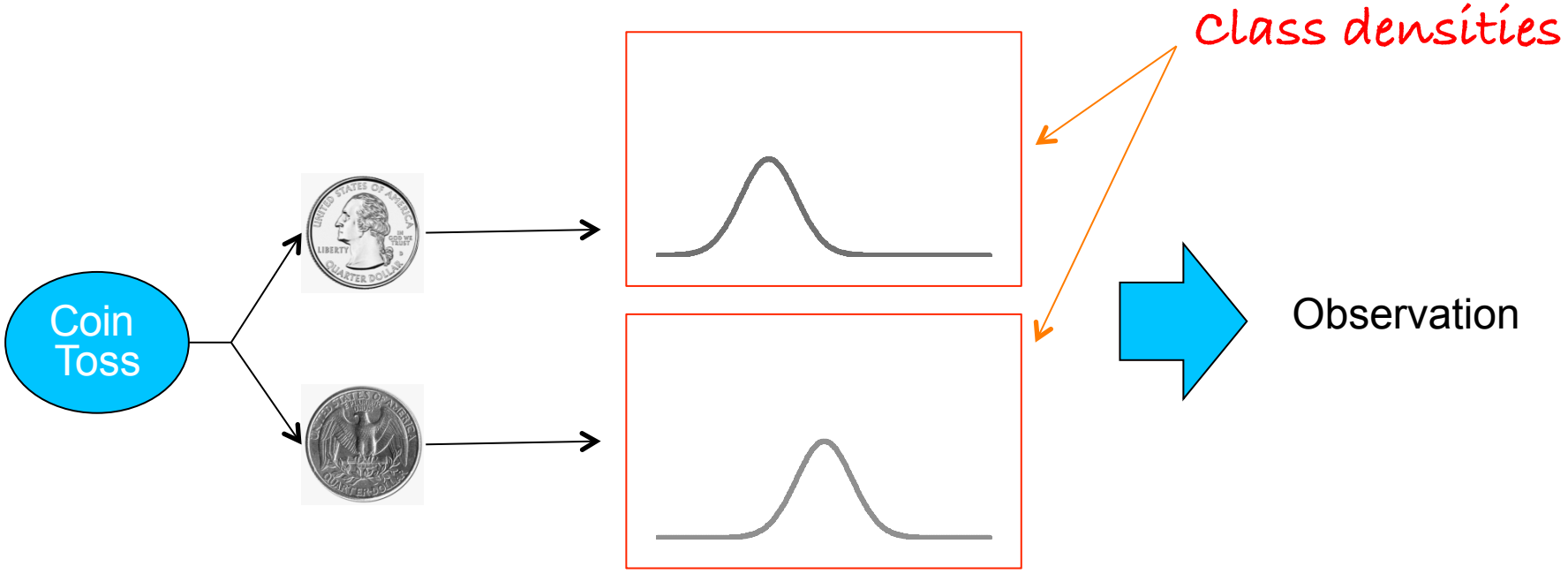
Optimal Decision Rule

The optimal decision rule that minimizes the total risk is to maximize the posterior probability

$$\text{Choose Class } k \text{ if } p(x \mid k)\pi_k > p(x \mid k')\pi_{k'} \text{ for all } k' \neq k;$$

We would like to see what the rule is when the two class densities are Gaussian and when we operate in a higher dimensional feature space (ie we deal with more than one feature at a time)

A Classification Problem



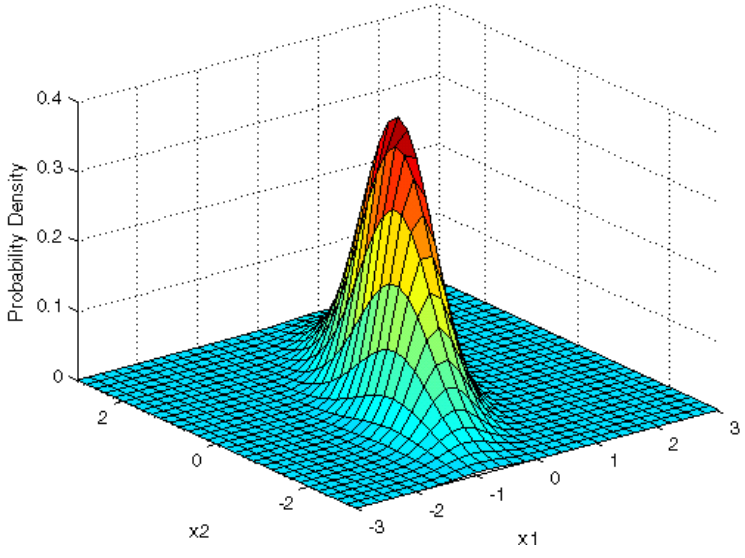
We want to look at the special case when:

- The class densities are Gaussian
- The class densities have the same variance
- The class densities differ in their means
- The observation may be a feature vector

Gaussian Density Function

p features organized as a feature vector

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}} \frac{1}{|\mathbf{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{m})}$$



The Gaussian density for a p-dimensional vector has two parameters: the mean vector and the covariance matrix.

- The density is centered at the mean.
- The spread of the density is controlled by the covariance matrix.

Maximum a Posteriori Decision Rule

Observe feature vector $\mathbf{x}$. We want to decide among K classes.

Choose $l = k$ to maximize the posterior probability $p(k \mid \mathbf{x})$.

Using Bayes' theorem, we can show that this is equivalent to maximizing $\pi_k p(\mathbf{x} \mid k)$.

When the class densities are Gaussian, we want to maximize

$$\pi_k p(\mathbf{x} \mid k) = \pi_k \frac{1}{(2\pi)^{p/2}} \frac{1}{|\mathbf{\Sigma}_k|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m}_k)^T \mathbf{\Sigma}_k^{-1} (\mathbf{x}-\mathbf{m}_k)}.$$

Gaussian Class Densities That Differ Only In Their Means

When the class densities are Gaussian, we want to choose $l = k$ to maximize

$$\pi_k p(\mathbf{x} \mid k) = \pi_k \frac{1}{(2\pi)^{p/2}} \frac{1}{|\mathbf{\Sigma}_k|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m}_k)^T \mathbf{\Sigma}_k^{-1}(\mathbf{x}-\mathbf{m}_k)}.$$

When the class densities differ only in their means, we want to maximize

$$\pi_k e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m}_k)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{m}_k)} \quad \text{or, equivalently,} \quad \log\pi_k - \frac{1}{2}(\mathbf{x}-\mathbf{m}_k)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{m}_k),$$

which is the same as minimizing

$$\frac{1}{2}(\mathbf{x}-\mathbf{m}_k)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{m}_k) - \log\pi_k.$$

This slide is optional

Gaussian Class Densities

When the class densities are Gaussian, we want to choose $l = k$ to maximize

$$\pi_k p(\mathbf{x} \mid k) = \pi_k \frac{1}{(2\pi)^{p/2}} \frac{1}{|\mathbf{\Sigma}_k|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m}_k)^T \mathbf{\Sigma}_k^{-1}(\mathbf{x}-\mathbf{m}_k)}.$$

When the class densities differ only in their means, we want to maximize

$$\pi_k e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m}_k)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{m}_k)} \quad \text{or, equivalently,} \quad \log\pi_k - \frac{1}{2}(\mathbf{x}-\mathbf{m}_k)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{m}_k),$$

which is the same as minimizing

$$\frac{1}{2}(\mathbf{x}-\mathbf{m}_k)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{m}_k) - \log\pi_k.$$

This slide is optional

Gaussian Class Densities for K=2

The term $\frac{1}{2}(\mathbf{x}-\mathbf{m}_k)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mathbf{m}_k)-\log\pi_k$ expands to

$$\frac{1}{2}\left(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}-\mathbf{m}_k{}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}-\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{m}_k+\mathbf{m}_k{}^T\boldsymbol{\Sigma}^{-1}\mathbf{m}_k\right)-\log\pi_k$$

so that the rule is to choose $l=k$ to minimize

$$\frac{1}{2}\mathbf{m}_k{}^T\boldsymbol{\Sigma}^{-1}\mathbf{m}_k-\mathbf{m}_k{}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}-\log\pi_k.$$

When $K=2$, the rule is to choose $l=0$ if

$$\frac{1}{2}\mathbf{m}_0{}^T\boldsymbol{\Sigma}^{-1}\mathbf{m}_0-\mathbf{m}_0{}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}-\log\pi_0<\frac{1}{2}\mathbf{m}_1{}^T\boldsymbol{\Sigma}^{-1}\mathbf{m}_1-\mathbf{m}_1{}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}-\log\pi_1.$$

This slide is optional

Gaussian Class Densities for K=2

When $K = 2$, the rule is to choose $l = 0$ if

$$\frac{1}{2}\mathbf{m}_0{}^T\boldsymbol{\Sigma}^{-1}\mathbf{m}_0 - \mathbf{m}_0{}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} - \log\pi_0 < \frac{1}{2}\mathbf{m}_1{}^T\boldsymbol{\Sigma}^{-1}\mathbf{m}_1 - \mathbf{m}_1{}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} - \log\pi_1.$$

Rearranging the terms, we have

$$\mathbf{m}_1{}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} - \mathbf{m}_0{}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} < \frac{1}{2}\left(\mathbf{m}_1{}^T\boldsymbol{\Sigma}^{-1}\mathbf{m}_1 - \mathbf{m}_0{}^T\boldsymbol{\Sigma}^{-1}\mathbf{m}_0\right) + \log\frac{\pi_0}{\pi_1},$$

or,

$$\left(\mathbf{m}_1 - \mathbf{m}_0\right)^T\boldsymbol{\Sigma}^{-1}\mathbf{x} < \frac{1}{2}\left(\mathbf{m}_1{}^T\boldsymbol{\Sigma}^{-1}\mathbf{m}_1 - \mathbf{m}_0{}^T\boldsymbol{\Sigma}^{-1}\mathbf{m}_0\right) + \log\frac{\pi_0}{\pi_1}.$$

Classification

This slide is optional

When $K = 2$, the rule is to choose $l = 0$ if

$$(\mathbf{m}_1 - \mathbf{m}_0)^T \boldsymbol{\Sigma}^{-1}\mathbf{x} < \frac{1}{2}\left(\mathbf{m}_1{}^T \boldsymbol{\Sigma}^{-1}\mathbf{m}_1 - \mathbf{m}_0{}^T \boldsymbol{\Sigma}^{-1}\mathbf{m}_0\right) + \log\frac{\pi_0}{\pi_1}.$$

Adding a term $\left(\mathbf{m}_1{}^T \boldsymbol{\Sigma}^{-1}\mathbf{m}_0 - \mathbf{m}_0{}^T \boldsymbol{\Sigma}^{-1}\mathbf{m}_1\right)$ to $\left(\mathbf{m}_1{}^T \boldsymbol{\Sigma}^{-1}\mathbf{m}_1 - \mathbf{m}_0{}^T \boldsymbol{\Sigma}^{-1}\mathbf{m}_0\right)$,

we have $\left(\mathbf{m}_1{}^T \boldsymbol{\Sigma}^{-1}\mathbf{m}_1 + \mathbf{m}_1{}^T \boldsymbol{\Sigma}^{-1}\mathbf{m}_0 - \mathbf{m}_0{}^T \boldsymbol{\Sigma}^{-1}\mathbf{m}_1 - \mathbf{m}_0{}^T \boldsymbol{\Sigma}^{-1}\mathbf{m}_0\right)$,

which simplifies to

$$\left(\mathbf{m}_1{}^T \boldsymbol{\Sigma}^{-1}(\mathbf{m}_1 + \mathbf{m}_0) - \mathbf{m}_0{}^T \boldsymbol{\Sigma}^{-1}(\mathbf{m}_1 + \mathbf{m}_0)\right) = (\mathbf{m}_1 - \mathbf{m}_0)^T \boldsymbol{\Sigma}^{-1}(\mathbf{m}_1 + \mathbf{m}_0).$$

The rule then becomes choose $l = 0$ if

$$(\mathbf{m}_1 - \mathbf{m}_0)^T \boldsymbol{\Sigma}^{-1}\mathbf{x} < (\mathbf{m}_1 - \mathbf{m}_0)^T \boldsymbol{\Sigma}^{-1}\frac{(\mathbf{m}_1 + \mathbf{m}_0)}{2} + \log\frac{\pi_0}{\pi_1}.$$

After considerable simplification...

Linear Classifier

Suppose $K = 2$.

When the class densities are Gaussian that differ only in their means, the optimal MAP rule is to choose $l = 0$ if

$$\left(\mathbf{m}_1 - \mathbf{m}_0\right)^T \boldsymbol{\Sigma}^{-1}\mathbf{x} < \left(\mathbf{m}_1 - \mathbf{m}_0\right)^T \boldsymbol{\Sigma}^{-1} \frac{\left(\mathbf{m}_1 + \mathbf{m}_0\right)}{2} + \log\frac{\pi_0}{\pi_1}.$$

Let $\Theta = \left(\mathbf{m}_1 - \mathbf{m}_0\right)^T \boldsymbol{\Sigma}^{-1} \frac{\left(\mathbf{m}_1 + \mathbf{m}_0\right)}{2} + \log\frac{\pi_0}{\pi_1}$ be the threshold term,

and $\mathbf{w} = \boldsymbol{\Sigma}^{-T}\left(\mathbf{m}_1 - \mathbf{m}_0\right)$ be a $(p \times 1)$ weight vector.

The decision rule is linear (in $\mathbf{x}$): Choose $l = 0$ if $\mathbf{w}^T\mathbf{x} < \Theta$.

Let a bias term w_p be defined as $w_p = -\Theta$.

Define a function g as $g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_p$.

In terms of g, the decision rule is to choose $l = 0$ if $g(\mathbf{x}) < 0$.

Geometric Interpretation of the Linear Classifier

The optimal MAP rule is to choose $l = 0$ if

$$\left(\mathbf{m}_1 - \mathbf{m}_0\right)^T \Sigma^{-1}\mathbf{x} < \left(\mathbf{m}_1 - \mathbf{m}_0\right)^T \Sigma^{-1} \frac{\left(\mathbf{m}_1 + \mathbf{m}_0\right)}{2} + \log\frac{\pi_0}{\pi_1}$$

A line joining the two means (in the feature space)

Length of the projection of a point ($\Sigma^{-1}\mathbf{x}$) on the line joining the two means (in the feature space)

Geometric Interpretation of the Linear Classifier

Simplify for now to assume (1) the two class prior probabilities are equal and (2) the covariance matrix is the identity matrix

The optimal MAP rule is to choose $l = 0$ if

$$\left(\mathbf{m}_1 - \mathbf{m}_0\right)^T \mathbf{x} < \left(\mathbf{m}_1 - \mathbf{m}_0\right)^T \frac{\left(\mathbf{m}_1 + \mathbf{m}_0\right)}{2}$$

A line joining the two means (in the feature space)

Length of the projection of the point $\mathbf{x}$ on the line joining the two means (in the feature space)

The rule says call Class 0 if the length of the projection of the point $\mathbf{x}$ on the line joining the two means is closer to the class mean 0 (smaller than the midpoint from $\mathbf{m}_0$ to $\mathbf{m}_1$)

When the two prior probabilities are not the same

The optimal MAP rule is to choose $l = 0$ if

$$\left(\mathbf{m}_1 - \mathbf{m}_0\right)^T \mathbf{x} < \left(\mathbf{m}_1 - \mathbf{m}_0\right)^T \frac{\left(\mathbf{m}_1 + \mathbf{m}_0\right)}{2} + \log \frac{\pi_0}{\pi_1}$$

When the two prior probabilities are not the same, $\log \frac{\pi_0}{\pi_1} \neq 0$.

Suppose $\pi_0 > \pi_1$, i.e. Class 0 is more likely to occur than Class 1.

Then $\frac{\pi_0}{\pi_1} > 1$ so that $\log \frac{\pi_0}{\pi_1} > 0$.

This pushes the deciding threshold away from $\mathbf{m}_0$ along the line joining it to $\mathbf{m}_1$.

And vice versa when $\pi_0 < \pi_1$.

When the covariance matrix is not identity

- What is the covariance matrix?
  - The (i,j)th element is the correlation of the ith and jth feature-pair
    - Is zero of the pair is uncorrelated
    - Is large if the pair is highly correlated
  - The ith diagonal term is the variance of the ith feature
    - Magnitude depends on the scale and the randomness of the feature

When the covariance matrix is not identity

- The covariance matrix is the identity matrix when:
  - The features are pairwise uncorrelated
  - All features have variance 1 (ie they have the same scale)

When the covariance matrix is not identity

- Given a feature vector with an arbitrary (ie non-identity) covariance matrix, we can transform the features by a linear transformation so that the transformed features will have an identity covariance matrix
  - Decorrelate the features
  - Scale each feature to unit variance
- This process is called “whitening” the data and is accomplished by

$$\mathbf{y} = \Sigma^{-1/2}\mathbf{x}$$

When the covariance matrix is not identity

Given an arbitrary feature vector, we (conceptually) whiten the data by $\mathbf{y} = \Sigma^{-1/2}\mathbf{x}$. We operate in the transformed space, so that the class means have to be transformed as well, to $\Sigma^{-1/2}\mathbf{m}_1$ and $\Sigma^{-1/2}\mathbf{m}_0$.

The optimal MAP rule is to choose $l = 0$ if

$$\left(\Sigma^{-1/2}\mathbf{m}_1 - \Sigma^{-1/2}\mathbf{m}_0\right)^T \Sigma^{-1/2}\mathbf{x} < \left(\Sigma^{-1/2}\mathbf{m}_1 - \Sigma^{-1/2}\mathbf{m}_0\right)^T \frac{\left(\Sigma^{-1/2}\mathbf{m}_1 + \Sigma^{-1/2}\mathbf{m}_0\right)}{2} + \log\frac{\pi_0}{\pi_1}$$

By factoring out $\Sigma^{-1/2}$ and writing $\left(\Sigma^{-1/2}\right)^T \Sigma^{-1/2} = \Sigma^{-1}$, we have

$$\left(\mathbf{m}_1 - \mathbf{m}_0\right)^T \Sigma^{-1}\mathbf{x} < \left(\mathbf{m}_1 - \mathbf{m}_0\right)^T \Sigma^{-1} \frac{\left(\mathbf{m}_1 + \mathbf{m}_0\right)}{2} + \log\frac{\pi_0}{\pi_1}.$$

Optimal Decision Rule

When the two classes

- Are Gaussian
- Have the same covariance matrix

the optimal MAP rule is a linear classifier

We call it a linear classifier because the feature vector values are linearly combined as a weighted sum and compared to a threshold.

If we know the class densities and the prior probabilities, the solution is known