Classification using Bayes Rule in 1-dimensional and N-dimensional feature spaces

Jihwan Lee

April 27, 2014

1 INTRODUCTION

In this slecture, we discuss Bayes rule that is widely used for many different kinds of applications, especially, pattern recognition. Due to its simplicity and effectiveness, we can use the method in both discrete values case and continuous values case, and it also usually works well in multi-dimensional feature space. So, we will take a look at what the definition of Bayes rule is, how it can be used for the classification task with examples, and how we can derive it in different cases.

2 BAYES THEOREM

Bayes theorem is a probabilistic theory that can explain a relationship between the prior probability and the posterior probability of two random variables or events. Give two events A and B, we may want to know $P(A|B)$ and it can be obtained if we know knowledge for other probabilities, $P(B|A)$, $P(A)$, and $P(B)$. By the definition of the conditional probability, a joint probability of A and B, $P(A,B)$, is the product of $P(A|B)P(B)$. We can also write $P(A,B) = P(B|A)P(A)$. From those two equations, $P(A|B)P(B) = P(B|A)P(A)$ and we can conclude

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

The formula above is called Bayes Theorem. Let's see an example. Suppose that a man told you that he had a nice conversation with a person. Assuming the populations of male and female are the same, the probability that his conversational partner was a woman is 0.5. However, if he gives you additional information that the conversational partner had long hair, then the probability of the partner being a woman will be higher than 0.5 because women are more likely to have long hair. So, let W be an event that the conversational partner was a women and L be an event that a person has a long haired person. Suppose that 75% of all women have long hair, 15% of men have long hair, and both genders are equally likely. That is, $P(W|L) = 0.75, P(W) = P(M) = 0.5$, and these are what we already know. With these information, we can say that the probability that the person who he had a conversation with was a woman is

$$\begin{aligned} P(W|L) &= \frac{P(L|W)P(W)}{P(L|W)P(W) + P(L|M)P(M)} \\ &= \frac{0.75 * 0.5}{0.75 * 0.5 + 0.15 * 0.5} \\ &= \frac{5}{6} \simeq 0.83 \end{aligned}$$

Here is another example. In a public university, 51% of the students are males. One adult is randomly selected for a survey. It turned out later that the selected survey subject was studying sciences. Also, 10% of students study sciences while 5% of females study sciences. What is the probability that the selected subject is a male? Let's use the following notations:

- $P(M) = 0.51$ because 51% of students are males
- $P(\overline{M}) = 0.49$ because 49% of students are females
- $P(S|M) = 0.1$ because 10% of the male students study sciences
- $P(S|\overline{M}) = 0.05$ because 5% of the female students study sciences

Now, in order to figure out the probability that the survey subject is a male given the subject studies sciences, $P(M|S)$, we can apply Bayes theorem as following

$$\begin{aligned} P(M|S) &= \frac{P(S|M)P(M)}{P(S)} \\ &= \frac{P(S|M)P(M)}{P(S|M)P(M) + P(S|\overline{M})P(\overline{M})} \\ &= \frac{0.1 * 0.51}{0.1 * 0.51 + 0.49 * 0.05} \\ &= \frac{0.051}{0.0755} \\ &\simeq 0.675 = 67.5\% \end{aligned}$$

3 Classification by Bayes Rule

In this section, we investigate how Bayes rule can be used for the task of classification. Suppose that there are many classes $\omega_1, \omega_2, \ldots, \omega_c$ and data that should belong to one of those classes. So, what we want to do is to classify new data that do not have any class information into one of the classes, based on training data whose class membership information is already known. Using Bayes rule, the probability that a new data x belongs to class ω_i is given by

$$\begin{aligned} P(\omega_i|x) &= \frac{\rho(x|\omega_i)P(\omega_i)}{\rho(x)} \\ &= \frac{\rho(x|\omega_i)P(\omega_i)}{\sum_{l=1} c\rho(x|\omega_l)P(\omega_l)} \end{aligned}$$

where ρ is a density function for continuous values. Now we can compare probability values of $P(\omega_i|x)$ for each class and make a decision by taking one with higher probability as following

$$\begin{aligned} & P(\omega_i|x) \geq P(\omega_j|x) \forall\, j = 1, \ldots, c \\ \Longleftrightarrow\ & \frac{\rho(x|\omega_i)P(\omega_i)}{\rho(x)} \geq \frac{\rho(x|\omega_j)P(\omega_j)}{\rho(x)} \forall\, j = 1, \ldots, c \\ \Longleftrightarrow\ & \rho(x|\omega_i)P(\omega_i) \geq \rho(x|\omega_j)P(\omega_j) \forall\, j = 1, \ldots, c \end{aligned}$$

From now on, we consider only two class classification problem for simplicity. Let $g_1(x) = \rho(x|\omega_1)P(\omega_1)$ and $g_2(x) = \rho(x|\omega_2)P(\omega_2)$. Then we can decide class ω_1 for x if $g_1(x) \geq g_2(x)$, otherwise, decide class ω_2. Equivalently, we can define a discriminant function $g(x) = g_1(x) - g_2(x)$ and decide class ω_1 if $g(x) \geq 0$, otherwise class ω_2. We here need to notice followings

- The discriminant function $g(x)$ is not unique. As long as the discriminant function is monotonically increasing, it will make the same decision. (e.g., $g(x) = \ln g_1(x) - \ln g_2(x)$)
- We do not need to know the actual function $g(x)$, but need to know just its sign.

Then what is the expected value of the error when we make decision by following Bayes rule?

$$\begin{aligned} E(error) &= \int_R P(error|x)\rho(x)dx \\ &= \int_R min(P(\omega_1|x), P(\omega_2|x))\rho(x)dx \\ &= \int_R min(\frac{\rho(x|\omega_1)P(\omega_1)}{\rho x}, \frac{\rho(x|\omega_2)P(\omega_2)}{\rho(x)})\rho(x)dx \\ &= \int_R min(\rho(x|\omega_1)P(\omega_1), \rho(x|\omega_2)P(\omega_2))dx \end{aligned}$$

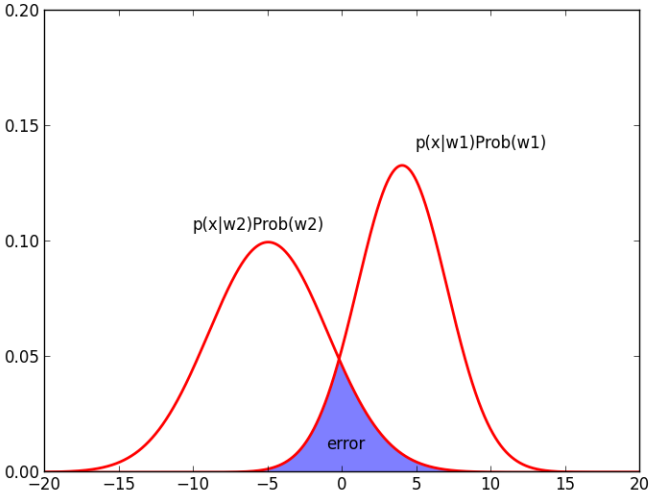

Figure 3.1: Error area of two distributions

Note that $g(x) = 0$ has at most 2 solutions, which means that those two distributions for ω_1 and ω_2 have one intersection or two intersections. Let's look the case of one intersection at the point t.

$$E(error) = \int_{-\infty}^{t} \rho(x|\omega_2)P(\omega_2)dx + \int_{t}^{\infty} \rho(x|\omega_1)P(\omega_1)dx$$
$$= P(\omega_2)\int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}\sigma_2} exp[-\frac{1}{2}(\frac{x-\mu_2}{\sigma_2})^2]dx + P(\omega_1)\int_{\infty}^{t} \frac{1}{\sqrt{2\pi}\sigma_1} exp[-\frac{1}{2}(\frac{x-\mu_1}{\sigma_1})^2]dx$$

Let $y_2 = \frac{x-\mu_2}{\sigma_2}$, $y_1 = \frac{x-\mu_1}{\sigma_1}$, and $\Phi(\alpha) = \int_{-\infty}^{\alpha} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$. Then,

$$E(error) = P(\omega_2)\int_{-\infty}^{\frac{t-\mu_2}{sigma_2}} \frac{1}{\sqrt{2\pi}} exp[-\frac{y_2^2}{2}]dy_2 + P(\omega_1)\int_{\frac{t-\mu_1}{\sigma_1}}^{\infty} \frac{1}{\sqrt{2\pi}} exp[-\frac{y_1^2}{2}]dy_1$$
$$= P(\omega_2)\Phi(\frac{t-\mu_2}{\sigma_2}) + P(\omega_1)\Phi(\frac{t-\mu_1}{\sigma_1})$$

In Figure 3.1, the shaded region represents the probability that data would be misclassified, that is, error. Let t be an intersection point. Based on Bayes rule, we should classify all data on the left of t into ω_2, and all other data that are on the right of t will be classified into ω_1. Therefore, as the region is larger, in other words, two distributions share larger overlapped area, the expected error will be increasing.

The following sections describe two cases of classification. One is where data exist in 1-dimensional feature space and the other is where data exist in N-dimensional feature space.

4 CASE 1: 1-DIMENSIONAL FEATURE SPACE

For simplicity, samples x_i of class ω_i are drawn from Gaussian distribution $x_i \sim N(\mu_i, \sigma_i^2)$. Then, the class-conditional density is given by

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

and let's take

$$\begin{aligned} g_i(x) &= \ln(\rho(x|\omega_i)P(\omega_i)) \\ &= \ln\rho(x|\omega_i) + \ln P(\omega_i) \\ &= \ln\frac{1}{\sqrt{2\pi}\sigma_i} - \frac{1}{2}(\frac{x-\mu_i}{\sigma_i})^2 + \ln P(\omega_i) \\ &= -\frac{1}{2}(\frac{x-\mu_i}{\sigma_i})^2 - \ln\sqrt{2\pi}\sigma_i + \ln P(\omega_i) \\ &= -\frac{1}{2}(\frac{x-\mu_i}{\sigma_i})^2 + C_i \end{aligned}$$

where C_i is a constant value which is independent of x. Finally, we can use the following discriminant function

$$g(x) = -\frac{1}{2}(\frac{x-\mu_1}{\sigma_1})^2 + \frac{1}{2}(\frac{x-\mu_2}{\sigma_2})^2 + C_1 - C_2$$

So, given a value of x, we can assign class to x based on the sign of the discriminant function above.

5 CASE 2: N-DIMENSIONAL FEATURE SPACE

Let's assume that each data is drawn from Multivariate Gaussian distribution with mean μ_i and standard deviation matrix Σ_i. Then the class-conditional density is given by

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

Again we use

$$\begin{aligned} g_i(\vec{x}) &= \ln\rho(\vec{x}|\omega_i)P(\omega_i) \\ &= \ln\rho(\vec{x}|\omega_i) + \ln P(\omega_i) \\ &= -\frac{1}{2}(\vec{x}-\vec{\mu_i})^T\Sigma_i^{-1}(\vec{x}-\vec{\mu_i}) + \ln\frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma_i|^{\frac{1}{2}}} + \ln P(\omega_i) \end{aligned}$$

Note that the first term is actually $-\frac{1}{2}$ square of Mahalanobis distance from $\vec{x}$ to $\vec{\mu}_i$ and the remaining terms are independent of $\vec{x}$, which means that the closer $\vec{x}$ is to $\vec{\mu}_i$, the larger $g_i(x)$ is. We can also observe that if the standard deviation matrix Σ is equal to I(identity matrix) the Mahalanobis distance at the first term of the equation above becomes Euclidean distance, that is,

$$\begin{aligned} g_i(x) &= -\frac{1}{2}\|\vec{x}-\vec{\mu}_i\|_2^2 + \ln\frac{1}{(2\pi)^{\frac{n}{2}}} + \ln\frac{1}{(\Sigma_i)^{\frac{1}{2}}} + \ln P(\omega_i) \\ &= -\frac{1}{2}\|\vec{x}-\vec{\mu}_i\|_2^2 + \ln\frac{1}{(2\pi)^{\frac{n}{2}}} + \ln P(\omega_i) \end{aligned}$$

and the second term can be ignored since it is constant and independent of $\vec{x}$ In two-class classification, we should decide ω_i if

$$-\frac{1}{2}\|\vec{x}-\vec{\mu}_1\|_2^2 + \ln P(\omega_1) \geq -\frac{1}{2}\|\vec{x}-\vec{\mu}_2\|_2^2 + \ln P(\omega_2)$$

otherwise, we should decide ω_2. One interesting observation here is that making decisions will be based on a linear separation, which is a hyperplane. The hyperplane can be defined by a set of points which are equidistant to μ_1 and μ_2 if all priors are equal($P(\omega_i) = \frac{1}{c}\forall\, i = 1,\ldots,c$). Thus, for a given data x, we can choose a class that has a nearest mean from x. Let's see the following derivation:

$$\begin{aligned} & -\frac{1}{2}\|\vec{x}-\vec{\mu}_1\|_2^2 + \ln P(\omega_1) = -\frac{1}{2}\|\vec{x}-\vec{\mu}_2\|_2^2 + \ln P(\omega_2) \\ \iff & -\frac{1}{2}(\vec{x}\cdot\vec{x} - 2\vec{x}\cdot\vec{\mu}_i + \vec{\mu}_i\cdot\vec{\mu}_i) + \ln P(\omega_i) = -\frac{1}{2}(\vec{x}\cdot\vec{x} - 2\vec{x}\cdot\vec{\mu}_j + \vec{\mu}_j\cdot\vec{\mu}_j) + \ln P(\omega_j) \\ \iff & \vec{x}\cdot(\vec{\mu}_i - \vec{\mu}_j) - \frac{1}{2}(\vec{\mu}_i\cdot\vec{\mu}_i - \vec{\mu}_j\cdot\vec{\mu}_j) + \ln P(\omega_i) - \ln P(\omega_j) = 0 \end{aligned}$$

The formula for a hyperplane is $n\cdot x + b = 0$ where n is a normal vector. Thus, the equation above can be representing a hyperplane and the normal vector is $n = \vec{\mu}_i - \vec{\mu}_j$.

We can also consider another case where each class ω_i has the same covariance matrix Σ, that is, $\Sigma_i = \Sigma_j$. Then $g_i(x)$ is given as follows,

$$g_i(\vec{x}) = -\frac{1}{2}(\vec{x}-\vec{\mu}_i)^T\Sigma^{-1}(\vec{x}-\vec{\mu}_i) + \ln P(\omega_i)$$

For this case, the decision boundaries are still hyperplanes, but they may no longer be normal to the lines between the respective class means $\vec{\mu}_i$ and $\vec{\mu}_j$.

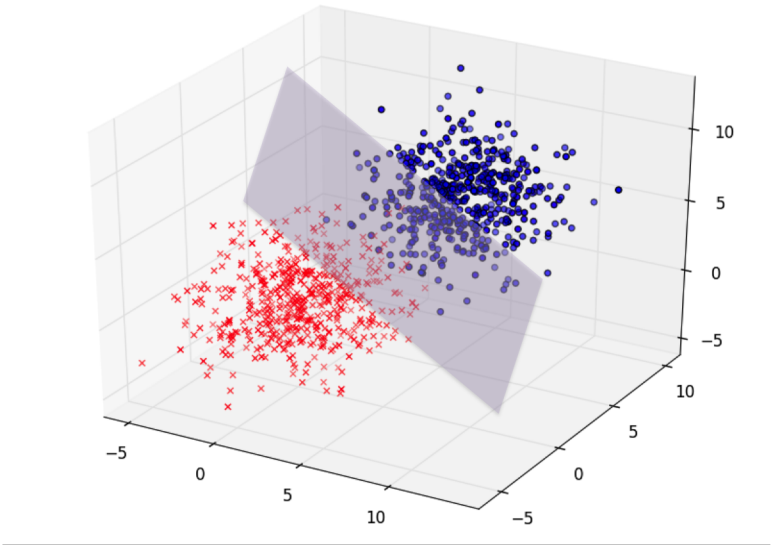

Figure 5.1: Two class classification in 3-dimensional feature space and hyperplane as decision boundary

The other possible case is where there are different covariance matrices for each class, which is actually the most general case. If each class ω_i has its own arbitrary covariance matrix Σ_i, then the decision boundaries are quadratic, specifically, hyperellipses. Figure 5.1 shows classification in three dimensional feature space. If data are drawn from Gaussian distributions with the identity covariance matrices, then the decision boundary between those two classes becomes a hyperplane and its normal vector should be $\vec{\mu}_1 - \vec{\mu}_2$.