Bayesian Decision Theory

An approach to two-category classification

Tim Haines
ECE486
April 8, 2008

Image courtesy of wikimedia.org (http://upload.wikimedia.org/wikipedia/en/thumb/f/fe/Unbalanced_scales.svg/400px-Unbalanced_scales.svg.png)

Bayesian Decision Theory Assumptions

1) There are only two states of nature.

We are examining something which has only two possible values. These values are referred to as ω_1 and ω_2.

For example, the book discusses classifying two kinds of fish: sea bass and salmon. These two types of fish are the only states of nature. Therefore, there can be no other kind of fish considered in this classification.

2) Both states are probabilistic.

The probabilities are comprised of a single random variable and are denoted as $P(\omega_1)$ for state one and $P(\omega_2)$ for state two. These probabilities are also known as the *a priori* probabilities or, simply, the prior probabilities. Because there are only two states of nature, then we know that the two probabilities must sum to one.

Bayesian Decision Theory Assumptions (cont.)

The Probability Density Function (PDF) of each state is fully known.

The PDF of each state describes the complete prior probability for the random variables ω_1 and ω_2. The PDF is a Gaussian (normal) distribution.

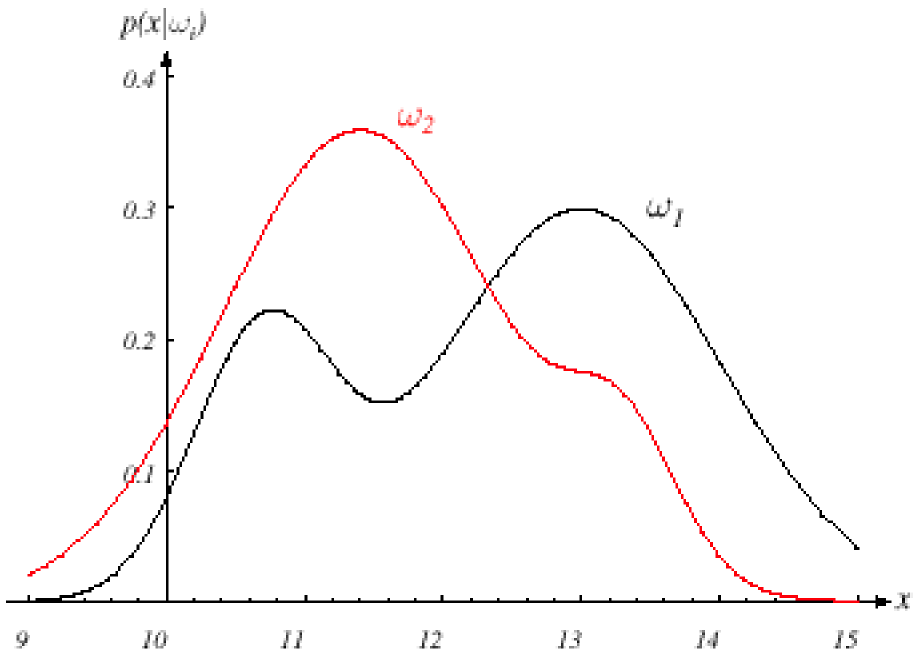
The decision rule

If $P(\omega_1) > P(\omega_2)$, then decide ω_1, else decide ω_2.

This rule works well if we are only going to make the decision once, but if we continually apply the rule to multiple instances of the states, then we always make the same decision. Therefore, we need to consider a factor which does not affect the prior probability but which can be observed each time we make a decision.

Class-Conditional Probability Density Function

Let there be some continuous, random variable *x* that describes an observation that is made each time we make a decision. This observation is dependent on the state of nature and is expressed as p(*x*|ω).



The two curves show the difference in the observation made for each state.

This is much better than using just the prior probabilities, but it still doesn't provide a clear distinction over the full range of observations.

Bayes Formula

Now that we know the class-conditional probability density function, how do we find the probabilities associated with the density function for certain values of x?

Recall:

$p(\omega_j, x) = P(\omega_j|x)p(x) = p(x|\omega_j)P(\omega_j)$ where $p(x|\omega_j)$ is called the likelihood.

Therefore,

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

Using this, we can now convert the prior probability $P(\omega_j)$ into the *a posteriori* probability $P(\omega_j|x)$ – the probability of the state of nature being ω_j given that we made the observation x.

Posterior Probabilities and Bayes Decision Rule

The old Bayes decision rule:

If $P(\omega_1) > P(\omega_2)$, then decide ω_1, else decide ω_2.

The new Bayes decision rule:

if $P(\omega_1|x) > P(\omega_2|x)$, then decide ω_1, else decide ω_2.

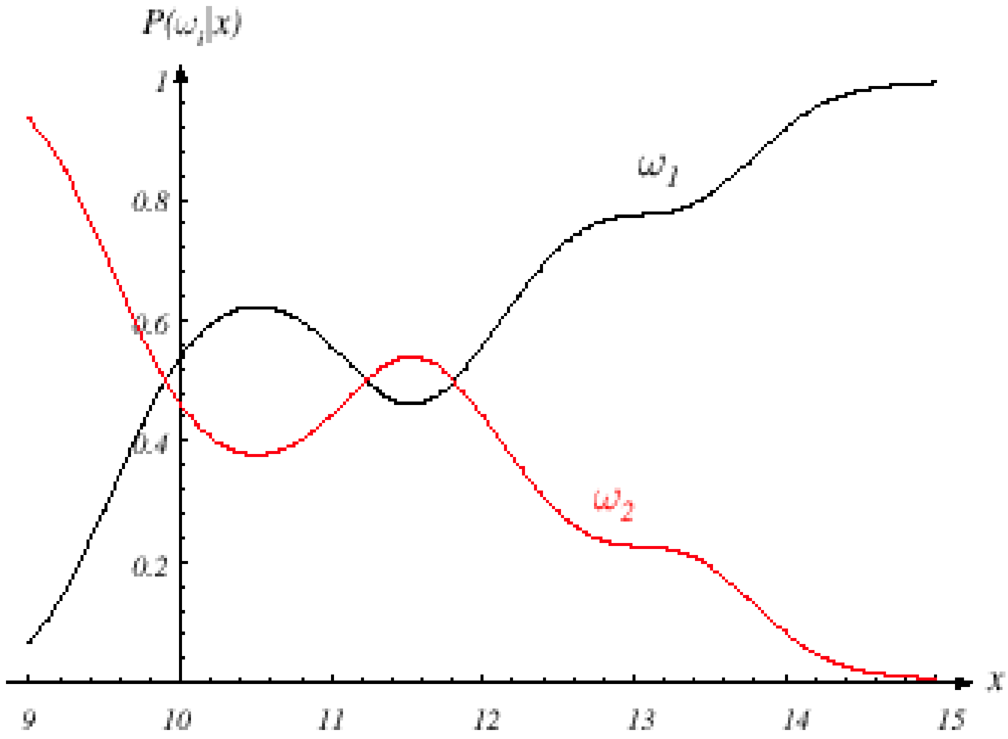
Therefore,

$$P(\text{error}|x) = \min[P(\omega_1|x) , P(\omega_2|x)].$$

Since the factor $p(x)$ is merely a scaling factor for the posterior probability and can thus be removed, then the Bayes decision rule can be written as

if $p(x|\omega_1)P(\omega_1) > p(x|\omega_2)P(\omega_2)$ then decide ω_1, else decide ω_2.

Posterior Probabilities and Bayes Decision Rule (cont.)



Using the posterior probabilities gives us a much larger difference between the two states of nature and thus a much better probability of making the correct decision.

Continuous Features - A More General Approach

Thus far we have restricted ourselves to a rather limited subset of the Bayesian Decision Theory. Now let us generalize in these ways:

Allow more than one feature
Allow more than two states of nature
Allow rejection
General loss function

Allow more than one feature

To do this, we simply replace the scalar x with the vector $\boldsymbol{x}$. This creates a feature space in $\Re^d$ where d is the length of $\boldsymbol{x}$.

Allow more than two states of nature

This involves merely replacing ω_1, ω_2 with ω_j for j=1...n.

Continuous Features - A More General Approach (cont.)

Allow rejection

Instead of being forced to make a decision, we allow ourselves to reject an observation if the probabilities are too close.

General loss function

We no longer assume that making an incorrect decision is equal for all states.

Let there be a finite set of states ω_j for j=1...n and a finite set of actions α_i for i=1...k. The loss function is denoted as $\lambda(\alpha_i|\omega j)$ and describes the loss incurred when we take action α_i when the state of nature is ω_j.

Bayes Risk

The generalized posterior probability

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

We can now compute the expected loss (conditional risk)

$$R(\alpha_i|\boldsymbol{x})=\sum_{j=1}^{n} p(\boldsymbol{x}|\omega_j)P(\omega_j|\boldsymbol{x})$$

To achieve the best performance, we must minimize the risk. Therefore, Bayes Decision rule can be reduced to

$$min(R(\alpha_i|\boldsymbol{x}))$$ for i=1...k

Two-Category Classification Using Bayes Risk

We define the risk for the two states of nature as being

$$R(\alpha_1|\boldsymbol{x})=\lambda_{11}P(\omega_1|\boldsymbol{x})+\lambda_{12}P(\omega_2|\boldsymbol{x})$$

$$R(\alpha_2|\boldsymbol{x})=\lambda_{21}P(\omega_1|\boldsymbol{x})+\lambda_{22}P(\omega_2|\boldsymbol{x})$$

where

$$\lambda_{ij}=\lambda(\alpha_i|\omega_j)$$

The Bayesian Decision Rule then becomes

$$(\lambda_{21}-\lambda_{11})P(\omega_1|\boldsymbol{x})>(\lambda_{12}-\lambda_{22})P(\omega_2|\boldsymbol{x})$$

Substituting the posterior probability and canceling the evidence term

$$(\lambda_{21}-\lambda_{11})p(\boldsymbol{x}|\omega_1)P(\omega_1)>(\lambda_{12}-\lambda_{22})p(\boldsymbol{x}|\omega_2)P(\omega_2)$$

Two-Category Classification Using Bayes Risk (cont.)

We can rearrange the Bayesian Decision rule to produce the likelihood ratio

$$\frac{p(\boldsymbol{x}|\omega_1)}{p(\boldsymbol{x}|\omega_2)} > \frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}} \frac{P(\omega_2)}{P(\omega_1)}$$

We can now use an arbitrary likelihood ratio as a decision factor that is independent of the observation **x**.