Multivariate Gaussian Classifier

Ole Marius Hoel Rindal

December 7, 2015

Bayes' Rule

We can use the Bayes' rule to find an expression for the class with the highest probability

$$p(\omega_j|x) = \frac{p(x|\omega_j)P(\omega_j)}{p(x)} \tag{1}$$

where the probability of being class ω_j given x is equal to the probability of x given ω_j times the prior probability of being class ω_j. This whole thing is over the probability of x, but we're going to argue that we don't need that in a moment. The prior probability of ω_j or $P(\omega_j)$ could be defined differently for each class although in this case we will be assuming the classes are equally likely or $\frac{1}{\#classes}$.

Gaussian Density

Any probability can be used to make $p(x|\omega_j)$, but we want to use the multi-variate Gaussian density.

$$p(\vec{x}|\omega_j) = \frac{1}{\sqrt{(2\pi)^d|\Sigma_j|}}e^{-\frac{1}{2}(\vec{x}-\vec{\mu_j})^\intercal \Sigma_j^{-1}(\vec{x}-\vec{\mu_j})} \tag{2}$$

where, $\vec{\mu_j}$ is the mean vector for class j for d features. Giving,

$$\vec{\mu_j} = \begin{bmatrix} \mu_j^1 \\ \mu_j^2 \\ \vdots \\ \mu_j^n \end{bmatrix} \tag{3}$$

And Σ_j is the covariance matrix for class j for d features, $|\Sigma_j|$ is its determinant, and Σ_j^{-1} is its inverse.

The Discriminant Function

From (1) we get the discriminant function

$$g_j(x) = p(\omega_j|\vec{x}) = \frac{p(\vec{x}|\omega_j)P(\omega_j)}{p(x)} \tag{4}$$

but we can ignore $p(x)$ since this is only a normalizing factor and simplify the discriminant function by tossing it out since it won't change which class has the highest probability which is all we care about.

$$g_j(x) = p(\vec{x}|\omega_j)P(\omega_j) \tag{5}$$

Our Discriminant Function

Before presenting our combined Gaussian density and Bayes' rule discriminant function we will want to simplify using the knowledge that

$$\ln(N \cdot M) = \ln(N) + \ln(M) \tag{6}$$

when applied to our simplified Bayes' rule classifier

$$g_j(\vec{x}) = \ln(p(\vec{x}|\omega_j)P(\omega_j)) \tag{7}$$
$$= \ln(p(\vec{x}|\omega_j)) + \ln(P(\omega_j)) \tag{8}$$

Now we can use this newly acquired knowledge to eliminate the exponent yielding a simplified multivariate Gaussian Classifier

$$g_j(\vec{x}) = -\frac{1}{2}(\vec{x} - \vec{\mu_j})^\intercal \Sigma_j^{-1}(\vec{x} - \vec{\mu_j}) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_j| + \ln P(\omega_j) \tag{9}$$

Special Cases

Equation (9) can sometimes be simplified given specific knowledge about the covariance matrices.

Case 1: $\Sigma_j = \sigma^2 I$

The features are uncorrelated (independent) and have the same variance. Recalling (9) we can discard everything that is common for all classes, so we get

$$g_j(x) = -\frac{||\vec{x} - \vec{\mu_j}||^2}{2\sigma^2} + \ln P(\omega_j) \tag{10}$$

where,

$$||\vec{x} - \vec{\mu_j}||^2 = (\vec{x} - \vec{\mu_j})^\intercal(\vec{x} - \vec{\mu_j}) \tag{11}$$

which is the Euclidian distance. This is known as a minimum distance classifier.

Case 2: Common covariance matrix

Again, recalling (9), but instead of $\Sigma_j = \sigma^2 I$ we have $\Sigma_j = \Sigma$. Since the covariance matrices are equal we can reduce the discriminant function to

$$g_j(\vec{x}) = -\frac{1}{2}(\vec{x} - \vec{\mu_j})^\intercal \Sigma_j^{-1}(\vec{x} - \vec{\mu_j}) + \ln P(\omega_j) \tag{12}$$

where,
$$(\vec{x} - \vec{\mu_j})^\intercal \Sigma_j^{-1} (\vec{x} - \vec{\mu_j}) \tag{13}$$
equation (13) is the Mahalanobis distance.