Intelligent Systems: Reasoning and Recognition

James L. Crowley

ENSIMAG 2 / MoSIG M1 Second Semester 2015/2016

Lesson 9 16 March 2016

The Normal Density Function

Bibliographical sources:

"Pattern Recognition and Machine Learning", C. M. Bishop, Springer Verlag, 2006.

"Pattern Recognition and Scene Analysis", R. E. Duda and P. E. Hart, Wiley, 1973.

Notation

x	a variable
X	a random variable (unpredictable value)
M	The number of possible values for X (Can be infinite).
$\vec{x}$	A vector of D variables.
$\vec{X}$	A vector of D random variables.
D	The number of dimensions for the vector $\vec{x}$ or $\vec{X}$
E	An observation. An event.
C_k	The class k
k	Class index
K	Total number of classes
ω_k	The statement (assertion) that $E \in C_k$
$P(\omega_k) = P(E \in C_k)$	Probability that the observation E is a member of the class k.
M_k	Number of examples for the class k.
M	Total number of examples. $M = \sum_{k=1}^{K} M_k$
$\{\vec{X}_m\}$	A set of training samples
$\{\vec{y}_m\}$	A set of indicator vectors for the training samples in $\{\vec{X}_m\}$ $\vec{y}_n$ is a vector of binary values, with 1 for the kth component and 0 elsewhere.
$p(X)$	Probability density function for X
$p(\vec{X})$	Probability density function for $\vec{X}$
$p(\vec{X}/\omega_k)$	Probability density for $\vec{X}$ the class k. $\omega_k = E \in C_k$.

Bayesian Classification

In the next few lectures we will look at techniques that learn to recognize classes with an explicit probabilistic model of the training data using variations of the Gaussian or Normal density functions.

As before, our problem is to build a box that maps a set of features $\vec{X}$ from an observation pnto a class C_k from a set of K possible classes.

Let ω_k be the proposition that the event belongs to class k: $\omega_k = E \in C_k$

ω_k Proposition that event E $\in$ the class C_k

In order to minimize the number of mistakes, we will maximize the probability that $\omega_k \equiv E \in C_k$

$$\hat{\omega}_k = \arg\max_{\omega_k} \left\{ P(\omega_k \mid \vec{X}) \right\}$$

Our primary tool for this is Baye's Rule : $P(\omega_k \mid \vec{X}) = \dfrac{P(\vec{X} \mid \omega_k) P(\omega_k)}{P(\vec{X})}$

To apply Baye's rule, we require a representation for the probabilities $P(\vec{X} \mid \omega_k)$, $P(\vec{X})$, and $p(\omega_k)$. The term $p(\omega_k)$ is a number that represents the a-priori probability of encountering an event of class C_k.

Today we will use the Normal Density function to represent probability:

$$p(\vec{X}) = \mathcal{N}(\vec{X}; \vec{\mu}, \Sigma) = \frac{1}{(2\pi)^{\frac{D}{2}} \det(\Sigma)^{\frac{1}{2}}} e^{-\frac{1}{2}(\vec{X}-\vec{\mu})^T \Sigma^{-1} (\vec{X}-\vec{\mu})}$$

and

$$p(\vec{X} \mid \omega_k) = \mathcal{N}(\vec{X}; \vec{\mu}_k, \Sigma_k) = \frac{1}{(2\pi)^{\frac{D}{2}} \det(\Sigma_k)^{\frac{1}{2}}} e^{-\frac{1}{2}(\vec{X}-\vec{\mu}_k)^T \Sigma_k^{-1} (\vec{X}-\vec{\mu}_k)}$$

Probability Density Functions

A probability density function *p(X),* is a function of a continuous variable *X* such that

1) *X* is a continuous real valued random variable with values between $[-\infty, \infty]$
2) $\int_{-\infty}^{\infty} p(X) = 1$

Classification using Bayes Rule also works with probability density functions

$$P(\omega_k \mid X) = \frac{p(X \mid \omega_k)}{p(X)} P(\omega_k) = \frac{p(X \mid \omega_k)}{\sum_{k=1}^{K} p(X \mid \omega_k)} P(\omega_k)$$

Note that *p(X)* is NOT a number but a continuous function.

A probability density function defines the relatively likelihood for a specific value of *X*. Because *X* is continuous, the value of *p(X)* for a specific *X* is infinitely small. To obtain a probability we must integrate over some range of *X*.

To obtain a probability we must integrate over some range V of X.
In the case of D=1, the probability that X is within the interval [A, B] is

$$P(X \in [A,B]) = \int_{A}^{B} p(x)dx$$

This integral gives a number that can be used as a probability.

Note that we use upper case $P(X \in [A,B])$ to represent a probability value,
and lower case *p(X)* to represent a probability density function.

Also note that the ratio $\frac{p(X \mid \omega_k)}{p(X)}$ IS a probability, provided that

$$p(X) = \sum_{k=1}^{K} p(X \mid \omega_k)$$

Probability density functions are easily generalized to vectors of random variables.
Let $\vec{X} \in R^D$, be a vector random variables.

A probability density function, $p(\vec{X})$, is a function of a vector of continuous variables

1) $\vec{X}$ is a vector of D real valued random variables with values between $[-\infty, \infty]$

2) $\int_{-\infty}^{\infty} p(\vec{x})d\vec{x} = 1$

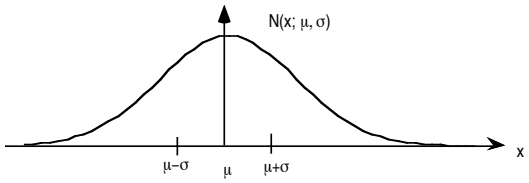
Note that while $p(\vec{X})$ and $p(\vec{X} \mid \omega_k)$ are NOT numbers, their ratio IS a number.

The Normal Density Function

Univariate Normal

Whenever a random variable is determined by a sequence of independent random events, the outcome will be a Normal or Gaussian density function. This is demonstrated by the Central Limit Theorem. The essence of the derivation is that repeated convolution of any finite density function will tend asymptotically to a Gaussian (or normal) function.

$$p(X) = \mathcal{N}(X;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(X-\mu)^2}{2\sigma^2}}$$



The parameters of $\mathcal{N}(x;\ \mu,\ \sigma)$ are the first and second moments, μ and σ of the function.

According to the Central Limit theorem, for any real density $p(X)$:

$$as\ M \to \infty \quad p(X)^{*M} \to \mathcal{N}(x;\ \mu,\ \sigma)$$

Multivariate Normal

For a vector of D features, $\vec{X}$, the Normal density has the form:

$$p(\vec{X}) = \mathcal{N}(\vec{X};\vec{\mu},\ \Sigma) = \frac{1}{(2\pi)^{\frac{D}{2}}\ \det(\Sigma)^{\frac{1}{2}}} e^{-\frac{1}{2}(\vec{X}-\vec{\mu})^T\ \Sigma^{-1}(\vec{X}-\vec{\mu})}$$

There are 3 parts to $\mathcal{N}(\vec{X};\vec{\mu},\ \Sigma)$:

$\frac{1}{(2\pi)^{\frac{D}{2}} \det(\Sigma)^{\frac{1}{2}}}$, e, and $d(\vec{X};\vec{\mu},\Sigma)^2 = -\frac{1}{2}(\vec{X}-\vec{\mu})^T \Sigma^{-1}(\vec{X}-\vec{\mu})$

1) e = 2.7818281828… Euler's Constant : $\int e^x dx = e^x$. Used to simplify the algebra.

2) The term $(2\pi)^{\frac{D}{2}} \det(\mathbf{\Sigma})^{\frac{1}{2}}$ is a normalization factor.

$$(2\pi)^{\frac{D}{2}} \det(\Sigma)^{\frac{1}{2}} = \int\int\ldots\int e^{-\frac{1}{2}(\vec{X}-\vec{\mu})^T \Sigma^{-1}(\vec{X}-\vec{\mu})} dx_1 dx_2 \ldots dx_D$$

The determinant, det($\mathbf{\Sigma}$) is an operation that gives the volume of $\mathbf{\Sigma}$.

for D=2 $\det\begin{pmatrix} a & b \\ c & d \end{pmatrix} = a\cdot b - c\cdot d$

for D=3 $\det\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} = a\cdot\det\begin{pmatrix} e & f \\ h & i \end{pmatrix} + b\cdot\det\begin{pmatrix} f & d \\ i & g \end{pmatrix} + c\cdot\det\begin{pmatrix} d & e \\ g & h \end{pmatrix}$

$= a(ei-fh) + b(fg-id) + c(dh-eg)$

for D > 3 this continues recursively.

3) The Mahalanobis distance. This is where the action is.

The Mahalanobis Distance

The exponent of $\mathcal{N}(\vec{X};\vec{\mu},\Sigma)$ is known as the "Mahalanobis Distance" named for a famous Indian statistician.

$$d(\vec{X};\vec{\mu},\Sigma)^2 = -\frac{1}{2}(\vec{X}-\vec{\mu})^T \Sigma^{-1}(\vec{X}-\vec{\mu})$$

This is a distance normalized by the covariance.
It is positive and 2nd order (quadratic).

The covariance is provides a distance metric that can be used when the features of $\vec{X}$ have different units (for example with height (m) and weigth (kg). In this case, the features are compared to the standard deviation of a population.

A training set $\{\vec{X}_m\}$ an be used to calculate a vector of average features $\vec{\mu}$

$$\vec{\mu} = E\{\vec{X}\} = \frac{1}{M}\sum_{m=1}^{M}\vec{X} = \begin{pmatrix}\mu_1\\ \mu_2\\ \dots\\ \mu_D\end{pmatrix} = \begin{pmatrix}E\{X_1\}\\ E\{X_2\}\\ \dots\\ E\{X_D\}\end{pmatrix}$$

The vector $\vec{\mu}$ the vector of averages for the components, X_d of $\vec{X}$.
It is also center of gravity (first moment) of the normal density function.

$$\mu_d = E\{X_{d,m}\} = \frac{1}{M}\sum_{m=1}^{M}X_{d,m} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty} p(x_1,x_2,\dots x_D)\cdot x_d\, dx_1,dx_2,\dots,dx_D$$

$$\vec{\mu} = E\{\vec{X}\} = \begin{pmatrix}E\{X_1\}\\ E\{X_2\}\\ \dots\\ E\{X_D\}\end{pmatrix} = \begin{pmatrix}\mu_1\\ \mu_2\\ \dots\\ \mu_D\end{pmatrix}$$

For D dimensions, the second moment is a co-variance matrix composed of D^2 terms:

$$\sigma_{ij}^2 = E\{(X_i-\mu_i)(X_j-\mu_j)\} = \frac{1}{M}\sum_{m=1}^{M}(X_{i,m}-\mu_i)(X_{j,m}-\mu_j)$$

This is often written

$$\Sigma = E\{(\vec{X}-E\{\vec{X}\})(\vec{X}-E\{\vec{X}\})^T\}$$

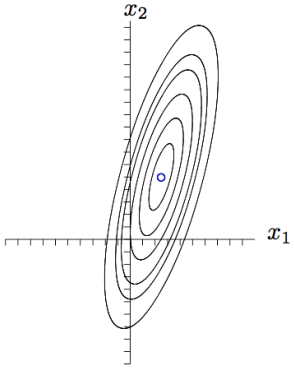
and gives

$$\Sigma = \begin{pmatrix}\sigma_{11}^2 & \sigma_{12}^2 & \dots & \sigma_{1D}^2\\ \sigma_{21}^2 & \sigma_{22}^2 & \dots & \sigma_{2D}^2\\ \dots & \dots & \dots & \dots\\ \sigma_{D1}^2 & \sigma_{D2}^2 & \dots & \sigma_{DD}^2\end{pmatrix}$$

This provides the parameters for

$$p(\vec{X}) = \mathcal{N}(\vec{X};\vec{\mu},\Sigma) = \frac{1}{(2\pi)^{\frac{D}{2}}\det(\Sigma)^{\frac{1}{2}}}e^{-\frac{1}{2}(\vec{X}-\vec{\mu})^T\Sigma^{-1}(\vec{X}-\vec{\mu})}$$

The result can be visualized by looking at the equi-probably contours.



Ellipses for 99%, 95%, 90%, 75%, 50%, and 20% of the mass

If x_i and x_j are statistically independent, then $\sigma_{ij}^2 = 0$

For positive values of σ_{ij}^2, x_i and x_j vary together.
For negative values of σ_{ij}^2, x_i and x_j vary in opposite directions.

For example, consider features x_1 = height *(meters)* and x_2 = weight *(kg)*

In most people height and weight vary together and so σ_{12}^2 would be positive

Linear Transforms of the Normal Multivariate Density

The Normal (Gaussian) function is a defined only by its moments.
It is thus invariant to transformations of its moments, that is affine transformations.
The affine transformations include rotation, translation, scale changes and other linear transformations.

For example consider a cosine vector $\vec{R}$, such that $\|\vec{R}\| = 1$

$$\vec{R} = \begin{pmatrix} \cos(\alpha_1) \\ \cos(\alpha_2) \\ \dots \\ \cos(\alpha_D) \end{pmatrix}$$

A vector $\vec{X}$ may be projected into a space $\vec{Y}$ by

$$\vec{Y} = \vec{R}^T \vec{X}$$

The Normal (Gaussian) function is defined only by its moments.
It is thus invariant to affine transformations of its moments.
The affine transformations include all linear transformations such as rotation, translation, scale changes and sheer.

For a projection onto a 1D vector Y, R is D x 1 : $\vec{Y} = \vec{R}^T \vec{X}$

$$\mu_y = \vec{R}^T \vec{\mu}_x, \qquad \sigma_y^2 = \vec{R}^T \mathbf{\Sigma}_x \vec{R}$$

Note that for the Covariance, projection requires pre- and post- multiplication by $\vec{R}$.
We can demonstrate this with a linear algebraic expression of the moments.

Linear Algebraic Form for Moment Calculation

Calculation of the mean and covariance are often expressed using linear algebra. Such expressions are widely used in machine learning.

As before, assume a set of M training examples $\{X_m\}$

Recall $$\vec{\mu} = E\{\vec{X}\} = \frac{1}{M}\sum_{m=1}^{M}\vec{X}_m = \begin{pmatrix}\mu_1\\ \mu_2\\ \dots\\ \mu_D\end{pmatrix} = \begin{pmatrix}E\{X_1\}\\ E\{X_2\}\\ \dots\\ E\{X_D\}\end{pmatrix}$$

We can compose a matrix with M columns and D rows from $\{X_m\}$.

$$X = \begin{pmatrix}x_{11} & x_{12} & \dots & x_{1M}\\ x_{21} & x_{22} & \dots & x_{2M}\\ \dots & \dots & & \dots\\ x_{D1} & x_{D2} & \dots & x_{DM}\end{pmatrix}$$ Let us define the unit vector : $\vec{u} = \begin{pmatrix}1\\ 1\\ \vdots\\ 1\end{pmatrix}$

Then $$\vec{\mu} = \begin{pmatrix}\mu_1\\ \mu_2\\ \dots\\ \mu_D\end{pmatrix} = \begin{pmatrix}x_{11} & x_{12} & \dots & x_{1N}\\ x_{21} & x_{22} & \dots & x_{2N}\\ \dots & \dots & & \dots\\ x_{D1} & x_{D2} & \dots & x_{DN}\end{pmatrix}\begin{pmatrix}1\\ 1\\ \vdots\\ 1\end{pmatrix} = X\cdot\vec{u}$$

Let us define $\vec{V}_m = \vec{X}_m - E\{\vec{X}_m\} = \vec{X}_m - \vec{\mu}$.

We can compose a matrix with N columns and D rows from $\{V_n\}$.

$$V = \begin{pmatrix}v_{11} & v_{12} & \dots & v_{1M}\\ v_{21} & v_{22} & \dots & v_{2M}\\ \dots & \dots & & \dots\\ v_{D1} & v_{D2} & \dots & v_{DM}\end{pmatrix}$$

From this: $\Sigma_x = E\{\vec{V}\vec{V}^T\}$ can be computed as a vector product.

$\Sigma_x \equiv V\ V^T$ is a D x D matrix that captures the "co-variance" of the elements of i,j of the vector X in $\{X_n\}$

This can be seen as

$$\Sigma_X = \mathbf{V}\mathbf{V}^T = \begin{pmatrix} \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{pmatrix} \begin{pmatrix} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix} = \begin{pmatrix} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix}$$

Note that we can also write $\Sigma_N = V^T V$ of size N x N.

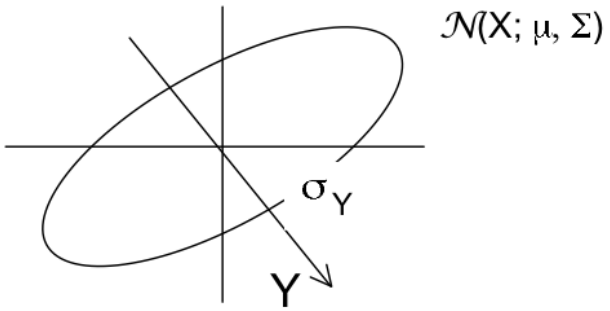
We can use this to show that projection of a covariance requires pre and post multiplication:

Note que $(\vec{R}^T V)^T = (V^T \vec{R})$

Thus
$$\Sigma_y = (\vec{R}^T V)(\vec{R}^T V)^T$$
$$\Sigma_y = (\vec{R}^T V)(V^T \vec{R})$$
$$\Sigma_y = \vec{R}^T (VV^T)\vec{R}$$
$$\Sigma_y = \vec{R}^T \Sigma_X \vec{R}$$

Thus projection of a covariance requires pre and post multiplication by $\vec{R}$.
In the case of projection to a 1D vector Y:

$$\sigma_y^2 = \vec{R}^T \Sigma_X \vec{R}$$



Quadratic Discrimination

The classification function can be decomposed into two parts: d() and $g_k()$:

$$\hat{\omega}_k = d\left(g_k\left(\vec{X}\right)\right)$$

$g(\vec{X})$: A discriminant function : $R^D \rightarrow R^K$
d() : a decision function $R^K \rightarrow \{\omega_K\}$

The discriminant is a vector of functions:

$$\vec{g}(\vec{X}) = \begin{pmatrix} g_1(\vec{X}) \\ g_2(\vec{X}) \\ \vdots \\ g_K(\vec{X}) \end{pmatrix}$$

Quadratic discrimination functions can be derived directly from $p(\omega_k \mid \vec{X})$

$$P(\omega_k \mid \vec{X}) = \frac{p(\vec{X} \mid \omega_k)P(\omega_k)}{p(\vec{X})}$$

To minimize the number of errors, we will choose k such that

$$\hat{\omega}_k = \underset{\omega_k}{\arg\max}\{\frac{p(\vec{X} \mid \omega_k)P(\omega_k)}{p(\vec{X})}\}$$

but because *p(X)* is constant for all k, it is common to use a likelihood function:

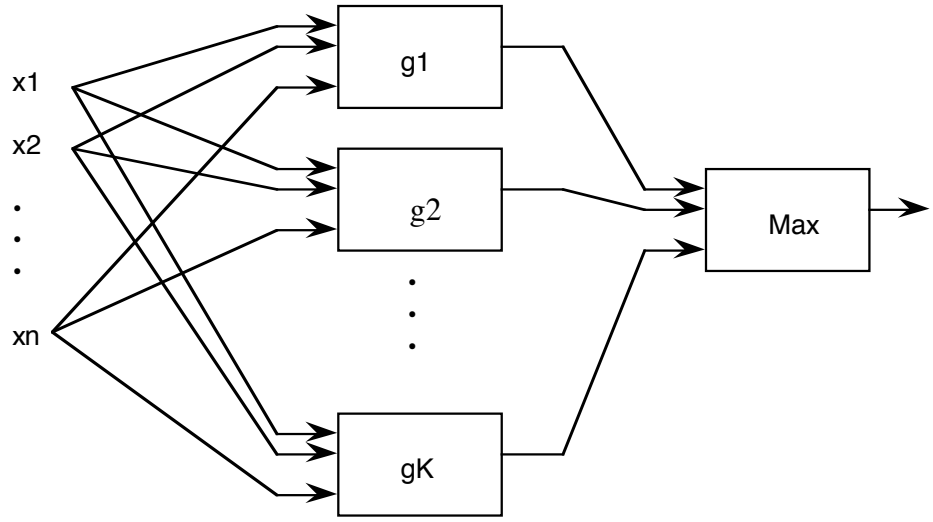
$$\hat{\omega}_k = \underset{\omega_k}{\arg\max}\{P(\vec{X} \mid \omega_k)p(\omega_k)\}$$

This is called a "Maximum Likelihood" classifier.

Warning: Maximum likelihood can sometimes give a highly improbable answer. Remember that confidence in the choice $\hat{\omega}_k$ is provided by the full probability:

$$CF_{\hat{\omega}_k} = p(\hat{\omega}_k \mid \vec{X}) = \frac{P(\vec{X} \mid \hat{\omega}_k)p(\hat{\omega}_k)}{P(\vec{X})}$$

The maximum likelihood classifier can be organized as a set of parallel discriminant functions. This gives a sort of parallel machine that resembles:



The functions $g_k()$ are commonly constructed by from the Log of the likelihood:

$$g_k(X) = Log\{p(\vec{X} \mid \omega_k)P(\omega_k)\}$$

This gives a simple function in the case where the pdf is a multivariate norm:

$$P(\vec{X} \mid \omega_k) = \mathcal{N}(\vec{X}; \vec{\mu}_k, \Sigma_k)$$

or even to a Gaussian Mixture Model

$$P(\vec{X} \mid \omega_k) = \sum_{n=1}^{M} \alpha_n \mathcal{N}(\vec{X}; \vec{\mu}_n, \Sigma_n)$$

Discrimination using Log Likelihood

As a simple example, let D=1

$$P(X) = \mathcal{N}(X;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(X-\mu)^2}{2\sigma^2}}$$

The Likelihood function takes the form:

$$L_k(X) = p(X \mid \omega_k) \cdot P(\omega_k) = P(\omega_k) \cdot \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{(X-\mu_k)^2}{2\sigma_k^2}}$$

We can simplify the math by working with the Logarithm of the likelihood.
We note that:

$$\hat{k} = \arg\max_k \{L_k(X)\} = \arg\max_k \{Log\{L_k(X)\}\}$$

because Log{} is a monotonic function. In this case we define $g_k()$ to be the Log-likelihood.

$$g_k(X) = Log\{\mathcal{N}(X;\mu_k,\sigma_k) \cdot P(\omega_k)\}$$

$$g_k(X) = Log\{\frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{(X-\mu_k)^2}{2\sigma_k^2}}\} + Log\{P(\omega_k)\}$$

$$g_k(X) = Log\{\frac{1}{\sqrt{2\pi}\sigma_k}\} + Log\{e^{-\frac{(X-\mu_k)^2}{2\sigma_k^2}}\} + Log\{P(\omega_k)\}$$

$$g_k(X) = -Log\{\sqrt{2\pi}\} - Log\{\sigma_k\} - \frac{(X-\mu_k)^2}{2\sigma_k^2} + Log\{P(\omega_k)\}$$

$$g_k(X) = -Log\{\sigma_k\} - \frac{(X-\mu_k)^2}{2\sigma_k^2} + Log\{P(\omega_k)\}$$

and

$$\hat{k} = \arg\max_k \{g_k(X)\}$$

Example for K > 2 and D > 1

In general, it is more effective to work with D>1 features.
In this case:

$$g_k(\vec{X}) = Log\{p(\omega_k \mid \vec{X})P(\omega_k)\}$$

Thus the classifier is a machine that calculates K functions $g_k(\vec{X})$
Followed by a maximum selection.

The discrimination function is $g_k(\vec{X}) = Log\{p(\omega_k \mid \vec{X})P(\omega_k)\}$

For a Gaussian (Normal) density function

$$p(\vec{X} \mid \omega_k) = \mathcal{N}(\vec{X}; \vec{\mu}_k, \Sigma_k)$$

$$Log(p(\vec{X} \mid \omega_k)) = Log(\frac{1}{(2\pi)^{\frac{D}{2}} \det(\Sigma_k)^{\frac{1}{2}}} e^{-\frac{1}{2}(\vec{X}-\vec{\mu}_k)^T \Sigma_k^{-1}(\vec{X}-\vec{\mu}_k)})$$

$$Log(p(\vec{X} \mid \omega_k)) = -\frac{D}{2}Log(2\pi) - \frac{1}{2}Log\{Det(\Sigma_k)\} - \frac{1}{2}(\vec{X}-\vec{\mu}_k)^T \Sigma_k^{-1}(\vec{X}-\vec{\mu}_k)$$

We can observe that $-\frac{D}{2}Log(2\pi)$ can be ignored because it is constant for all k.

The discrimination function becomes:

$$g_k(\vec{X}) = -\frac{1}{2}Log\{\det(\Sigma_k)\} - \frac{1}{2}(\vec{X}-\vec{\mu}_k)^T \Sigma_k^{-1}(\vec{X}-\vec{\mu}_k) + Log\{p(\omega_k)\}$$

Different families of Bayesian classifiers can be defined by variations of this formula.
This becomes more evident if we reduce the equation to a quadratic polynomial.

Canonical Form for the discrimination function

The quadratic discriminant can be reduced to a standard (canonical) form.

$$g_k(\vec{X}) = -\frac{1}{2} Log\{\det(\Sigma_k)\} - \frac{1}{2}(\vec{X} - \vec{\mu}_k)^T \Sigma_k^{-1}(\vec{X} - \vec{\mu}_k) + Log\{p(\omega_k)\}$$

Let us start with the term $(\vec{X} - \vec{\mu}_k)^T \Sigma_k^{-1}(\vec{X} - \vec{\mu}_k)$.

This can be rewritten as :

$$(\vec{X} - \vec{\mu}_k)^T \Sigma_k^{-1}(\vec{X} - \vec{\mu}_k) = \vec{X}^T \Sigma_k^{-1}\vec{X} - \vec{X}^T \Sigma_k^{-1}\vec{\mu}_k - \vec{\mu}_k{}^T \Sigma_k^{-1}\vec{X} + \vec{\mu}_k{}^T \Sigma_k^{-1}\vec{\mu}_k$$

We note that $\vec{X}^T \Sigma_k^{-1}\vec{\mu}_k = \vec{\mu}_k{}^T \Sigma_k^{-1}\vec{X}$
and thus : $-\vec{X}^T \Sigma_k^{-1}\vec{\mu}_k - \vec{\mu}_k{}^T \Sigma_k^{-1}\vec{X} = -2\vec{\mu}_k{}^T \Sigma_k^{-1}\vec{X}$

we define: $\vec{W}_k = \frac{1}{2}\left(-2\vec{\mu}_k{}^T \Sigma_k^{-1}\right) = -\vec{\mu}_k{}^T \Sigma_k^{-1}$
Let us also define $D_k = -\frac{1}{2}\Sigma_k^{-1}$

The remaining terms are constant. Let us defined the constant

$$b_k = -\frac{1}{2}\vec{\mu}_k{}^T \Sigma_k^{-1}\vec{\mu}_k - \frac{1}{2} Log\{\det(\Sigma_k)\} + Log\{p(\omega_k)\}$$

which gives a quadratic polynomial

$$g_k(\vec{X}) = \vec{X}^T D_k \vec{X} + \vec{W}_k{}^T \vec{X} + b_k$$

where: $D_k = -\frac{1}{2} C_k^{-1}$

$\vec{W}_k = -\vec{\mu}_k{}^T \Sigma_k^{-1}$

and $b_k = -\frac{1}{2}\vec{\mu}_k{}^T \Sigma_k^{-1}\vec{\mu}_k - \frac{1}{2} Log\{\det(\Sigma_k)\} + Log\{p(\omega_k)\}$

A set of K discrimination functions $g_k(\vec{X})$ partitions the space $\vec{X}$ into a disjoint set of regions with quadratic boundaries. The boundaries are the functions $g_i(\vec{X}) - g_j(\vec{X}) = 0$

The boundaries are defined by points for which

$$g_i(\vec{X}) = g_j(\vec{X}) \geq \; g_k(\vec{X}) \; \forall k \neq i, j$$