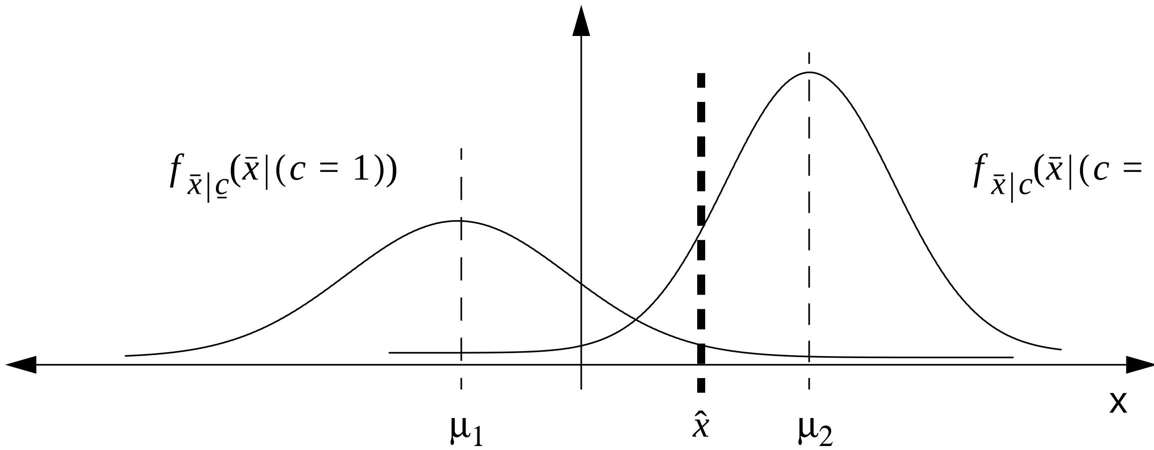
Maximum Likelihood Classification

Consider the problem of assigning a measurement to one of two sets:



What is the best criterion for making a decision?

Ideally, we would select the class for which the conditional probability is highest:

$$c^* = \underset{c}{\operatorname{argmax}} \; P((c = \hat{c})|(\bar{x} = \hat{\bar{x}}))$$

However, we can't estimate this probability directly from the training data. Hence, we consider:

$$c^* = \underset{c}{\operatorname{argmax}} \; P((\bar{x} = \hat{\bar{x}})|(c = \hat{c}))$$

By definition

$$P((c = \hat{c})|(\bar{x} = \hat{\bar{x}})) = \frac{P((c = \hat{c}),(\bar{x} = \hat{\bar{x}}))}{P(\bar{x} = \hat{\bar{x}})}$$

and

$$P((\bar{x} = \hat{\bar{x}})|(c = \hat{c})) = \frac{P((c = \hat{c}),(\hat{x} = \hat{\bar{x}}))}{P(c = \hat{c})}$$

from which we have

$$P((c = \hat{c})|(\bar{x} = \hat{\bar{x}})) = \frac{P((\bar{x} = \hat{\bar{x}})|(c = \hat{c}))P(c = \hat{c})}{P(\bar{x} = \hat{\bar{x}})}$$

Clearly, the choice of c that maximizes the right side also maximizes the left side. Therefore,

$$c^* = \underset{c}{\operatorname{argmax}}\ [P((\bar{x} = \hat{\bar{x}}) | (c = \hat{c}))]$$

$$= \underset{c}{\operatorname{argmx}}[P((\bar{x} = \hat{\bar{x}}) | (c = \hat{c}))P(c = \hat{c})]$$

if the class probabilities are equal,

$$c^* = \underset{c}{\operatorname{argmx}}[P((\bar{x} = \hat{\bar{x}}) | (c = \hat{c}))]$$

A quantity *related* to the probability of an event which is used to make a decision about the occurrence of that event is often called a *likelihood measure*.

A decision rule that maximizes a likelihood is called a maximum likelihood decision.

In a case where the number of outcomes is not finite, we can use an analogous continuous distribution. It is common to assume a multivariate Gaussian distribution:

$$f_{\bar{x}|c}(x_1,\ldots,x_N | c) = f_{\bar{x}|c}(\hat{\bar{x}} | \hat{c})$$

$$= \frac{1}{\sqrt{2\pi |C_{\bar{x}|c}|}} \exp\left\{-\frac{1}{2}(\hat{\bar{x}} - \bar{\mu}_{\bar{x}|c})^T \underline{C}_{\bar{x}|c}^{-1} (\hat{\bar{x}} - \bar{\mu}_{\hat{\bar{x}}|c})\right\}$$

We can elect to maximize the log, $\ln[f_{\bar{x}|c}(\bar{x}|c)]$ rather than the likelihood (we refer to this as the log likelihood). This gives the decision rule:

$$c^* = \underset{c}{\operatorname{argmin}}\left[(\hat{\bar{x}} - \bar{\mu}_{\bar{x}|c})^T \underline{C}_{\bar{x}|c}{}^{-1} (\hat{\bar{x}} - \bar{\mu}_{\hat{\bar{x}}|c}) + \ln\left\{\left|\underline{C}_{\bar{x}|c}^{-1}\right|\right\}\right]$$

(Note that the maximization became a minimization.)

We can define a distance measure based on this as:

$$d_{ml}(\bar{x}, \bar{\mu}_{\bar{x}|c}) = (\hat{\bar{x}} - \bar{\mu}_{\bar{x}|c})^T \underline{C}_{\bar{x}|c}{}^{-1} (\hat{\bar{x}} - \bar{\mu}_{\hat{\bar{x}}|c}) + \ln\left\{\left|\underline{C}_{\bar{x}|c}^{-1}\right|\right\}$$

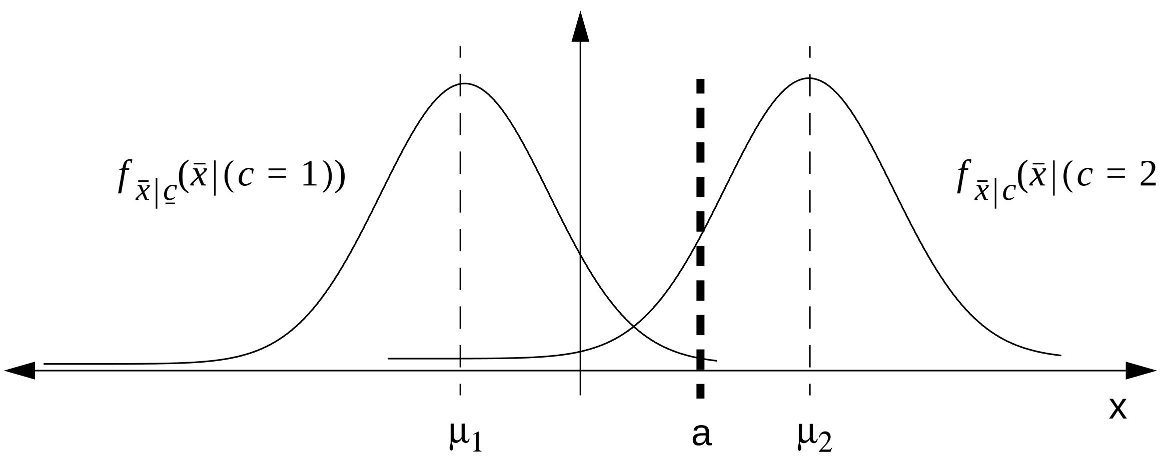
Note that the distance is conditioned on each class mean and covariance. This is why “generic” distance comparisons are a joke.

If the mean and covariance are the same across all classes, this expression simplifies to:

$$d_M(\bar{x}, \bar{\mu}_{\bar{x}|c}) = (\hat{\bar{x}} - \bar{\mu}_{\bar{x}|c})^T \underline{C}_{\bar{x}|c}^{-1} (\hat{\bar{x}} - \bar{\mu}_{\hat{\bar{x}}|c})$$

This is frequently called the *Mahalanobis distance*. But this is nothing more than a weighted Euclidean distance.

This result has a relatively simple geometric interpretation for the case of a single random variable with classes of equal variances:



The decision rule involves setting a threshold:

$$a = \left(\frac{\mu_1 + \mu_2}{2}\right) + \frac{\sigma^2}{\mu_1 - \mu_2} \ln\left(\frac{P(c = 2)}{P(c = 1)}\right)$$

and,

$$\begin{array}{lll} if & x < a & x \in (c = 1) \\ else & & x \in (c = 2) \end{array}$$

If the variances are not equal, the threshold shifts towards the distribution with the smaller variance.

What is an example of an application where the classes are not equiprobable?

Probabilistic Distance Measures

How do we compare two probability distributions to measure their overlap?

Probabilistic distance measures take the form:

$$J = \int_{-\infty}^{\infty} g\{ f_{\bar{x}|c}(\hat{\bar{x}}|\hat{c}), P(c = \hat{c}), \hat{c} = 1, 2, \ldots, K\} d\hat{\bar{x}}$$

where

1. J is nonnegative
2. J attains a maximum when all classes are disjoint
3. J=0 when all classes are equiprobable

Two important examples of such measures are:

(1) Bhattacharyya distance:

$$J_B = -\ln\left[\int_{-\infty}^{\infty} \sqrt{f_{\bar{x}|c}(\hat{\bar{x}}|1)\, f_{\bar{x}|c}(\hat{\bar{x}}|2)}\, d\hat{\bar{x}}\right]$$

(2) Divergence

$$J_D = \int_{-\infty}^{\infty} [f_{\bar{x}|c}(\hat{\bar{x}}|1) - f_{\bar{x}|c}(\hat{\bar{x}}|2)] \ln\left[\frac{f_{\bar{x}|c}(\hat{\bar{x}}|1)}{f_{\bar{x}|c}(\hat{\bar{x}}|2)}\right] d\hat{\bar{x}}$$

Both reduce to a Mahalanobis-like distance for the case of Gaussian vectors with equal class covariances.

Such metrics will be important when we attempt to cluster feature vectors and acoustic models.

Probabilistic Dependence Measures

A probabilistic dependence measure indicates how strongly a feature is associated with its class assignment. When features are independent of their class assignment, the class conditional pdf's are identical to the mixture pdf:

$$f_{\bar{x}|c}(\hat{\bar{x}}|\hat{c}) = f_{\bar{x}}(\hat{\bar{x}}) \qquad \forall c$$

When their is a strong dependence, the conditional distribution should be significantly different than the mixture. Such measures take the form:

$$J = \int_{-\infty}^{\infty} g\{ f_{\bar{x}|c}(\hat{\bar{x}}|\hat{c}), f_{\bar{x}}(\hat{\bar{x}}), P(c = \hat{c}), \hat{c} = 1, 2, \ldots, K\} d\hat{\bar{x}}$$

An example of such a measure is the average mutual information:

$$M_{avg}(c, \hat{\bar{x}}) = \sum_{c=1}^{K} P(c = \hat{c}) \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f_{\bar{x}|c}(\hat{\bar{x}}|\hat{c}) \log \frac{f_{\bar{x}|c}(\hat{\bar{x}}|\hat{c})}{f_{\bar{x}}(\hat{\bar{x}})} d\hat{\bar{x}}$$

The discrete version of this is:

$$M_{avg}(c, \hat{\bar{x}}) = \sum_{c=1}^{K} P(c = \hat{c}) \sum_{i=1}^{L} P(\bar{x} = \bar{x}_l) \log_2 \frac{P(\bar{x} = \bar{x}_l | c = \hat{c})}{P(\bar{x} = \bar{x}_l)}$$

Mutual information is closely related to entropy, as we shall see shortly.

Such distance measures can be used to cluster data and generate vector quantization codebooks. A simple and intuitive algorithms is known as the K-means algorithm:

Initialization: Choose K centroids

Recursion:

1. Assign all vectors to their nearest neighbor.
2. Recompute the centroids as the average of all vectors assigned to the same centroid.
3. Check the overall distortion. Return to step 1 if some distortion criterion is not met.