Tutorial on Bayesian Decision Theory with Gaussians

Introduction

Bayesian Decision Theory is one method used to solve *Pattern Recognition* problems, when those problems are posed in a particular way.

Suppose that an observer is standing over a conveyer belt that leads into a grocery store. The belt carries 2 types of fruits - oranges and apples. It is up to the observer to determine which of the 2 fruits is on the belt at a particular moment. For humans and for machines, this is done by examining certain features of the fruits and then classifying them as either an orange or an apple based on those features. This is exactly a pattern classification problem. If the fruits enter the store at random, and we have no other information, then the only way to classify a certain fruit would be by guessing.

Bayesian decision theory plays a role when there is some *a priori* information about the things we are trying to classify.

For example, suppose that you didn't know anything about the fruits, but you knew that 80% of the fruits that the conveyer belt carried were apples, and the rest were oranges. If this is the only information that you are given in order to make your decision, then you would want to classify a random fruit as an apple. The a priori information in this case is the probabilities of either an apple or an orange being on the conveyer belt. If a decision must be made with so little information, it makes sense to use the following rule:

Decide 'apple' if P(apple) > P(orange), otherwise, decide 'orange'.

where P(apple) is the probability of their being an apple on the belt. In this case, P(apple) = 0.8 (80%). This may seem strange, because if you use this rule, then you will classify every random fruit as an apple. But by doing this, you still ensure that you will be right about 80% of the time.

The above example is very simple, and should just be used to understand the basic idea of a pattern recognition problem involving some probability information. In general, there is a lot more known about the things we are trying to classify. For example, we may know that most apples are red, and therefore if we observe a red fruit, we should classify it as an apple. In this case, we could then use the color of the fruit to determine what it is. We would usually also have the probability distribution of the color property for apples and oranges. This means we would know exactly how rare it is for there an orange to have the same color as your typical apple. Obviously this is important, because if oranges were more or less the same color as apples then this feature would not be very useful to a pattern classifier. These probability distributions play an important role in formalizing the decision rule. (For background information, see Stats glossary).

The next section will describe in greater detail the importance of the a priori information in the pattern recognition problem, and will provide some exact formulas for decision rules that use probability distributions.

Decision Rules

Notation:

As was discussed in the introduction, in most circumstances there is some information about the objects we are trying to classify.

For example, we may have the probability distribution for the color of apples, as well as that for oranges. To introduce some notation, let w_{app} represent the state of nature where the fruit is an apple, let w_{org} represent that state where the fruit is an orange, and let x be the continuous random variable that represents the color of a fruit. Then the expression $p(x|w_{app})$ represents the density function for x given that the state of nature is an apple.

In a typical problem, we would know (or be able to calculate) the conditional densities $p(x|w_j)$ for j either an apple or an orange. We would also typically know the prior probabilities $P(w_{app})$ and $P(w_{org})$, which represent simply the total number of apples versus oranges that are on the conveyer belt. What we are looking for is some formula that will tell us about the probability of a fruit being an apple or an orange *given* that we observe a certain color x. If we had such a probability, then for some given color that we observed we would classify the fruit by comparing the probability that an orange had such a color versus the probability that an apple had such a color. If it were more probable that an apple had such a color, the fruit would be classified as an apple. Fortunately, we can use Baye's Formula which states that :

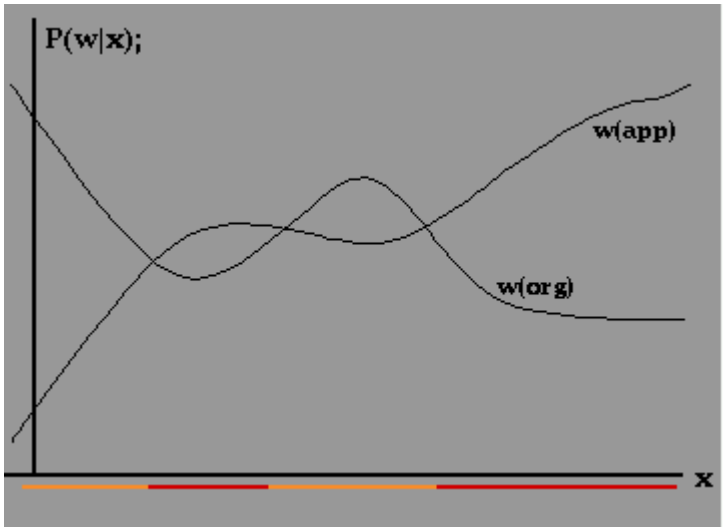
$$\mathbf{P(w_j|x) = p(x|w_j)\ P(w_j)/p(x)}$$

What this is means, is that using our a priori information, we can calculate the *a posteriori* probability of the state of nature being in state w_j given that the feature value x has been measured. So, if you observe a certain x for a random fruit on the conveyer belt, then by calculating $P(w_{app}|x)$ and $P(w_{org}|x)$ we would be inclined to decide that the fruit was an apple if the first value were greater than the second one. Similiarly, if $P(w_{org}|x)$ is greater, we would decide that the fruit was most likely an orange. Therefore, *Baye's Decision Rule* can be stated as:

Decide w_{org} if $P(w_{org}|x) > P(w_{app}|x)$; otherwise decide w_{app},

Since p(x) occurs on both sides of the comparison, the rule is equivalent to saying:

Decide w_{org} if $p(x|w_{org})P(w_{org}) > p(x|w_{app})P(w_{app})$;
otherwise decide w_{app}.



The following graph shows the a posteriori probabilities for the 2 class decision problem. At every x, the posteriors must sum to 1. The red region on the x axes depicts values for x for which the decision rule would decide 'apple'. The orange region represents values for x for which you would decide 'orange'.

The probability that we make an error is just the minimum of the 2 curves at any point, since that represents the smaller probability that we didn't pick. So

$$\mathbf{P(}error\mathbf{|x) = min[p(}w_{app}\mathbf{|x), p(}w_{org}\mathbf{|x)].}$$

This formula represents the probability of making an error for a specific measurement x. But it is often useful to know the *average* probability of error over all possible measurements. This can be calculated using Bayes' **Law of total Probabilities**, which implies that

$$P(\text{error}) = \int_{all\,x} [p(x)min[p(w_{app}|x), p(w_{org}|x)]$$

Allowing more than 1 feature and more than 1 class:

In a more general case, there are several different features that we measure, so instead of x we have a *feature vector* **x** in R^d for d different features. We also allow for more than 2 possible states of nature, where w_1 ... w_c represent the c states of nature. Bayes' formula can be computed in the same way as:

$$P(w_j|\mathbf{x}) = p(x|w_j)P(w_j)/\ p(\mathbf{x}), \text{ for } j=1..c$$

but now p(**x**) can be calculated using the Law of Total Probabilities so that

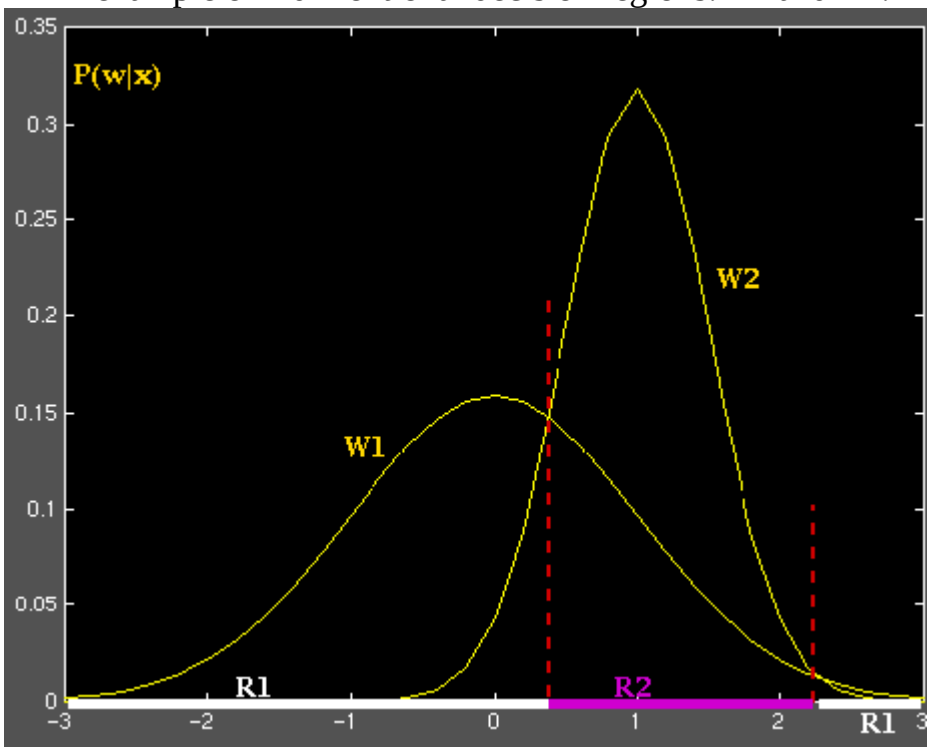
$$p(\mathbf{x}) = \sum_{j=1}^{c} p(\mathbf{x}|w_j)P(w_j)$$

As before, if we measure feature vector **x**, we want to classify an object into class j if $p(w_j|\mathbf{x})$ is the maximum of all the probability densities for j=1..c. This is the same as the Bayes' decision rule for the 2 category case.

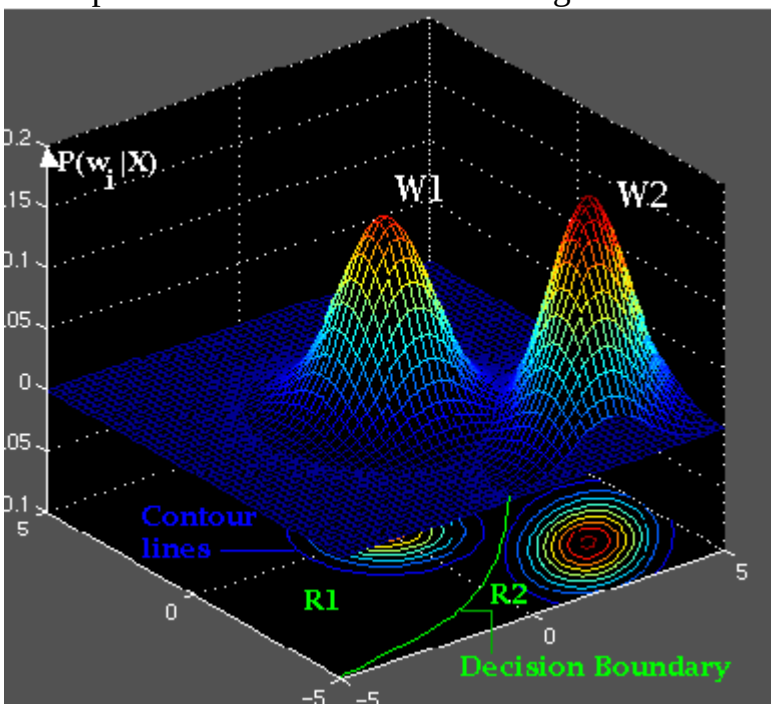
Decision Regions

When any decision rule is applied to the d-dimensional feature space R^d, the result is that the space is split up into c decision regions R_1, ..., R_c. In the above graph for the 2 category case, the decision regions were marked in red and orange at the bottom of the graph. In general, if x lies in decision region R_i then it means that the pattern classifier selected the function $g_i(x)$ to be the maximum of all the discriminant functions. The decision regions are any subset of the space R^d. For example, if the feature vector is a 2-dimentional vector, then the discriminant functions $g_i(\mathbf{x})$ will be functions of 2 variables and will be mapped in 3-D. The decision regions for this case will be subsets of the x-y plane. Here are 2 simple examples:

An example of 1 dimentional decision regions: R1 and R2.



An example of 2 dimensional decision regions: R1 and R2.



In general,decision regions may be any subsets of R^d and it is common to have a region R_i that is disconnected.

Obviously, the shape of the decision bondary depends on the functions $P(w_i|\mathbf{x})$. The next section takes a closer look at discriminant functions and their corresponding decision regions for the Normal Density in particular.

Decision Rules for the Normal Distribution

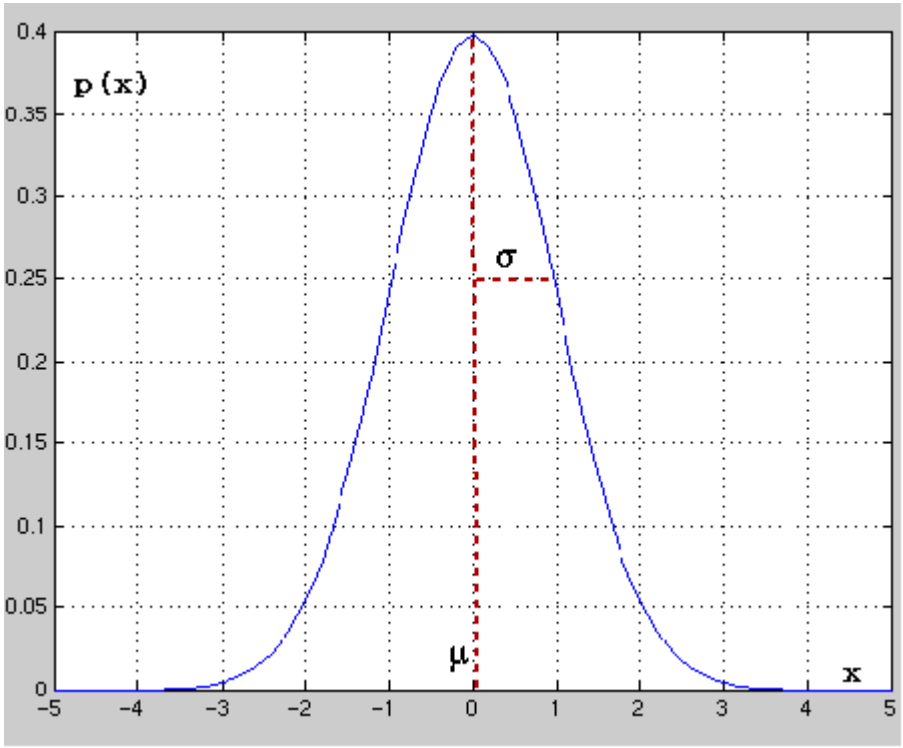
Definitions

The multivariate normal density is typically an appropriate model for most pattern recognition problems where the feature vectors **x** for a given class w_i are continuous valued, mildly corrupted versions of a single mean vector $\boldsymbol{u_i}$. In this case, the conditional densities $p(x|w_i)$ and the a priori probabilities $P(w_i)$ are normally distributed. (For background information on the normal density, see Normal Distribution). As a reminder, the density function for the univariate normal is given by

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)}$$

$$\text{for } \sigma > 0 \text{ and } -\text{inf} < \mu < \text{inf}$$

The 2 parameters called the *mean* and the *variance* completely specify the normal distribution. Samples from this type of distribution tend to cluster about the mean, and the extend to which they spread out depends on the variance.



The general multivariate normal density is given by a d-dimensional *mean vector* and a d-by-d *covariance matrix*:

$$\mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_d \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1d} \\ \vdots & & \vdots \\ \sigma_{d1} & \cdots & \sigma_{dd} \end{bmatrix}$$

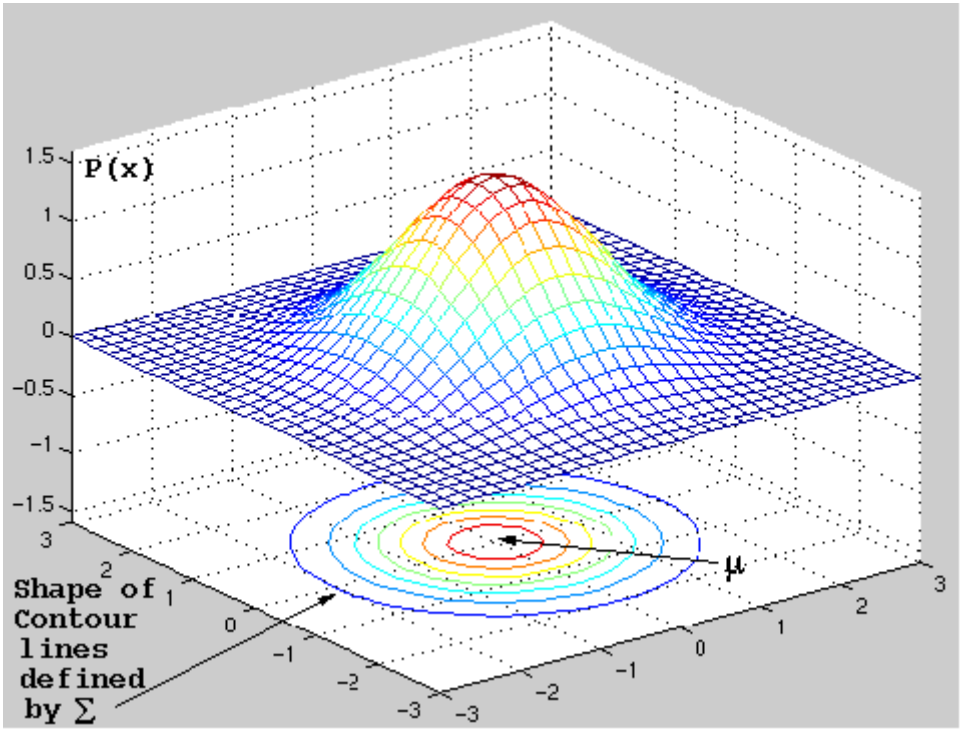
$$p(x) = \frac{1}{\sqrt{2\Pi}\sqrt{\Sigma}} \exp\left[\frac{-1}{2}(x-\mu)^t \Sigma^{-1}(x-\mu)\right]$$

The *mean vector* is just a collection of single means u_i where the ith mean represents the mean for the ith feature that we are measuring. For example, if

we decided to measure the color and weight of a random fruit, then u_1 = the mean of all the colors and u_2 = the mean of all the weights.

The covariance matrix is similar to the variance in the univariate case. The diagonal elements represent the variances for the different features we measure. For example, the ith diagonal element represents the variance for the ith feature we measure. The off-diagonal elements represent the covariance between 2 different features. In other words, the element o_{ij} in the above matrix represents the covariance between feature i and feature j. This is important because the features that we measure are not necessarily independent. Suppose that the color of some fruit depended on the weight of the fruit. The exact value of the covariance for color and weight would depend on exactly *how* they vary together. For more information on these values, see Covariance.

As with the univariate density, samples from a normal population tend to fall in a single cluster centered about the mean vector, and the shape of the cluster depends on the covariance matrix:



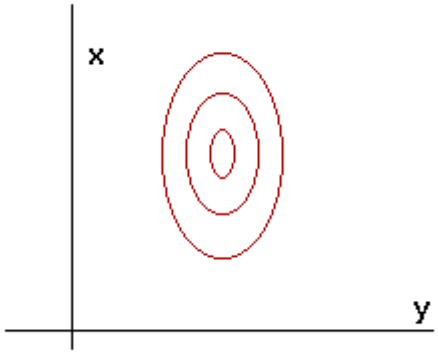
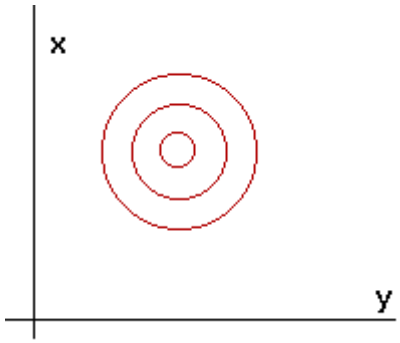
The contour lines in the above diagram show the regions for which the function has constant density. From the equation for the normal density, it is apparent that points which have the same density must have the same constant term:

$$(x - \mu)^t \Sigma^{-1} (x - \mu)$$

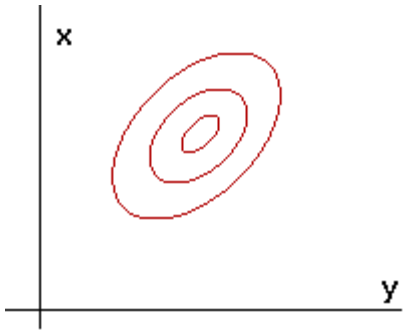
This quantity is often called the squared Mahalanobis distance from **x** to ***u***. This term depends on the contents of the covariance matrix, which explains

why the shape of the contour lines (lines of constant Mahalanobis distance) is determined by this matrix. Since this distance is a quadratic function, the contours of constant density are hyperellipsoids of constant Mahalanobis distance to ***u***.

In simple cases, there is some intuition behind the shape of the contours, depending on the contents of the covariance matrix:

Description	Diagram of the contour lines on the xy plane.
The covariance matrix for 2 features x and y is diagonal (which implies that the 2 features don't covary), but feature x varies more than feature y. The contour lines are stretched out in the x direction to reflect the fact that the distance spreads out at a lower rate in the x direction than it does in the y direction. The reason that the distance decreases slower in the x direction is because the variance for x is greater and thus a point that is far away in the x direction is not quite as *distant* from the mean as a point that is far away in the y direction.	
The covariance matrix for 2 features x and y is diagonal, *and* x and y have the exact same variance. This results in euclidean distance contour lines.	

The covariance matrix is *not* diagonal. Instead, x and y have the same variance, but x varies with y in the sense that x and y tend to increase together. So the covariance matrix would have identical diagonal elements, but the off-diagonal element would be a strictly positive number representing the covariance of x and y.



Discriminant Functions for the Normal Density

One of the discriminant functions that was listed in the previous section on decision rules, was

$$g_i(x) = \ln p(x|w_i) + \ln P(w_i)$$

. When the densities $p(x|w_i)$ are each normally distributed then the discriminant function becomes :

$$g_i(x) = \frac{-1}{2}(x - \mu_i)^t \Sigma_i^{-1}(x - \mu_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma| + \ln P(w_i) \qquad (0)$$

where u_i is the mean vector for the distribution of class i, and Σ_i is the covariance matrix for the distribution of class i.

Here let's check a special case where the above discriminant function can be significantly simplified.

Suppose that each of the features that we are measuring are independent of each other. For example, if we were once again trying to recognize an apple from an orange, and we measured the color and the weight as our feature vector, then chances are that there is no relationship between these 2 properties. The non-diagonal elements of the covariance matrix are the covariances of the 2 features x_1 = color and x_2 = weight. But because these features are independent, their covariances would be 0. Therefore, the covariance matrix for both classes would be diagonal.

As a second simplification, assume that the variance of colors is the same is the variance of weights. This means that there is the same degree of spreading out from the mean of colors as there is from the mean of weights. If this is true for some class i then the covariance matrix for that class will have identical diagonal elements.
Finally, suppose that the variance for the color and weight features is the same in both classes. This means that the degree of spreading for these 2 features is independent of the class from which you draw your samples. If this is true, then the covariance matrices Σ_i for i=1,2 be identical. When normal distributions are plotted that have a diagonal covariance matrix that is just a constant multplied by the identity matrix, their cluster points about the mean are spherical in shape, (see Covariance matrix examples).

To calculate the functions $g_i(x)$ now becomes very easy. Suppose that the common covariance matrix for all the classes is given by $\Sigma = \sigma I$ where I is the identity matrix. Then the determinant is just σ^{2d} and the inverse of Σ_i is $\frac{1}{\sigma^2} I$. So the discriminant functions now become:

$$g_i(x) = \frac{-1}{2}(x - \mu_i)^t \frac{1}{\sigma^2} I (x - \mu_i) - \frac{d}{2} \ln 2\pi - \frac{1}{2} \ln \sigma^{2d} + \ln P(w_i).$$

The above 2 terms have been crossed out because they are additive constants that are the same for each g_i. So the final version of the discriminant functions can be given by :

$$g_i(x) = \frac{-1}{2\sigma^2}(x - \mu_i)^t (x - \mu_i) + \ln P(w_i).$$

This discriminant function does make some intuitive sense. The first term is just the *Euclidean norm* that has been normalized by dividing by the variance . The second term is just the a priori probability of class i. To understand how this discriminant function works, suppose that some x is equally near 2 different mean vectors. Then the first term for both the discriminant function for apples and that for oranges, will be the same. Thus the decision rule will rely on the second term alone and will favor the class that has the greater a priori probability. As an example, imagine that you observe a fruit that has the color and weight properties that lie exactly between the average color and weight for oranges and that for apples. Then these measured features don't help in your decision making. But if also have the information that 80% of all the fruits are apples, then it would make sense to classify that fruit as an apple.

If you expand out the euclidean distance term, you have a discriminant function that looks like:

$$g_i(x) = \frac{-1}{2\sigma^2}[x^t x - 2\mu_i^t x + \mu_i^t \mu_i] + \ln P(w_i).$$

The first quadratic term of x was crossed out because it is the same for every $g_i(x)$. So the discriminant functions are actually *linear* and are of the form

$$g_i(x) = w_i^t x + w_{i0}$$

where

$$w_i = \frac{1}{\sigma^2}\mu_i \qquad (1)$$

and

$$w_{i0} = \frac{-1}{2\sigma^2}\mu_i^t \mu_i + \ln P(w_i).$$

The decision boundaries for these discriminant functions are found by intersecting the functions $g_i(x)$ and $g_j(x)$ where i and j represent the 2 classes with the highest a posteriori probabilities. As in the univariate case, this is equivalent to determining the region for which $g_i(x)$ is the maximum of all the discriminant functions. By setting $g_i(x) = g_j(x)$ we have that:

$$(w_i^t - w_j^t)x + (w_{i0} - w_{j0}) = 0 \qquad (2)$$

Consider the term $w_{i0} - w_{j0}$:

$$= \frac{-1}{2\sigma^2}\mathbf{u}_i^t\mathbf{u}_i + \ln \mathbf{P}(\mathbf{w}_i) + \frac{1}{2\sigma^2}\mathbf{u}_j^t\mathbf{u}_j - \ln \mathbf{P}(\mathbf{w}_j)$$

$$= \frac{-1}{2\sigma^2}\mathbf{u}_i^t\mathbf{u}_i + \ln \frac{P(w_i)}{P(w_j)} + \frac{1}{2\sigma^2}\mathbf{u}_j^t\,\mathbf{u}_j$$

$$= \frac{-1}{2\sigma^2}\mathbf{u}_i^t\mathbf{u}_i + \frac{1}{2\sigma^2}\mathbf{u}_j^t\mathbf{u}_j + \frac{(u_i^t - u_j^t)(u_i - u_j)}{||u_i - u_j||^2} \ln \frac{P(w_i)}{P(w_j)}$$

$$= \frac{-1}{2\sigma^2}\mathbf{u}_i^t\mathbf{u}_i + \frac{1}{2\sigma^2}\mathbf{u}_j^t\mathbf{u}_j +$$

$$\frac{u_i^t(u_i - u_j)}{||u_i - u_j||^2} \ln \frac{P(w_i)}{P(w_j)} - \frac{u_j^t(u_i - u_j)}{||u_i - u_j||^2} \ln \frac{P(w_i)}{P(w_j)}$$

Now, by adding and subtracting the same term, we get:

$$= \frac{-1}{2\sigma^2}\mathbf{u}_i^t\mathbf{u}_i + \frac{1}{2\sigma^2}\mathbf{u}_j^t\mathbf{u}_i + \frac{u_i^t(u_i - u_j)}{||u_i - u_j||^2} \ln \frac{P(w_i)}{P(w_j)}$$

$$+ \frac{1}{2\sigma^2}\mathbf{u}_j^t\mathbf{u}_j - \frac{1}{2\sigma^2}\mathbf{u}_j^t\mathbf{u}_i - \frac{u_j^t(u_i - u_j)}{||u_i - u_j||^2} \ln \frac{P(w_i)}{P(w_j)}$$

$$= \frac{u_j^t}{\sigma^2}\left[\frac{1}{2}(\mathbf{u}_j + \mathbf{u}_i) - \sigma^2 \frac{u_i - u_j}{||u_i - u_j||^2} \ln \frac{P(w_i)}{P(w_j)}\right]$$

$$- \frac{u_i^t}{\sigma^2}\left[\frac{1}{2}(\mathbf{u}_j + \mathbf{u}_i) - \sigma^2 \frac{u_i - u_j}{||u_i - u_j||^2} \ln \frac{P(w_i)}{P(w_j)}\right]$$

By letting:

$$\mathbf{x}_0 = \frac{1}{2}(\mathbf{u}_j + \mathbf{u}_i) - \sigma^2 \frac{u_i - u_j}{||u_i - u_j||^2} \ln \frac{P(w_i)}{P(w_j)}$$

the result is:

$$\left(\frac{u_j^t}{\sigma^2} - \frac{u_i^t}{\sigma^2}\right)\mathbf{x}_0$$

But because of the way we define w_i and w_j, this is just:

$(w_j - w_i)^t x_0$

So from the original equation (2) we have :

$$(w_i - w_j)^t x - (w_i - w_j)^t x_0$$

and after multiplying through by sigma^2 the final decision boundary is given by:

$$(u_i - u_j)x - (u_i - u_j)x_0$$

No let $W = u_i - u_j$. Then this boundary is just :

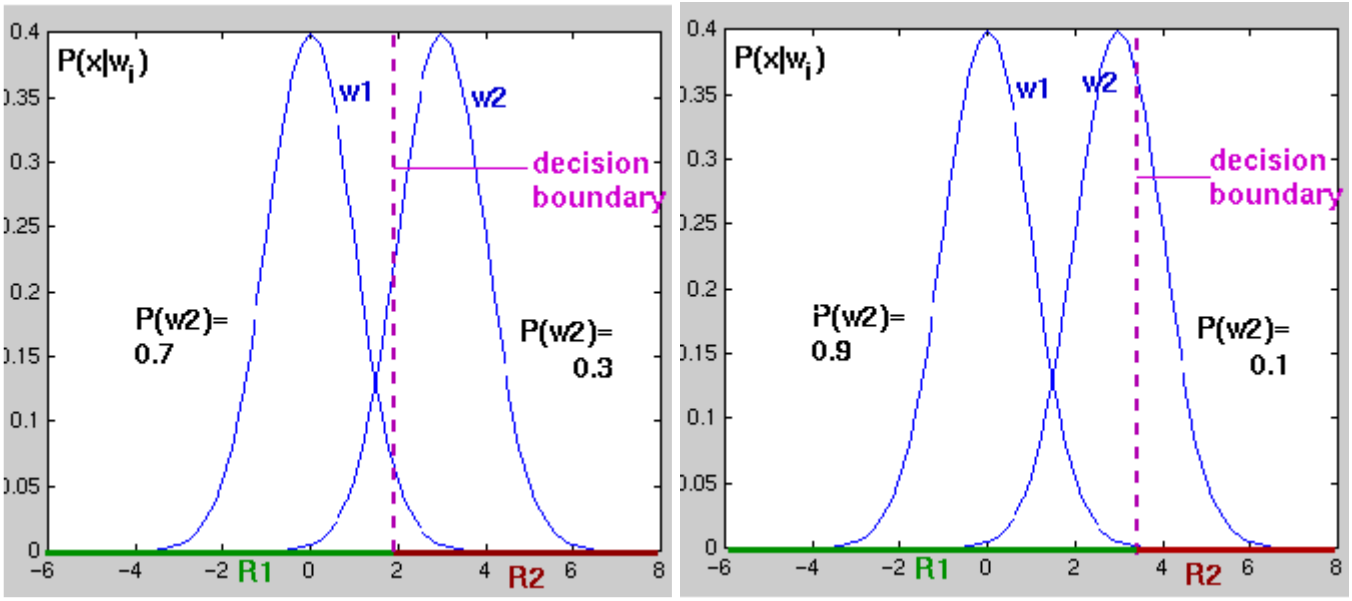
$$W^t(x - x_0)$$

This is called the *normal form* of the boundary equation. Geometrically, it defines a hyperplane through the point x_0 that is orthogonal to the vector W. But since $W = u_i - u_j$ then the plane which separates the decision regions for classes i and j is orthogonal to the line that links their means.

Imagine now that $P(w_i) = P(w_j)$. In this case, the second term in the expression for x_0 completely disappears, leaving only the first term which defines the midpoint of u_i and u_j. Therefore x_0 will lie exactly halfway between the means for class i and j, and therefore the decision boundary will equally divide the distance between the 2 means, with a decision region on either side. This makes intuitive sense. For example, suppose there are exactly the same number of oranges as apples on the conveyer belt entering the grocery store. If you then observe some color and weight features that are closer to the average color and weight for apples than they are to the average for oranges, the observer should classify the fruit as an apple. In general if the $P(w_i)$ are the same for all c classes, then the decision rule is based entirely on the distance from the feature vector x to the different mean vectors. The object will be classified to class i if it is closest to the mean vector for that class. This type of decision rule is called the *minimum-distance classifier*.

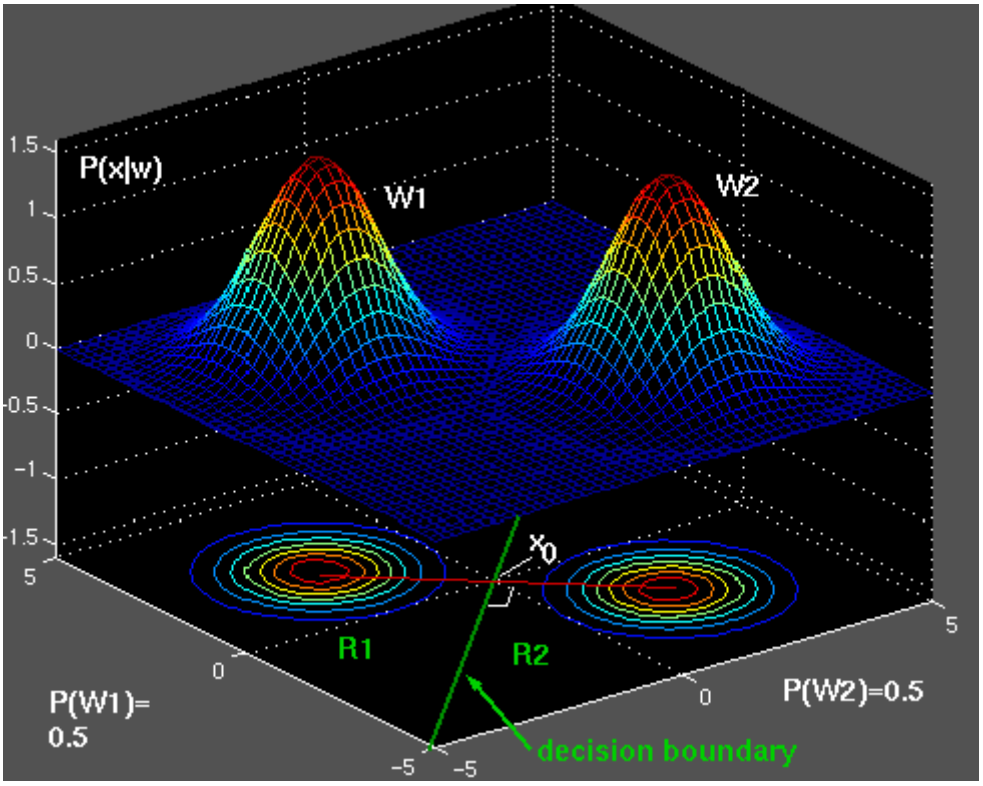
But what if P(w_i) != P(w_j)? In other words, there are 80% apples entering the store. If you observe some feature vector of color and weight that is just a little closer to the mean for oranges than the mean for apples, should the observer classify the fruit as an orange? The answer depends on how far from the apple mean the feature vector lies. In fact, if P(w_i) > P(w_j) then the second term in the equation for x_0 will subtract a positive amount from the first term. This will move point x_0 away from the mean for class i. If P(w_i) < P(w_j) then x_0 would tend to move away from the mean for class j. So for the above example and using the above decision rule, the observer will classify the fruit as an apple, simply because it's not *very* close to the mean for oranges, and because we know there are 80% apples in total. Below are 2 examples showing this fact:

Note that as the priors change in these 2 examples, the decision boundary through point x_0 shifts away from the more common class mean.

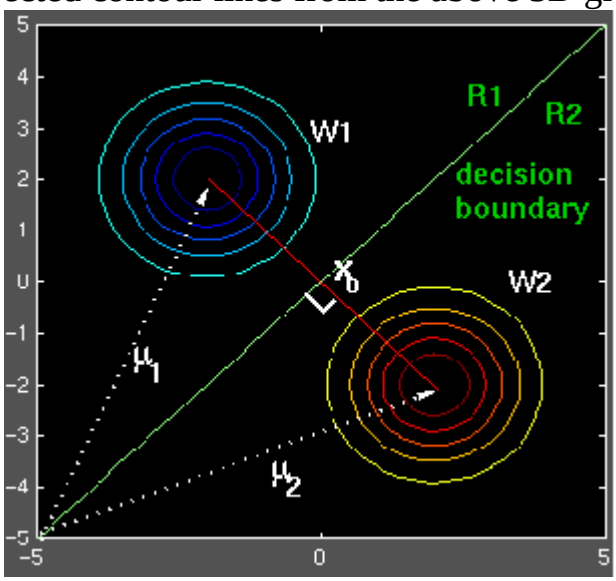


Below shows 2 bivariate normal distributions, whose priors are exactly the same. Therefore, the decision boundary is exactly at the midpoint between the 2 means. Note also that the decision boundary is a line orthogonal to the line joining the 2 means.

Tutorial on Bayesian Decision Theory with Gaussians

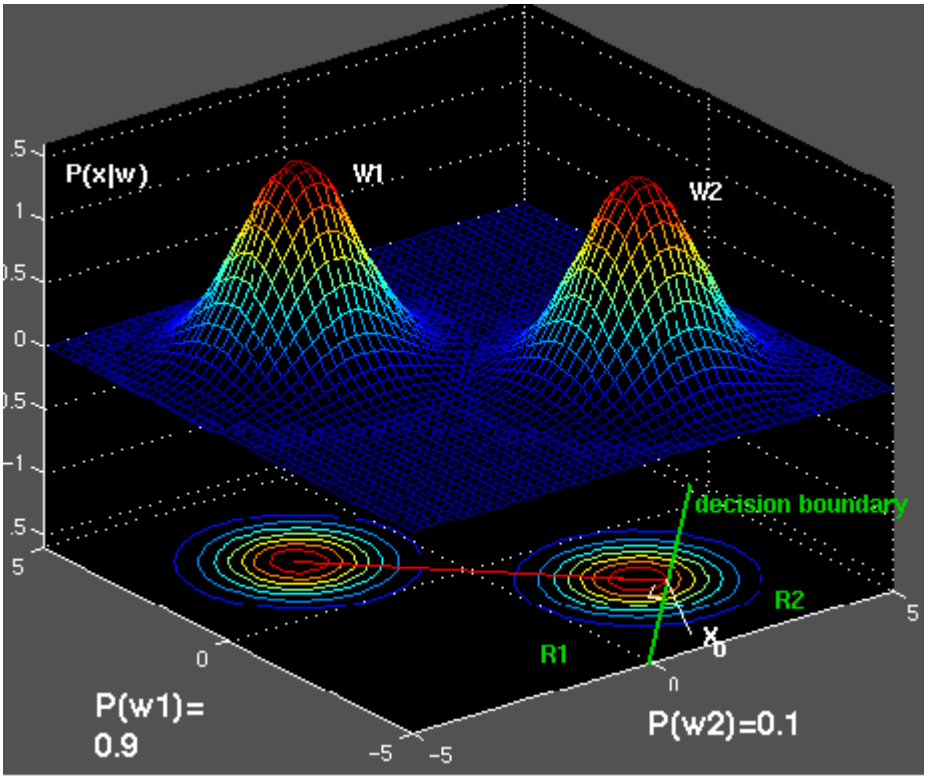


Projected contour lines from the above 3D graph.



A Second example where the priors have been changed. Note that the decision boundary has shifted away from the more likely class, although it is still orthogonal to the line joining the 2 means.

Note in the last 2 examples that because the bivariate normal densities have



diagonal covariance matrices, that their contours are spherical in shape. Secondly, because each class has the exact same covariance matrix, the circular lines forming the contours are the same size for both classes. This is because identical covariance matrices imply that the 2 classes have identically shaped clusters about their mean vectors.

The position of the decision boundary is not always greatly effected by the prior probabilities. After referring again to to the formula for $\underline{x_0}$, suppose that the variance constant was much smaller than the squared term $\|(u_i - u_j)\|$. Then the second term in the equation for x_0 would be multiplied by an extremely small constant. Thus even if $P(w_i)$ was much greater than $P(w_j)$, it would not effect the final position of x_0 to such an extent because the second term would only shift x_0 a very small amount away from the mean for class i.

Once again, consider the problem of classifying a random fruit. Suppose you know that almost all oranges are *exactly* one color of orange, and that almost all apples are *exactly* one color of red. Then if you observe a color that is closer to orange than to red, you should classify the fruit as an orange, even if there may be 80% apples in total. This is because it is now much less likely that your orange fruit is actually an apple, then it was when the color distribution for apples had a greater variation. In other words, the better our features are, the more we take them into account and ignore the a priori information. The worse our features are, the more the decision rule tends to ignore them and listens more to the a priori information.

Obviously not all features that we measure can be guaranteed to be independent of each other. What if the weight of a fruit depended somehow on its color? Perhaps ripe fruits of

a deeper color actually weighed more. In these cases, the covariance matrix will no longer be diagonal. The next section will consider this situation.

(The tutorial is based on http://www.cs.mcgill.ca/~mcleish/644/main.html. More case studies are available from that Web site.)