Gaussian Naive Bayesian

(Taken mostly from the book by Duda, Hart, and Stork.)

Gaussian density/distribution

The Gaussian density function of n-dimensional vectors is:

$$g(x;\mu,C) = \frac{1}{(\sqrt{2\pi})^n |C|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T C^{-1}(x-\mu)}$$

Here μ is the distribution mean and C is the Covariance matrix. The value $|C|$ is the determinant of the matrix C. The parameters μ, C can be estimated from the data by:

$$\mu = \frac{\sum_{i=1}^m x_i}{m}, \quad C = \frac{\sum_i (x_i-\mu)(x_i-\mu)^T}{m} \quad \text{or} \quad C = \frac{\sum_i (x_i-\mu)(x_i-\mu)^T}{m-1}$$

$$\text{The special } n\text{=1 case is:} \qquad g(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

The parameters μ, σ can be estimated from the data by:

$$\mu = \frac{\sum_{i=1}^m x_i}{m}, \quad C = \frac{\sum_i (x_i-\mu)^2}{m} \quad \text{or} \quad C = \frac{\sum_i (x_i-\mu)^2}{m-1}, \quad \sigma = \sqrt{C}$$

Discriminant functions for Gaussian density

Suppose the positive examples have Gaussian density with parameters μ_1, C_1, and the negative examples have Gaussian density with parameters μ_2, C_2. When should we decide that x is positive and not negative? Call h_1 the hypothesis that x is positive, and h_2 the hypothesis that x is negative. We should decide that x is positive if: $P(h_1|x) > P(h_2|x)$. Observe that $P(x|h_1) = g(x;\mu_1,C_1)$, and similarly $P(x|h_2) = g(x;\mu_2,C_2)$. Therefore:

ML classifier: Classify as positive if	$g(x;\mu_1,C_1) > g(x;\mu_2,C_2)$
MAP classifier: Classify as positive if	$g(x;\mu_1,C_1)P(h_1) > g(x;\mu_2,C_2)P(h_2)$

Concentrating on the MAP classifier the condition $g(x;\mu_1,C_1)P(h_1) > g(x;\mu_2,C_2)P(h_2)$ gives the following after taking logarithm:

$$\begin{aligned} &-\frac{1}{2}(x-\mu_1)^T C_1^{-1}(x-\mu_1) - \frac{n\ln 2\pi}{2} - \frac{\ln|C_1|}{2} + \ln P(h_1) \\ &> -\frac{1}{2}(x-\mu_2)^T C_2^{-1}(x-\mu_2) - \frac{n\ln 2\pi}{2} - \frac{\ln|C_2|}{2} + \ln P(h_2) \end{aligned}$$

This simplifies to:

$$(x-\mu_2)^T C_2^{-1}(x-\mu_2) - (x-\mu_1)^T C_1^{-1}(x-\mu_1) + \ln|C_2| - \ln|C_1| + 2(\ln P(h_1) - \ln P(h_2)) > 0$$

This is a quadratic discriminant function. The main problem with this approach is that the Covariance matrices have $O(n^2)$ variables that need to be determined, and reliable estimation may require a lot of training data. Instead, in many situations this model is used with additional simplifying assumptions that reduce the quadratic expression to a linear expression.

Case 1: $C_1 = C_2 = \sigma^2 I$

Here the discriminant function is:

$$|x - \mu_2|^2 - |x - \mu_1|^2 + 2\sigma^2(\ln P(h_1) - \ln P(h_2)) > 0$$

This can be further simplified to:

$$\begin{aligned} d(x) =& w^T x + b > 0 \quad \text{where:} \\ & w = (\mu_1 - \mu_2) \\ & b = \frac{1}{2}(|\mu_2|^2 - |\mu_1|^2) + \sigma^2(\ln P(h_1) - \ln P(h_2)) \end{aligned}$$

Typically only the value of w is computed. The values of $w^T x_i$ are calculated for all i, and the value of b is determined using an ad-hoc technique for computing a threshold.

Case 2: $C_1 = C_2 = C$

In this case it can be shown that the discriminant function is:

$$\begin{aligned} d(x) =& w^T x + b > 0 \quad \text{where:} \\ & w = C^{-1}(\mu_1 - \mu_2) \\ & b = \frac{1}{2}(\mu_2^T C^{-1} \mu_2 - \mu_1^T C^{-1} \mu_1) + (\ln P(h_1) - \ln P(h_2)) \end{aligned}$$

Typically only the value of w is computed. The values of $w^T x_i$ are calculated for all i, and the value of b is determined using an ad-hoc technique for computing a threshold.

Case 3: arbitrary C_1**,** C_2

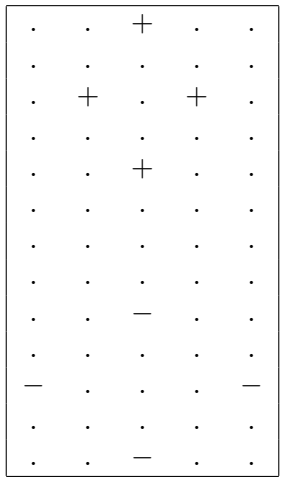
Here we can get an arbitrary quadratic function as the discriminant.

$$\begin{aligned} d(x) &= (x - \mu_2)^T C_2^{-1}(x - \mu_2) - (x - \mu_1)^T C_1^{-1}(x - \mu_1) + \ln|C_2| - \ln|C_1| + 2(\ln P(h_1) - \ln P(h_2)) \\ &= (x - \mu_2)^T C_2^{-1}(x - \mu_2) - (x - \mu_1)^T C_1^{-1}(x - \mu_1) + b > 0 \end{aligned}$$

Typically the value of b is determined using an ad-hoc technique for computing a threshold.

Example

x_1	x_2	y
2	6	$+$
3	8	$+$
4	6	$+$
3	4	$+$
1	-2	$-$
3	0	$-$
5	-2	$-$
3	-4	-



Case 1:

$$\mu_1 = \begin{pmatrix} 3 \\ 6 \end{pmatrix}, \quad \mu_2 = \begin{pmatrix} 3 \\ -2 \end{pmatrix}$$

$$w = \begin{pmatrix} 0 \\ 8 \end{pmatrix}, \quad b = -16$$

$$d(x) = 8x_2 - 16, \quad \text{or} \quad d(x) = x_2 - 2$$

The value of b was computed by considering the sorted values of $w^T x$. They are:

-4 · 8	-2 · 8	-2 · 8	0 · 8	4 · 8	6 · 8	6 · 8	8 · 8
$-$	$-$	$-$	$-$	$+$	$+$	$+$	$+$

The value of b that gives the smallest error and maximizes the margins.

Case 2:

$$\mu = \begin{pmatrix} 3 \\ 2 \end{pmatrix}$$

$$C = \frac{1}{8}\begin{pmatrix} 10 & 0 \\ 0 & 144 \end{pmatrix} = \begin{pmatrix} 10/8 & 0 \\ 0 & 18 \end{pmatrix}$$

To compute w solve the following linear system:

$$Cw = \mu_1 - \mu_2$$

This gives:

$$w = \begin{pmatrix} 0 \\ 4/9 \end{pmatrix}$$

$$d(x) = \frac{4}{9}x_2 + b \quad \text{or} \quad d(x) = x_2 + b = x_2 - 2$$

The value of b was computed in the same way as in Case 1. The result is the same as in Case 1: $b = -2$.

Case 3:

$$C_1 = \begin{pmatrix} 1/2 & 0 \\ 0 & 2 \end{pmatrix}, \quad C_1^{-1} = \begin{pmatrix} 2 & 0 \\ 0 & 1/2 \end{pmatrix}$$
$$C_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}, \quad C_2^{-1} = \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix}$$

We have:

$$(x-\mu_1)^T C_1^{-1}(x-\mu_1) = 2(x_1-3)^2 + \frac{1}{2}(x_2-6)^2$$
$$(x-\mu_2)^T C_2^{-1}(x-\mu_2) = \frac{1}{2}(x_1-3)^2 + \frac{1}{2}(x_2+2)^2$$

This gives the following expression for the discriminant:

$$d(x) = \frac{1}{2}(x_1-3)^2 + \frac{1}{2}(x_2+2)^2 - 2(x_1-3)^2 - \frac{1}{2}(x_2-6)^2 + b = d_q(x) + b$$

Simplifying $d_q(x)$ we get:

$$d_q(x_1, x_2) = -1.5x_1^2 + 9x_1 + 8x_2 - 29.5$$

To compute b we consider the sorted values of $d_q(x)$:

$d_q(3,-4)$ $= -48$	$d_q(5,-2)$ $= -38$	$d_q(1,-2)$ $= -38$	$d_q(3,0)$ $= -16$	$d_q(3,4)$ $= 16$	$d_q(4,6)$ $= 30.5$	$d_q(2,6)$ $= 30.5$	$d_q(3,8)$ $= 48$
$-$	$-$	$-$	$-$	$+$	$+$	$+$	$+$

The value of b that gives the smallest error and maximizes the margins is $b = 0$. Observe that this result is different from cases 1,2. For example, in Case 3 we classify an example as negative whenever x_1 is large positive or large negative.