_2) P(\lambda_{2)}$$
$$r_2 = \lambda_{21} p(\mathbf{x}|\omega_1) P(\omega_1) + \lambda_{22} p(\mathbf{x}|\omega_2) P(\lambda_{2)}$$

For an unknown vector $\mathbf{x}$ we give label ω_1 if $r_1 < r_2$ otherwise label ω_2. Assuming that $\lambda_{ij} > \lambda_{ii}$ we can write the above as:

$$label(x) = \begin{cases} \omega_1 & \text{if } \left(\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} > \frac{P(\omega_2)}{P(\omega_1)} \frac{\lambda_{21} - \lambda_{22}}{\lambda_{12} - \lambda_{11}} \right) \\ \omega_2 & \text{otherwise} \end{cases}$$

Notice that this is again Bayes rule if we assume $\lambda_{11} = \lambda_{22} = 0$ and $\lambda_{12} = \lambda_{21}$. For better intuition consider the effect of different values of λ_{12} and λ_{21} on the threshold $\mathbf{t}$ in figure 2.1. Assuming $(P(\omega_1) = P(\omega_2))$, we get $\mathbf{x} \in \omega_1$ if $(p(\mathbf{x}|\omega_1) > p(\mathbf{x}|\omega_2) \frac{\lambda_{21}}{\lambda_{12}})$ otherwise it is in ω_2. So, threshold $\mathbf{t}$ moves left if $(\lambda_{21} > \lambda_{12})$ otherwise it moves right.