

Feb 4 (next Monday)

- No class

Topics

Optimal decision theory

Unsupervised methods

- Projection methods
 - Principal component analysis
 - Independent component analysis
- Clustering
- Self-organizing maps

Supervised methods

- Linear classifiers
- Artificial neural networks
 - Deep learning
- Kernel methods
- Nearest neighbor classifier
- Learning vector quantizers

Implementations

- Python

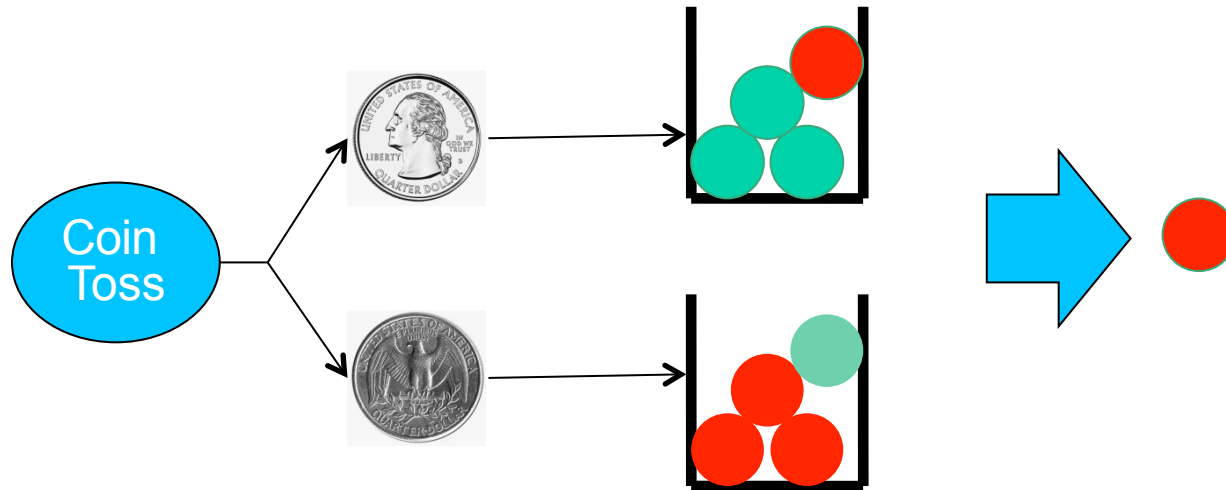
What is Pattern Recognition?

- “The field of pattern recognition is concerned
- with the automatic discovery of regularities in data through the use of computer algorithms and
 - with the use of these regularities to take actions such as classifying the data into different categories”

An Example

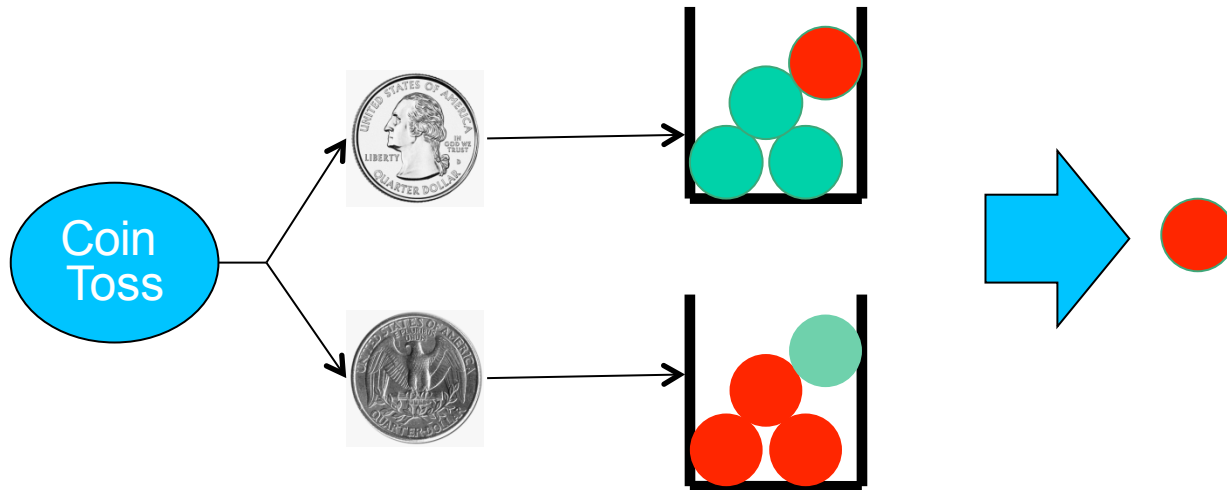
- Events
 - Red ball, green ball

Defining the probability of drawing a red ball or green ball is not so straightforward.



Probabilities

- What is the probability of observing ● ?

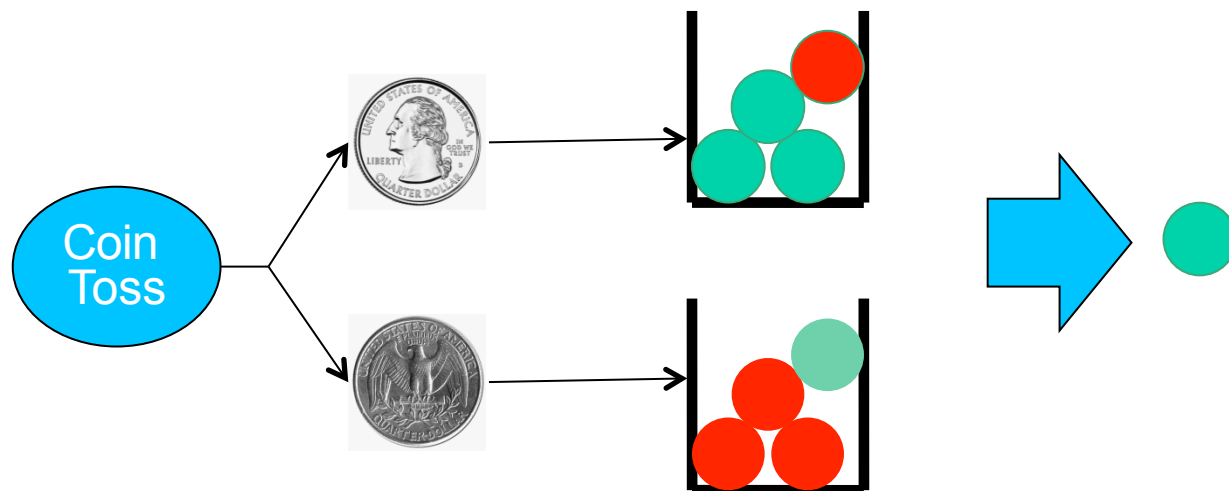


$$P(\text{red}) = P(\text{red} \mid \text{head})P(\text{head}) + P(\text{red} \mid \text{tail})P(\text{tail})$$

$$= \frac{1}{4} \frac{1}{2} + \frac{3}{4} \frac{1}{2} = \frac{1}{2}.$$

Probability (● observed)

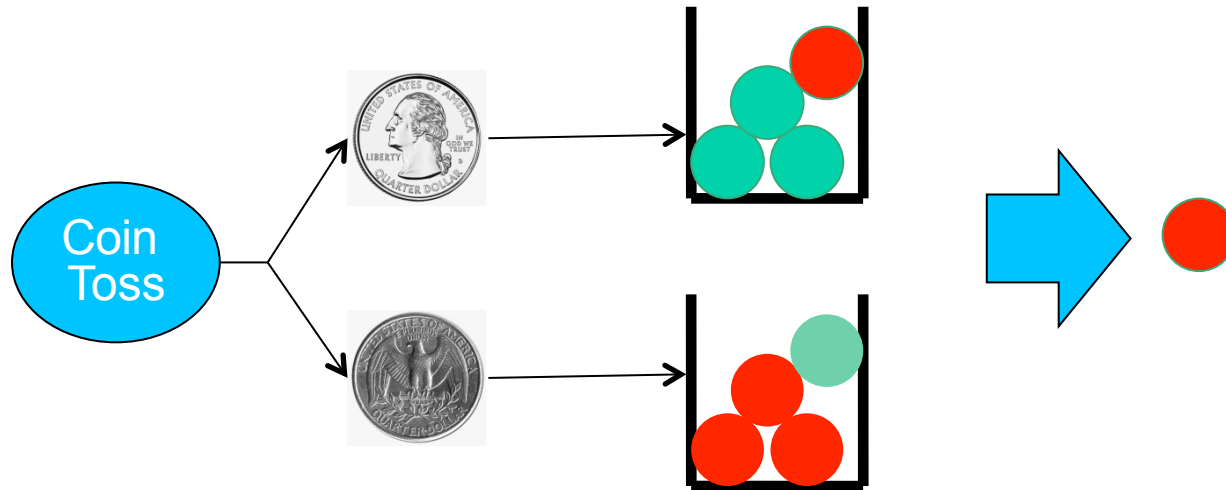
Similarly,



$$P(\text{green}) = \frac{1}{4} \frac{1}{2} + \frac{3}{4} \frac{1}{2} = \frac{1}{2}$$

Decision Rule

- Events
 - Red ball, green ball



Suppose a red ball is observed, how do we decide the outcome of the coin toss?

Keeping Score

- Suppose A pays B \$1 for every correct call, and B pays A \$1 for every incorrect call.
- After, say, 100 calls by B , how much money do you think B will have?

Expected Value

“Suppose A pays B \$1 for every correct call, and B pays A \$1 for every incorrect call.”

Define x as the amount of money that B has.

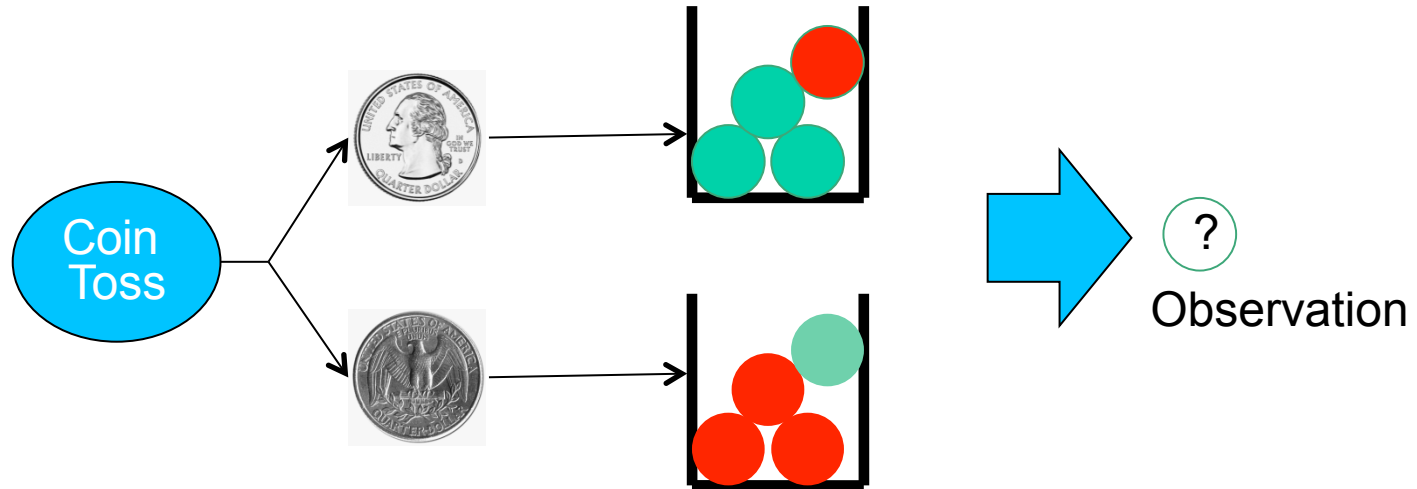
From the rule, x is given by

$$x = \begin{cases} +1 & \text{if the decision is correct} \\ -1 & \text{if the decision is incorrect} \end{cases}$$

What is $E[x]$?

When is a decision correct?

- Need a decision rule



- Let us make up a decision rule

$$d(\text{observation}) = \begin{cases} \text{head} & \text{if observation is green} \\ \text{tail} & \text{if observation is red} \end{cases}$$

Expected Value

“Suppose A pays B \$1 for every correct call, and B pays A \$1 for every incorrect call.”

Define x as the amount of money that B has.

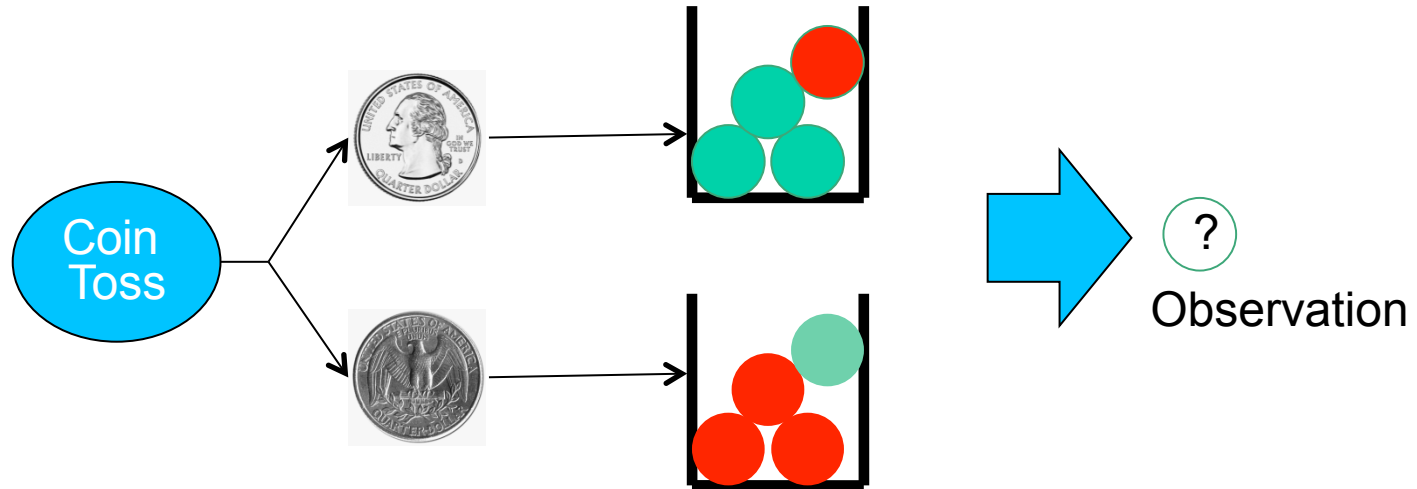
From the rule, x is given by

$$x = \begin{cases} +1 & \text{if the decision is correct} \\ -1 & \text{if the decision is incorrect} \end{cases}$$

The expected value of x is

$$\begin{aligned} E[x] &= +1 \times P(\text{correct decision}) - 1 \times P(\text{incorrect decision}) \\ &= +1 \times \frac{3}{4} - 1 \times \frac{1}{4} = \frac{2}{4} = \frac{1}{2}. \end{aligned}$$

Pattern Recognition

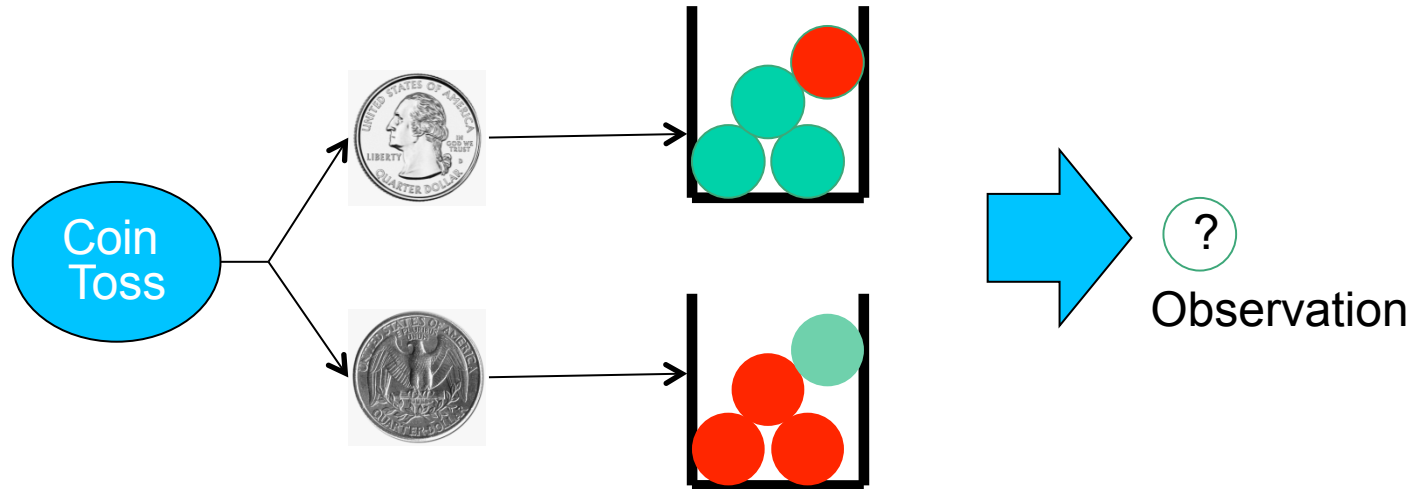


Pattern Recognition Problem

Based on the observed color of the drawn ball, and knowing the probability of the coin toss and composition of each of the buckets, decide the outcome of the coin toss.

In practice, we do not know the exact composition of each of the buckets. That is why we need to do learning, supervised or unsupervised.

Pattern Recognition



Pattern Recognition Problem

Based on the observed color of the drawn ball, and knowing the probability of the coin toss and composition of each of the buckets, decide the outcome of the coin toss.

What is the optimal decision rule? ←

What do we mean by “optimal”

How do we optimize?

Optimal Decision Rule

Suppose A pays B \$1 for every correct call, and B pays A \$1 for every incorrect call.”

Assume B is making the decision

A decision rule is optimum when the expected value of the payout is maximized

Theoretically we have to search the entire decision rule space to find the optimal one.

In practice, if the decision rule is parameterized, we optimize over the parameters

Optimal Decision Rule

Suppose A (the “world”) penalizes B \$1 for every incorrect call.”

Assume B is making the decision

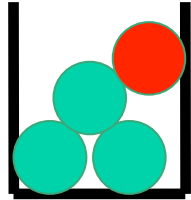
Believe it or not this is the more realistic assumption

A decision rule is optimum when the expected value of the penalty is minimized

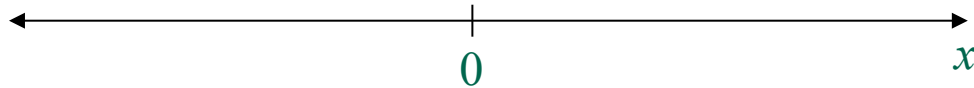
Theoretically we have to search the entire decision rule space to find the optimal one.

In practice, if the decision rule is parameterized, we optimize over the parameters

Events and Random Variables



$\{c = \text{red}, c = \text{green}\}$

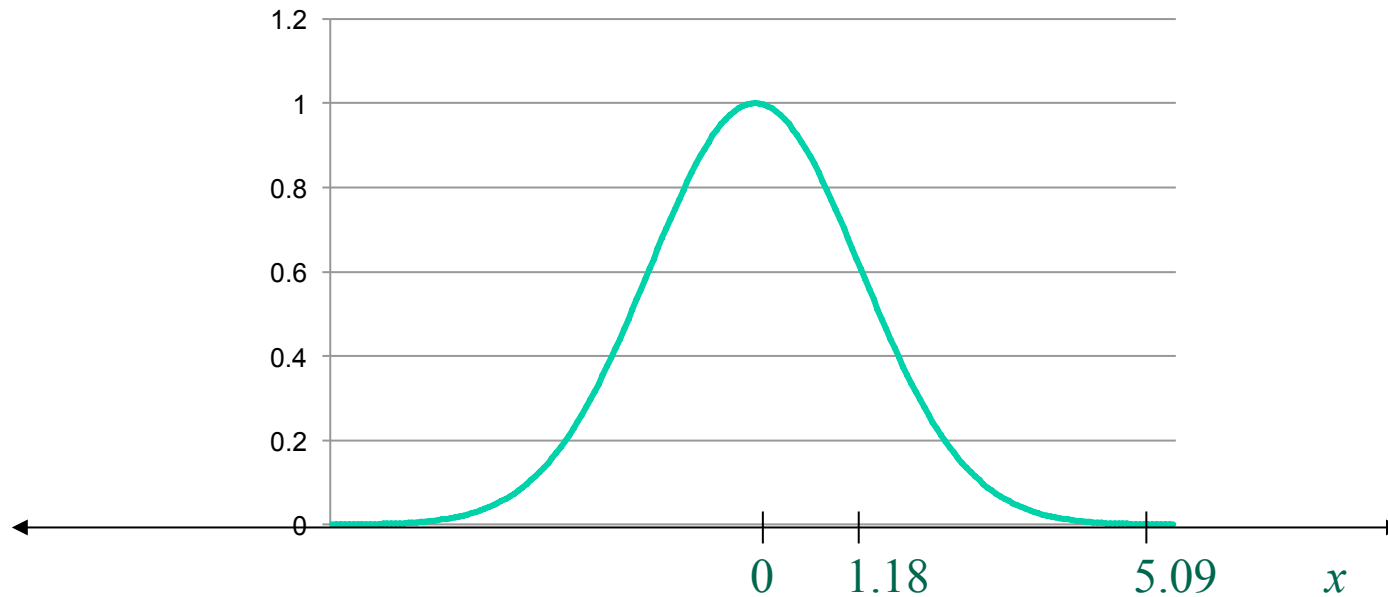


$\{x = 5.09, 1.18 \leq x < 5.09\}$



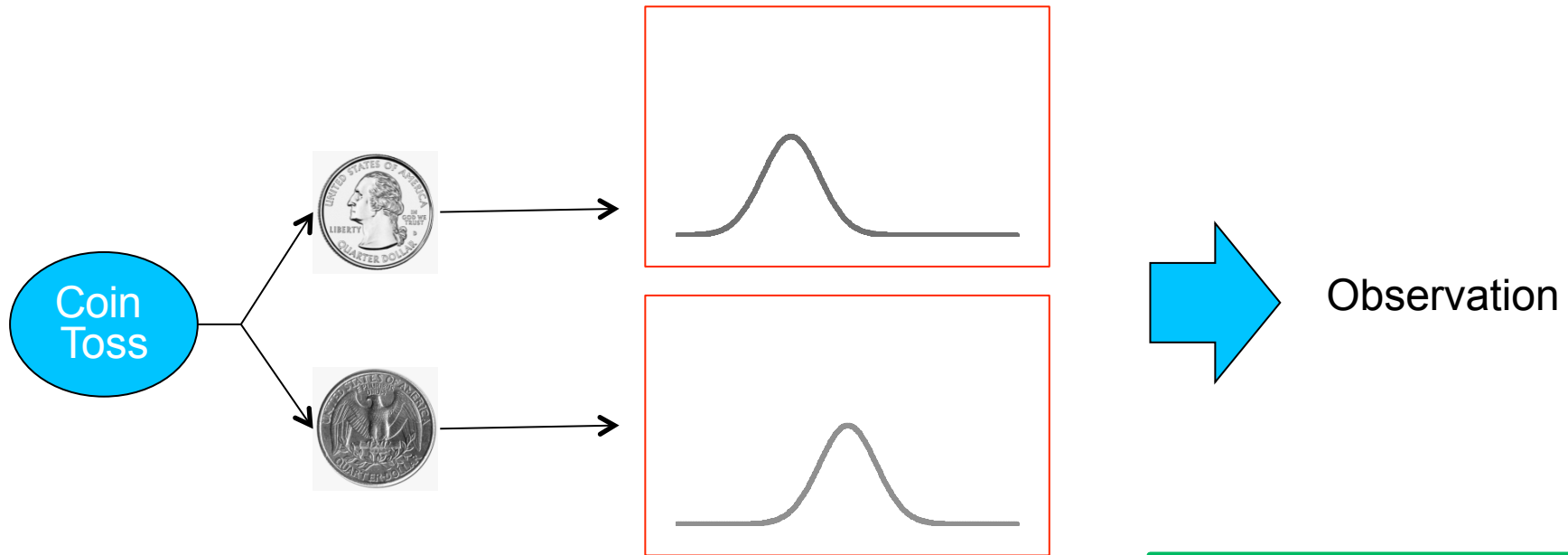
$\{d = 1, d = 2, \dots, d = 6\}$

Gaussian density



$P\{1.18 \leq x < 5.09\}$ = area under the density curve from 1.18 to 5.09

Another Pattern Recognition Problem



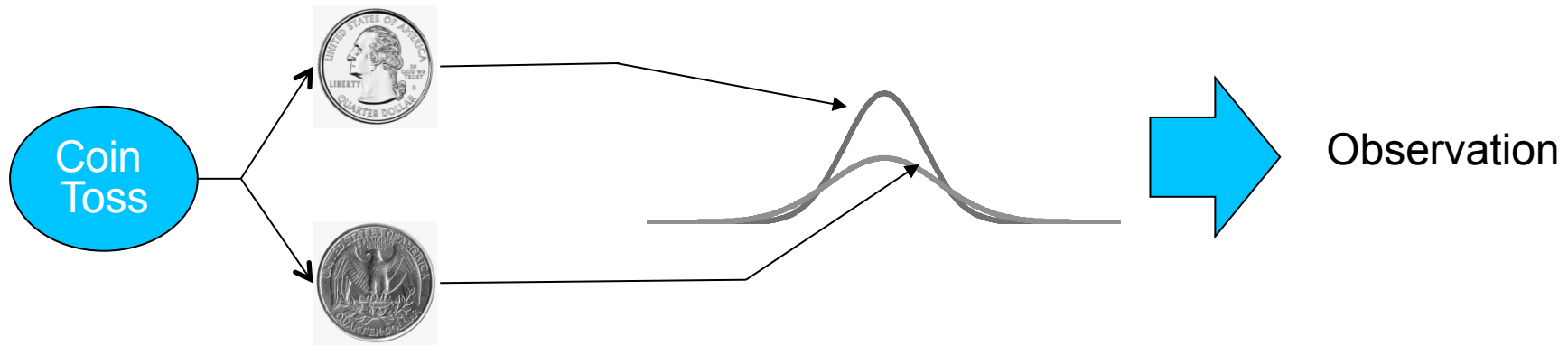
Pattern Recognition Problem

Based on the observed measurement,
and knowing the probability of the coin toss
and the densities of the measurement process,

decide the outcome of the coin toss.

In practice, we do not know
the exact densities.
That is why we need to do
learning, supervised or
unsupervised.

Yet Another Pattern Recognition Problem



Pattern Recognition Problem

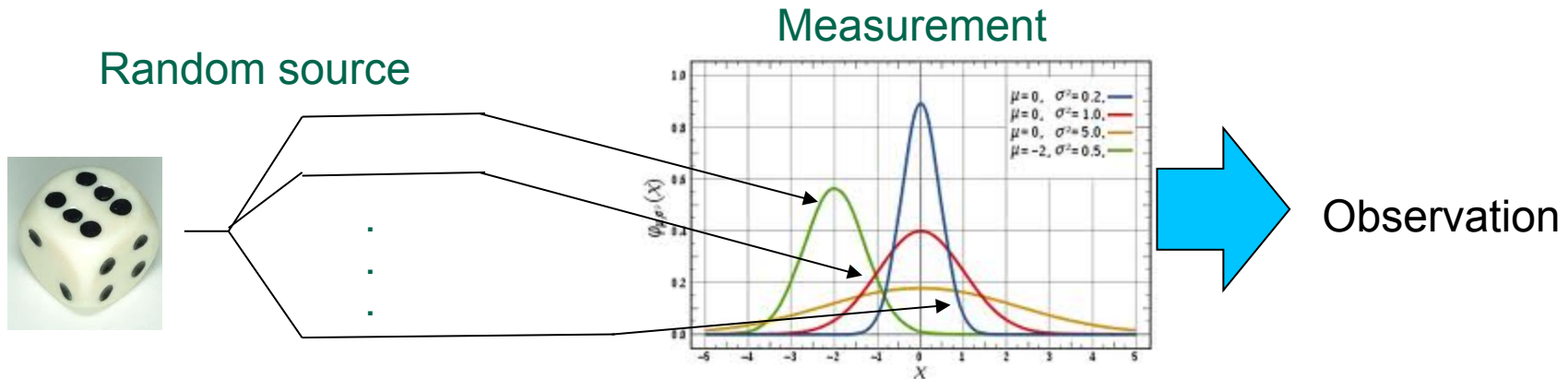
Based on the observed measurement,
and knowing the probability of the coin toss
and the densities of the measurement process,

decide the outcome of the coin toss.

In practice, we do not know
the exact densities.
That is why we need to do
learning, supervised or
unsupervised.

K-class problem

A K -class problem with a K -sided dice:



Pattern Recognition Problem

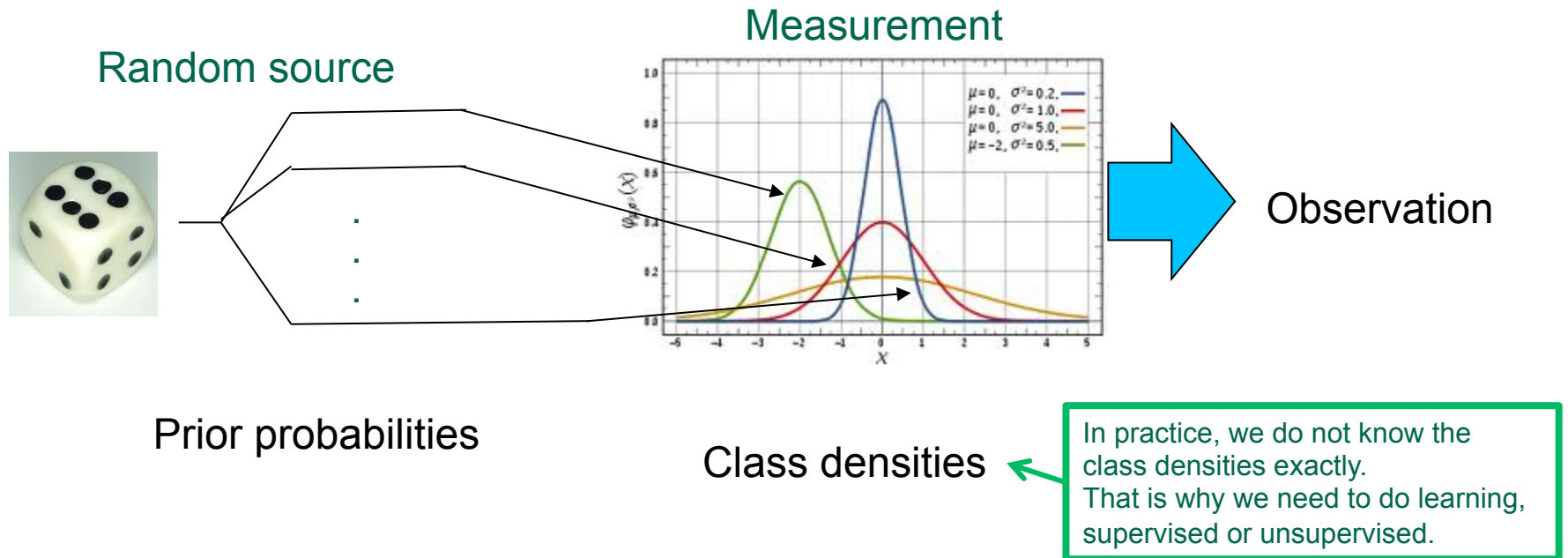
Based on the observed measurement,
and knowing the probability of the dice toss
and the densities of the measurement process,

decide the outcome of the dice toss.

In practice, we do not know
the exact densities.
That is why we need to do
learning, supervised or
unsupervised.

Optimum Decision Rule

- We want to derive an optimum decision rule
- If we know the prior probabilities and the class densities, the optimum decision rule is known



Optimal Decision Rule

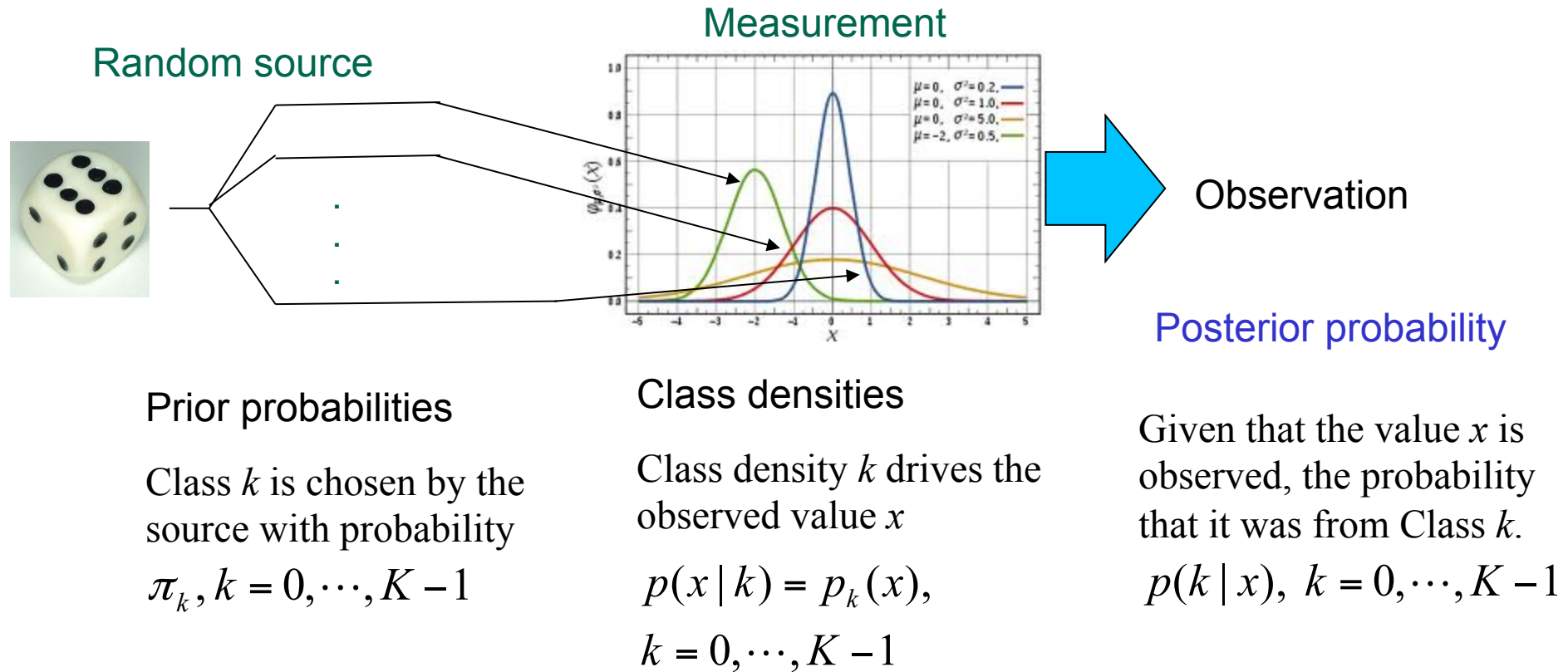
- What are we optimizing?
- Candidates
 - Maximize probability of correct decision for a particular class
 - Maximize probability of correct decision for all classes

Optimal Decision Rule

A Theory:

- Define a loss function for an incorrect decision for each class
- Define the total risk as the expected value of the total loss function (over all classes)
- The optimal decision rule that minimizes the total risk

Probabilities



Probabilities

Prior probabilities

Class k is chosen by the source with probability

$$\pi_k, k = 0, \dots, K - 1$$

Class densities

Class density k drives the observed value x

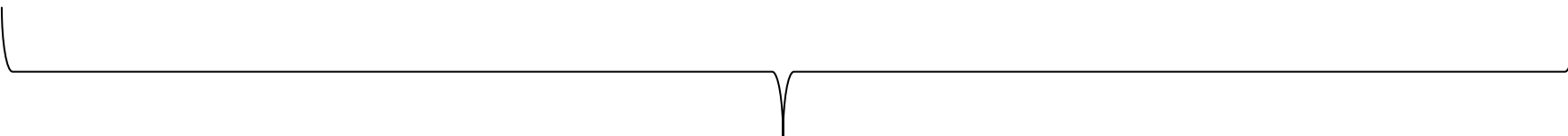
$$p(x | k) = p_k(x),$$

$$k = 0, \dots, K - 1$$

Posterior probability

Given that the value x is observed, the probability that it was from Class k .

$$p(k | x), k = 0, \dots, K - 1$$

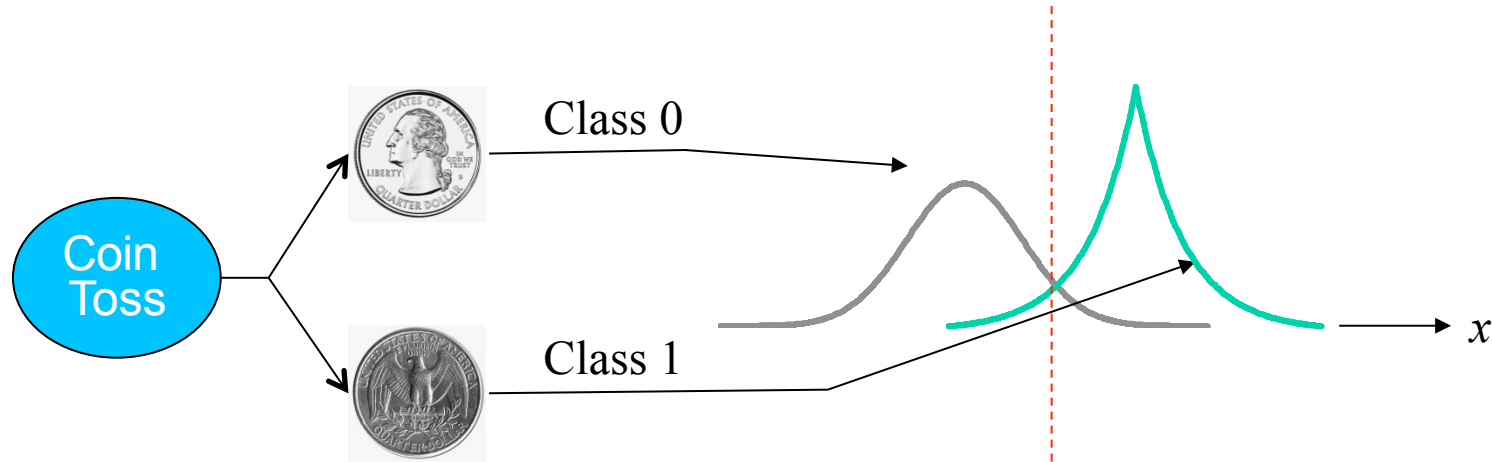

$$p(k | x) = \frac{p(k, x)}{p(x)} = \frac{p(x | k)\pi_k}{p(x)}$$

Maximum a posteriori rule says:

Choose Class k if $p(k | x) > p(k' | x)$ for all $k' \neq k$;

i.e., choose Class k if $p(x | k)\pi_k > p(x | k')\pi_{k'}$ for all $k' \neq k$;

Optimal Decision Rule



Bayesian Rule

Decide Class 0 if x is on this side \leftarrow \rightarrow Decide Class 1 if x is on this side

Optimal Decision Rule

A Theory:

- Define a loss function for an incorrect decision for each class
- Define the total risk as the expected value of the total loss function (over all classes)
- The optimal decision rule is the one that minimizes the total risk

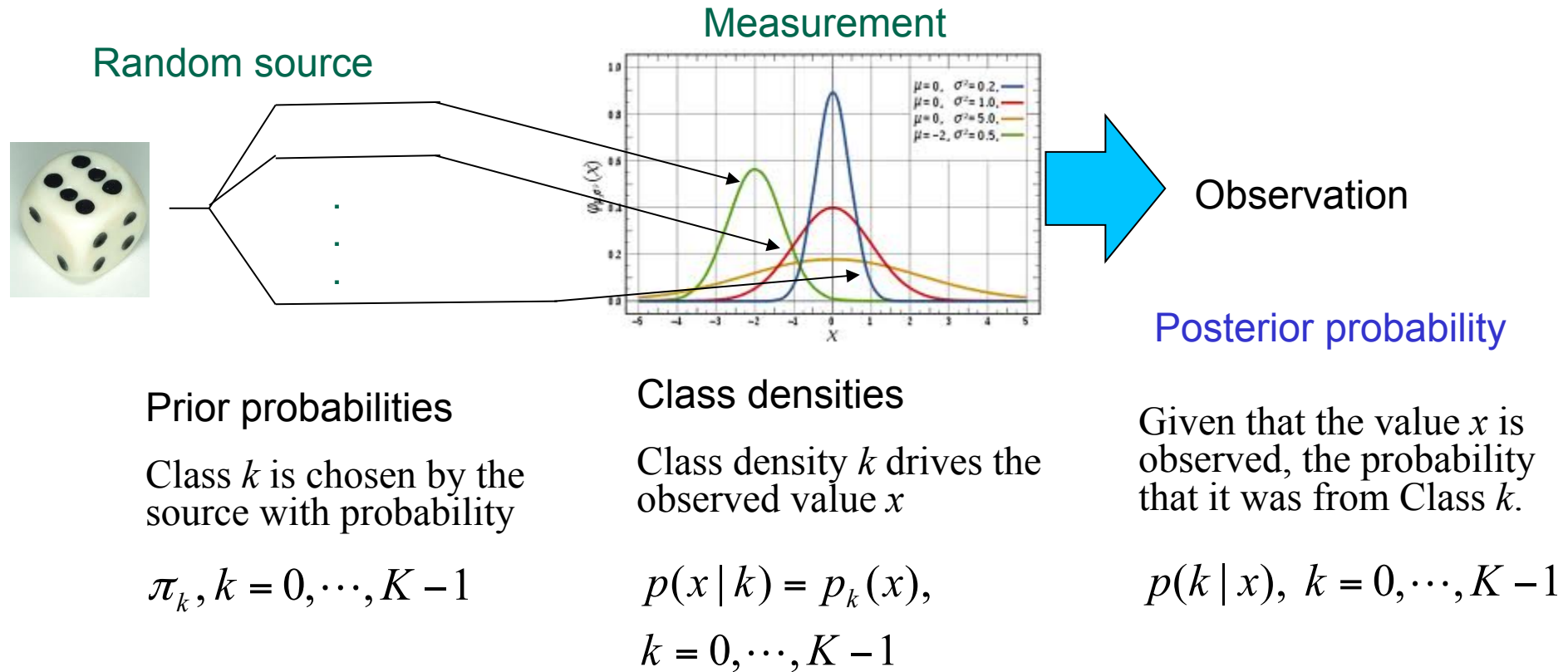
Optimal Decision Rule

A Theory:

- Define a loss function for an incorrect decision for each class
- Define the total risk as the expected value of the total loss function (over all classes)
- The optimal decision rule is the one that minimizes the total risk

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

Probabilities



Probabilities

Prior probabilities

Class k is chosen by the source with probability

$$\pi_k, k = 0, \dots, K - 1$$

Class densities

Class density k drives the observed value x

$$p(x | k) = p_k(x),$$

$$k = 0, \dots, K - 1$$

Posterior probability

Given that the value x is observed, the probability that it was from Class k .

$$p(k | x), k = 0, \dots, K - 1$$

$$p(k | x) = \frac{p(k, x)}{p(x)} = \frac{p(x | k)\pi_k}{p(x)}$$

Maximum a posteriori rule says:

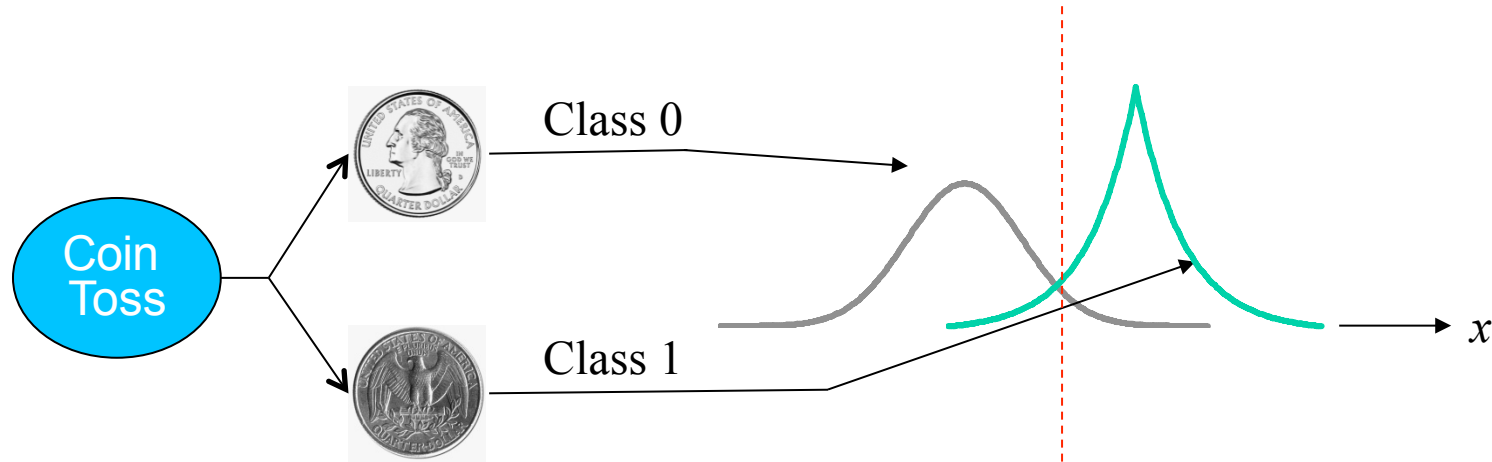
Choose Class k if $p(k | x) > p(k' | x)$ for all $k' \neq k$;

i.e., choose Class k if

$$p(x | k)\pi_k > p(x | k')\pi_{k'} \text{ for all } k' \neq k;$$

Optimal Decision Rule

Simplest case: Two class; single feature



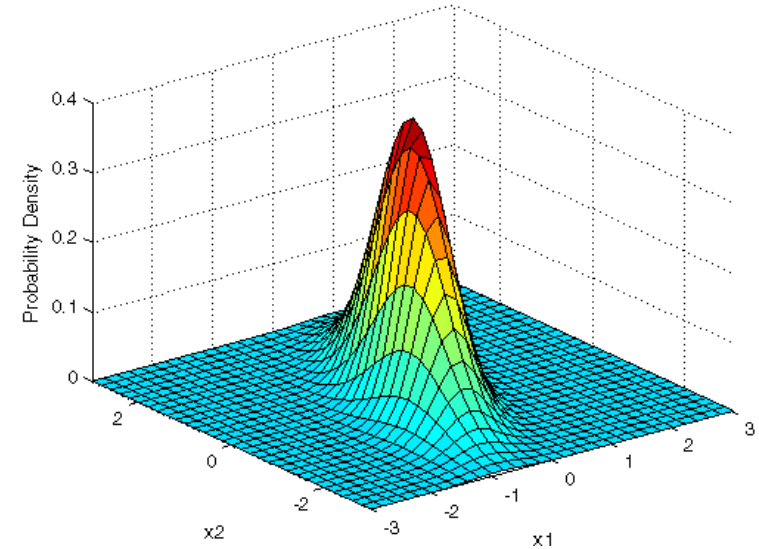
Bayesian Rule

Decide Class 0 if x is on this side \leftarrow \rightarrow Decide Class 1 if x is on this side

Gaussian Density Function

p features organized as a feature vector

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



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 $p(k | \mathbf{x})$.

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

When the class densities are Gaussian, we want to maximize

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

Gaussian Class Densities That Differ Only In Their Means

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

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

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

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

which is the same as minimizing

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

Gaussian Class Densities

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

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

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

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

which is the same as minimizing

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

Gaussian Class Densities for $K=2$

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

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

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

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

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

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

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

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

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

Adding a term $(\mathbf{m}_1^T \Sigma^{-1} \mathbf{m}_0 - \mathbf{m}_0^T \Sigma^{-1} \mathbf{m}_1)$ to $(\mathbf{m}_1^T \Sigma^{-1} \mathbf{m}_1 - \mathbf{m}_0^T \Sigma^{-1} \mathbf{m}_0)$,
we have $(\mathbf{m}_1^T \Sigma^{-1} \mathbf{m}_1 + \mathbf{m}_1^T \Sigma^{-1} \mathbf{m}_0 - \mathbf{m}_0^T \Sigma^{-1} \mathbf{m}_1 - \mathbf{m}_0^T \Sigma^{-1} \mathbf{m}_0)$,

which simplifies to

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

The rule then becomes choose $l = 0$ if

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

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

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

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

and $\mathbf{w} = \boldsymbol{\Sigma}^{-T} (\mathbf{m}_1 - \mathbf{m}_0)$ 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 $g(\mathbf{x}) < 0$.

Geometry of the Linear Classifier

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

where \mathbf{w} is a $(p \times 1)$ weight vector and w_p is the bias.

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

This decision rule is linear because $g(\mathbf{x})$ is a linear combination of the components of \mathbf{x} : $g(\mathbf{x}) = w_0 x_0 + w_1 x_1 + \cdots + w_{p-1} x_{p-1} + w_p$.

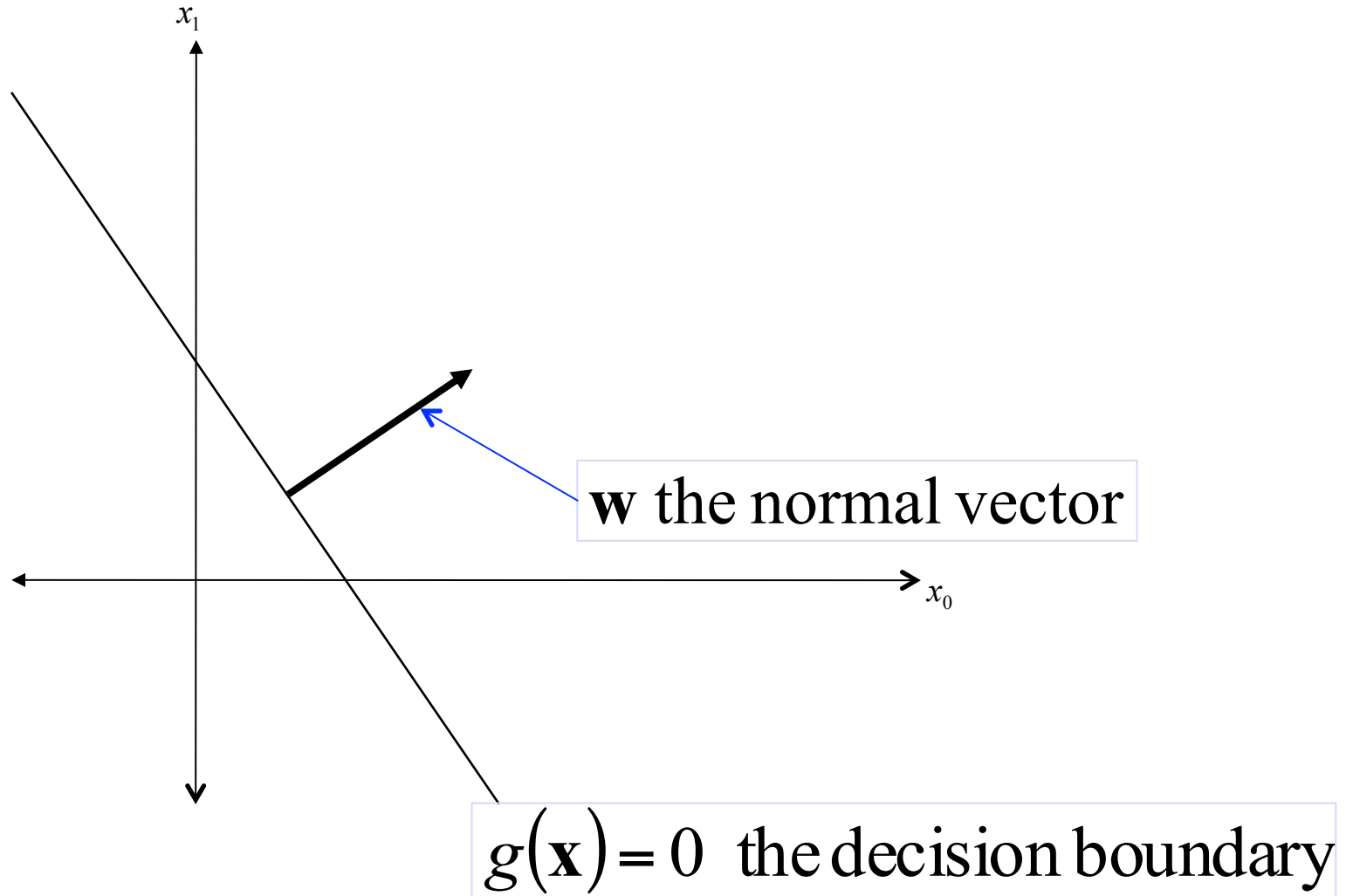
When $p = 2$, $g(\mathbf{x})$ is a line.

When $p = 3$, $g(\mathbf{x})$ is a plane.

When $p > 3$, $g(\mathbf{x})$ is a hyperplane.

\mathbf{w} is the normal vector

Linear Classifier



Distance of a Point to the Decision Boundary

The decision boundary is $g(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + w_p = 0$.

Let \mathbf{x} be an arbitrary point in the feature space and let δ be the distance of \mathbf{x} from the decision boundary. What is δ ?

Let \mathbf{x}_\perp be the point on the decision boundary that is closest to \mathbf{x} ,

so that $\mathbf{x} = \mathbf{x}_\perp + \delta \frac{\mathbf{w}}{\|\mathbf{w}\|}$. Perform the inner product of \mathbf{w} with both sides

of the equation and add the bias term to both inner products. We have

$$\underbrace{\mathbf{w}^T \mathbf{x} + w_p}_{g(\mathbf{x})} = \underbrace{\mathbf{w}^T \mathbf{x}_\perp + w_p}_{g(\mathbf{x}_\perp)} + \delta \frac{\mathbf{w}^T \mathbf{w}}{\|\mathbf{w}\|}, \text{ so that } \delta = \frac{g(\mathbf{x})}{\|\mathbf{w}\|}.$$

$g(\mathbf{x}_\perp) = 0$ because by definition \mathbf{x}_\perp sits on the decision boundary

Distance of a Point to the Decision Boundary

The decision boundary is $g(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + w_p = 0$.

Let \mathbf{x} be an arbitrary point in the feature space.

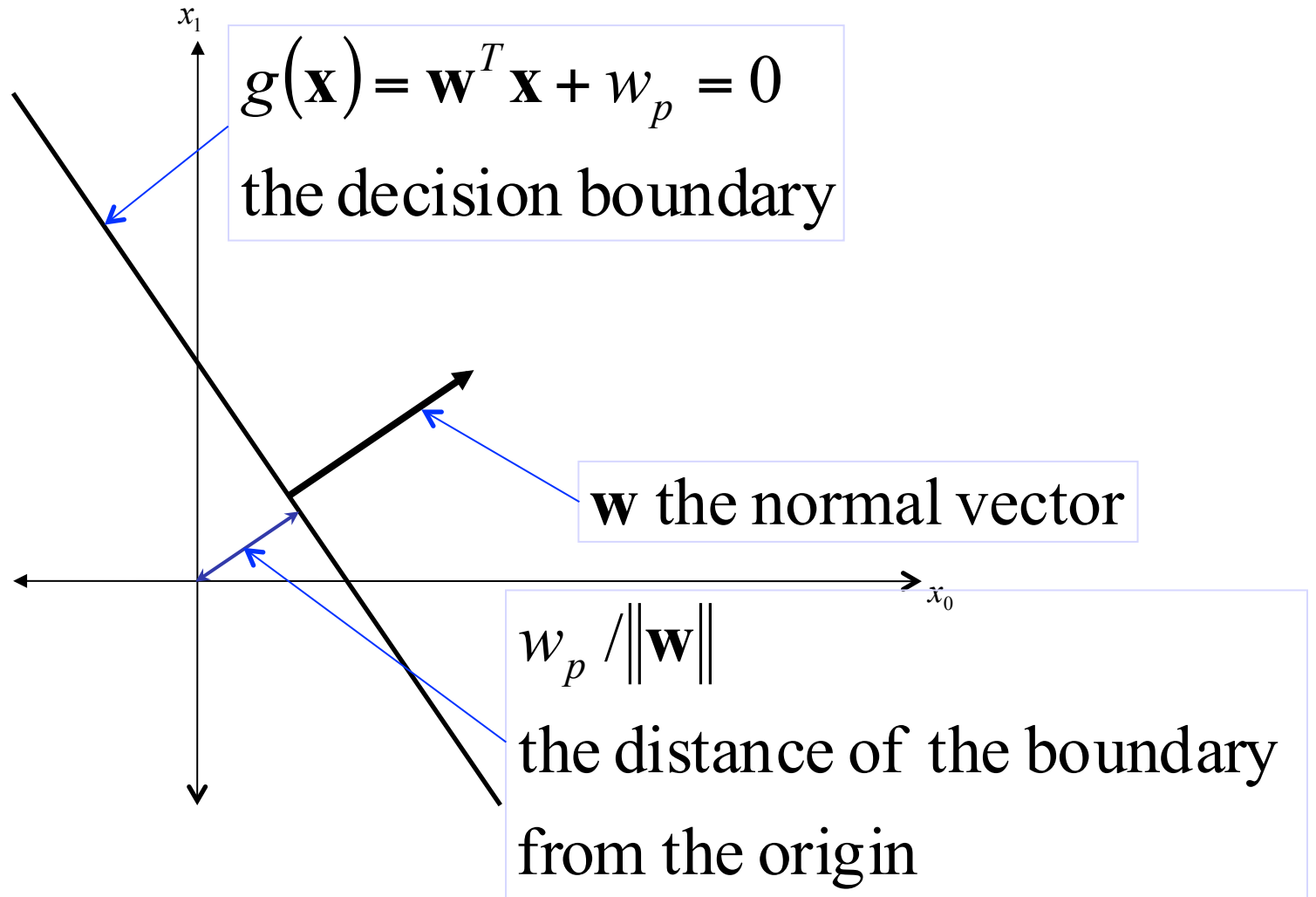
The distance of \mathbf{x} from the decision boundary is $\delta(\mathbf{x}) = \frac{g(\mathbf{x})}{\|\mathbf{w}\|}$.

What is the distance of the decision boundary from the origin?

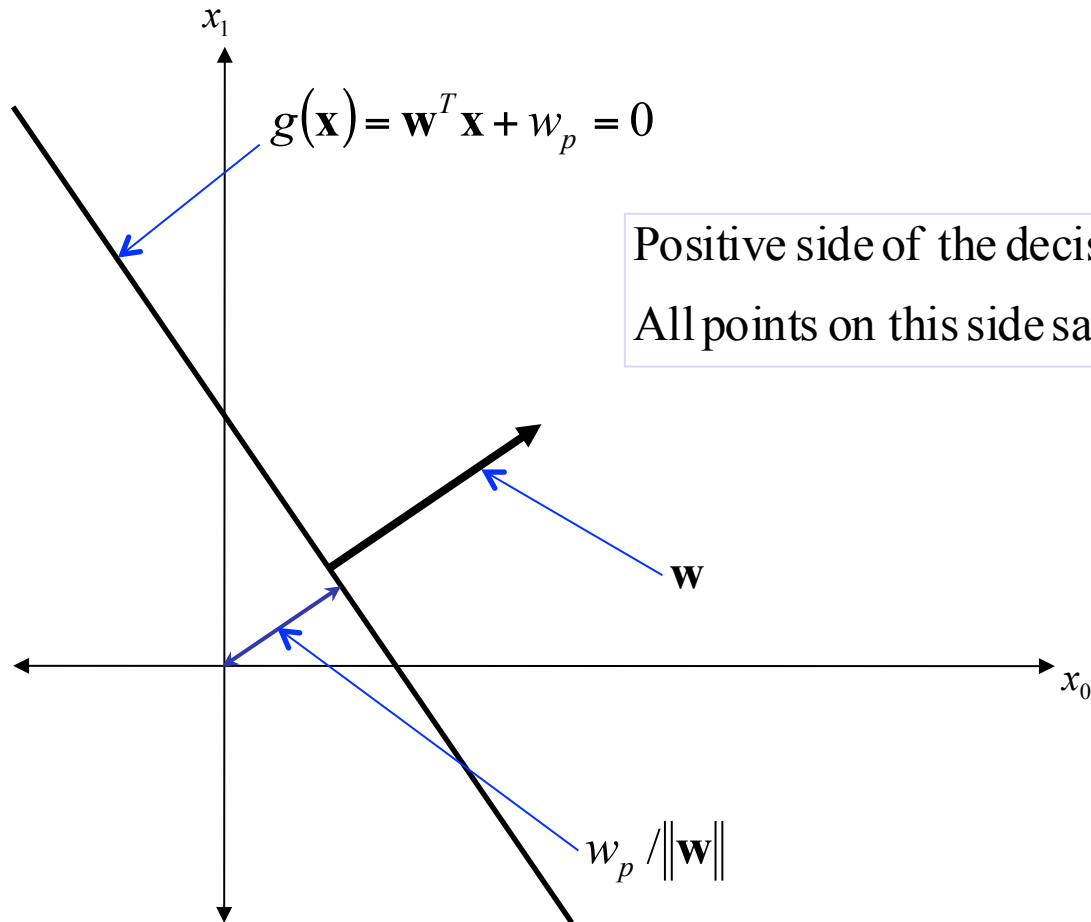
Let $\mathbf{x} = \mathbf{0}$, the origin. Then

$$\delta(\mathbf{0}) = \frac{g(\mathbf{0})}{\|\mathbf{w}\|} = \frac{\mathbf{w}^T \mathbf{0} + w_p}{\|\mathbf{w}\|} = \frac{w_p}{\|\mathbf{w}\|}.$$

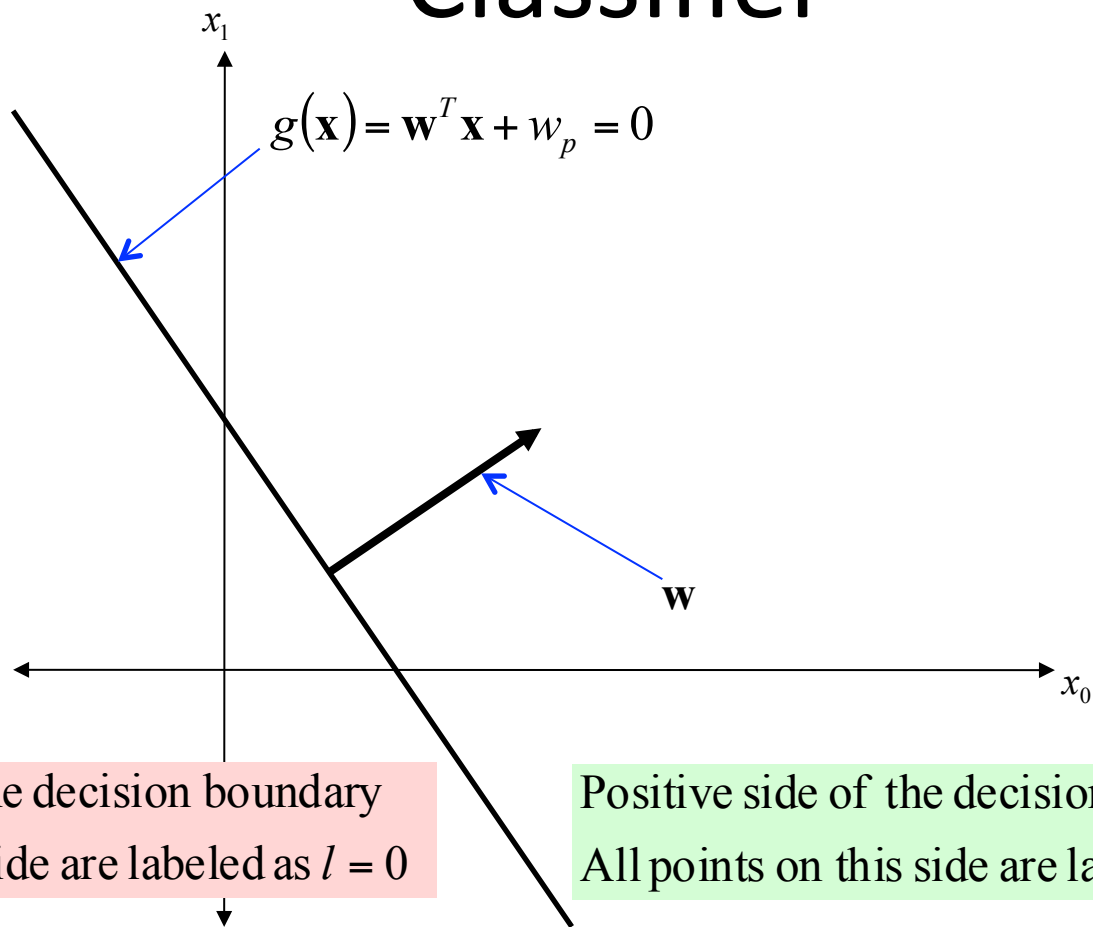
Linear Classifier



Linear Classifier



Decision Regions of a Linear Classifier



Negative side of the decision boundary
All points on this side are labeled as $l = 0$

Positive side of the decision boundary
All points on this side are labeled as $l = 1$

Linear Classifier

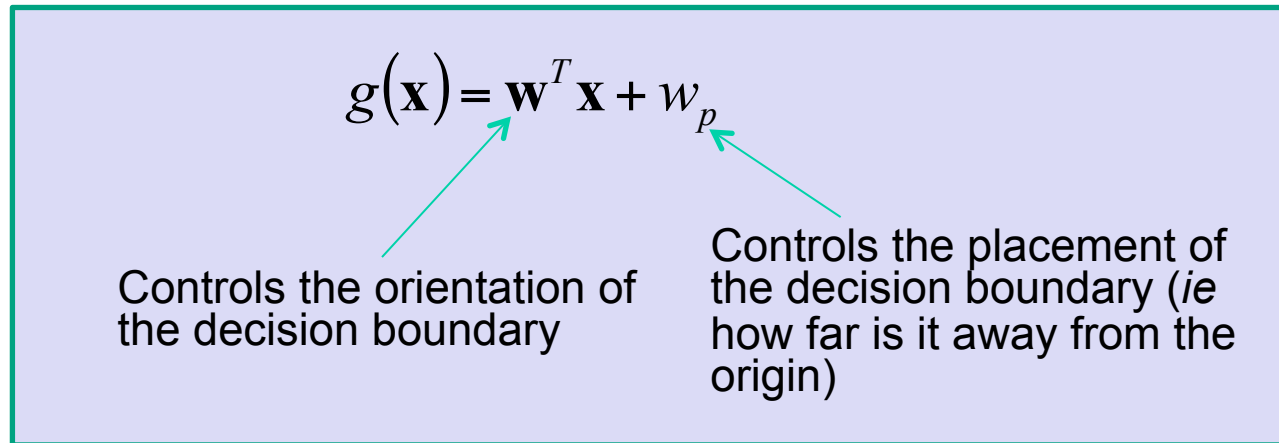
Suppose $K = 2$.

Let the class densities be Gaussian that differ only in their means \mathbf{m}_1 and \mathbf{m}_0 ; let the covariance matrix be Σ .

Let $\mathbf{w} = \Sigma^{-T}(\mathbf{m}_1 - \mathbf{m}_0)$ and $w_p = -(\mathbf{m}_1 - \mathbf{m}_0)^T \Sigma^{-1} \frac{(\mathbf{m}_1 + \mathbf{m}_0)}{2} + \log \frac{\pi_0}{\pi_1}$.

Define $g(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + w_p$. The hyperplane $g(\mathbf{x}) = 0$ is the decision boundary.

The optimal decision rule is to choose $l = 0$ if $g(\mathbf{x}) < 0$ and $l = 1$ if $g(\mathbf{x}) > 0$.



Classification when the priors are the same

Choose $l = k$ to minimize

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

When all classes have the same prior probability, then the decision is choose $l = k$ to minimize

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

Let $d_M^2(\mathbf{x}, \mathbf{m}_k) = (\mathbf{x} - \mathbf{m}_k)^T \mathbf{\Sigma}^{-1}(\mathbf{x} - \mathbf{m}_k)$; the term $d_M(\mathbf{x}, \mathbf{m}_k)$ is called the Mahalanobis distance between \mathbf{x} and \mathbf{m}_k .

Classification when the priors are the same

Choose $l = k$ to minimize

$$(\mathbf{x} - \mathbf{m}_k)^T \Sigma^{-1} (\mathbf{x} - \mathbf{m}_k),$$

the squared Mahalanobis distance between \mathbf{x} and \mathbf{m}_k .

Mahalanobis Distance

The Mahalanobis distance $d_M(\mathbf{x}, \mathbf{y})$ between \mathbf{x} and \mathbf{y} is given by

$$d_M^2(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \mathbf{y}).$$

When the components are pairwise uncorrelated so that

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_0^2 & & \\ & \ddots & \\ & & \sigma_{p-1}^2 \end{bmatrix}, \text{ the Mahalanobis distance is}$$

$$d_M^2(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \mathbf{y}) = (\mathbf{x} - \mathbf{y})^T \begin{bmatrix} \frac{1}{\sigma_0^2} & & \\ & \ddots & \\ & & \frac{1}{\sigma_{p-1}^2} \end{bmatrix} (\mathbf{x} - \mathbf{y}) = \sum_{i=0}^{p-1} \frac{(x_i - y_i)^2}{\sigma_i^2}.$$

Mahalanobis Distance

The Mahalanobis distance $d_M(\mathbf{x}, \mathbf{y})$ between \mathbf{x} and \mathbf{y} is given by

$$d_M^2(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \mathbf{y}).$$

When the components are pairwise uncorrelated and each has unit variance, so that $\mathbf{\Sigma} = \mathbf{I}$,

the Mahalanobis distance is the same as the Euclidean distance

$$d_M^2(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^T \mathbf{I}^{-1} (\mathbf{x} - \mathbf{y}) = (\mathbf{x} - \mathbf{y})^T (\mathbf{x} - \mathbf{y}) = \sum_{i=0}^{p-1} (x_i - y_i)^2.$$

Covariance Matrix

Let the covariance matrix be $\Sigma = E[(\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^T]$.

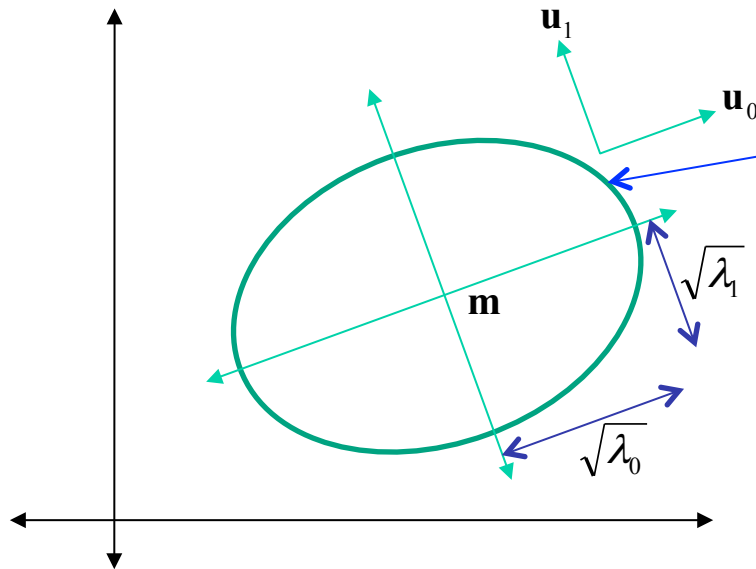
The ij th element is defined as $\sigma_{ij} = E[(x_i - m_i)(x_j - m_j)]$

Since $\sigma_{ij} = \sigma_{ji}$, the covariance matrix Σ is symmetric.

Covariance Matrix in a Gaussian Density

The density of a Gaussian vector with mean \mathbf{m} and covariance matrix Σ is

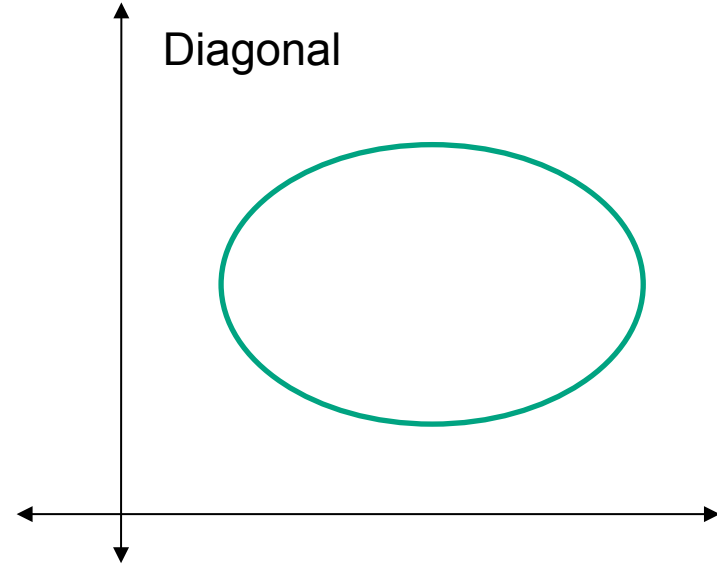
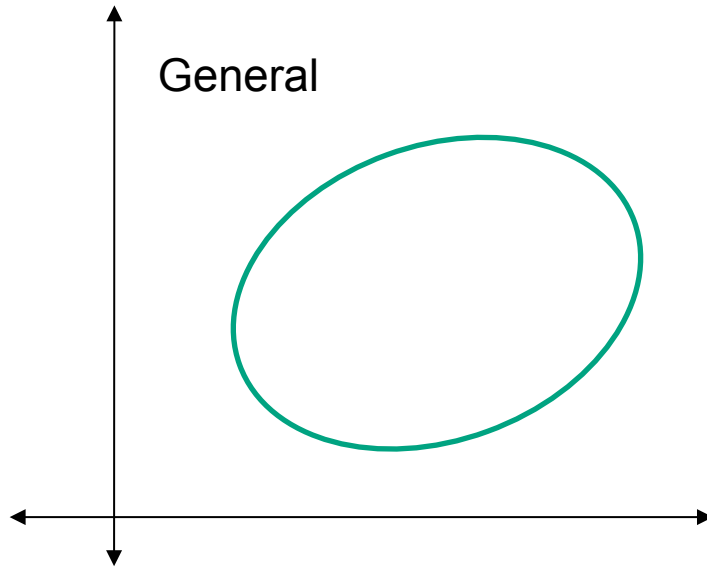
$$p(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}} \frac{1}{|\Sigma|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^T \Sigma^{-1}(\mathbf{x}-\mathbf{m})}.$$



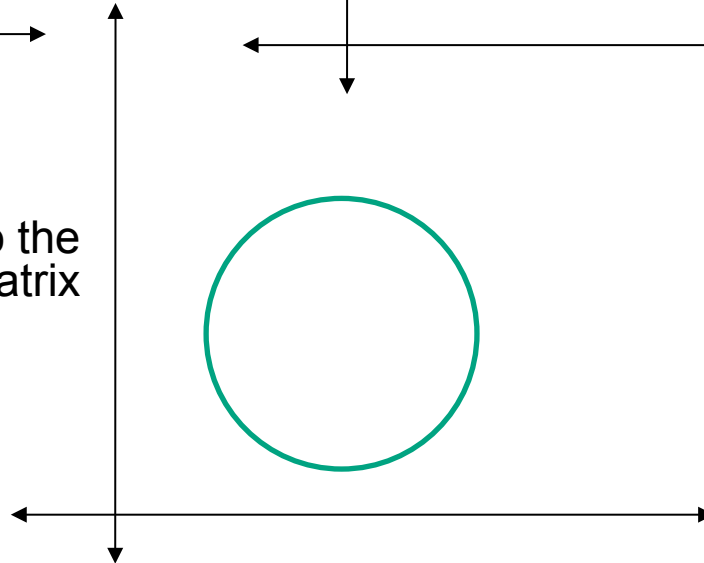
The surface of constant density on which $p(\mathbf{x})$ is $e^{-1/2}$ of $p(\mathbf{m})$

The covariance matrix Σ has eigenvectors \mathbf{u}_0 and \mathbf{u}_1 and corresponding eigenvalues λ_0 and λ_1 .

Covariance Matrices



Proportional to the
identity matrix



Examples

- Class densities are Gaussian that differ only in their means

Two classes

Class means:

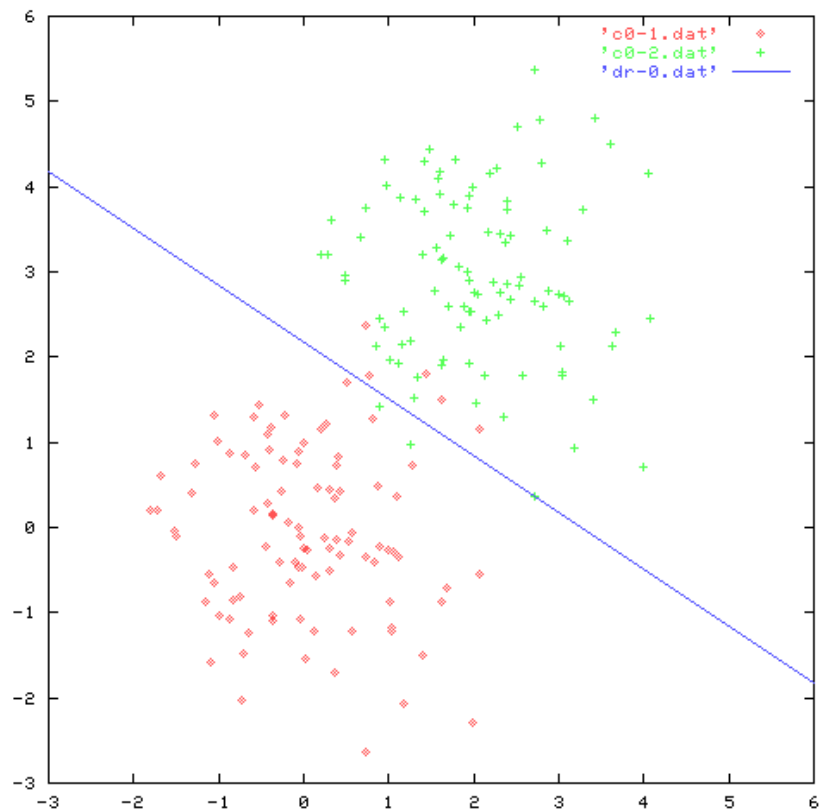
$$\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\Sigma^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Decision boundary:

$$2x_0 + 3x_1 = 6.5$$

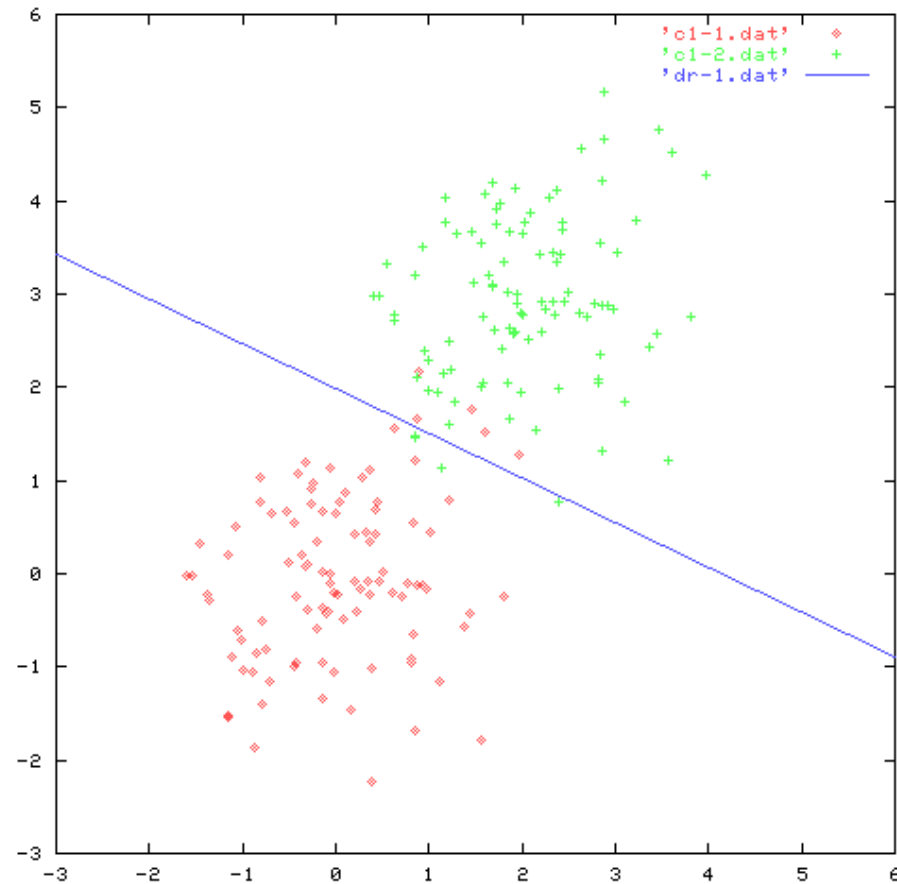


$$\Sigma = \begin{bmatrix} 0.82 & 0.20 \\ 0.20 & 0.79 \end{bmatrix}$$

$$\Sigma^{-1} = \begin{bmatrix} 1.30 & -0.32 \\ -0.32 & 1.35 \end{bmatrix}$$

Decision boundary:

$$1.63x_0 + 3.40x_1 = 6.73$$

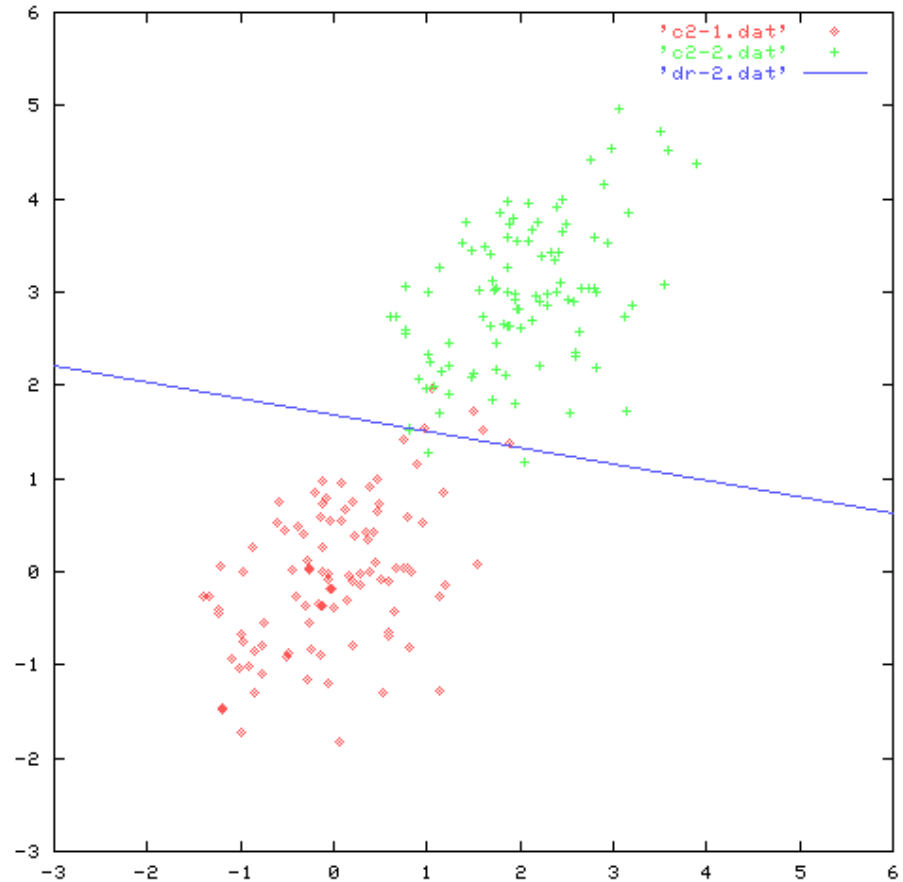


$$\Sigma = \begin{bmatrix} 0.68 & 0.34 \\ 0.34 & 0.64 \end{bmatrix}$$

$$\Sigma^{-1} = \begin{bmatrix} 2.03 & -1.10 \\ -1.10 & 2.17 \end{bmatrix}$$

Decision boundary:

$$0.76x_0 + 4.31x_1 = 7.23$$

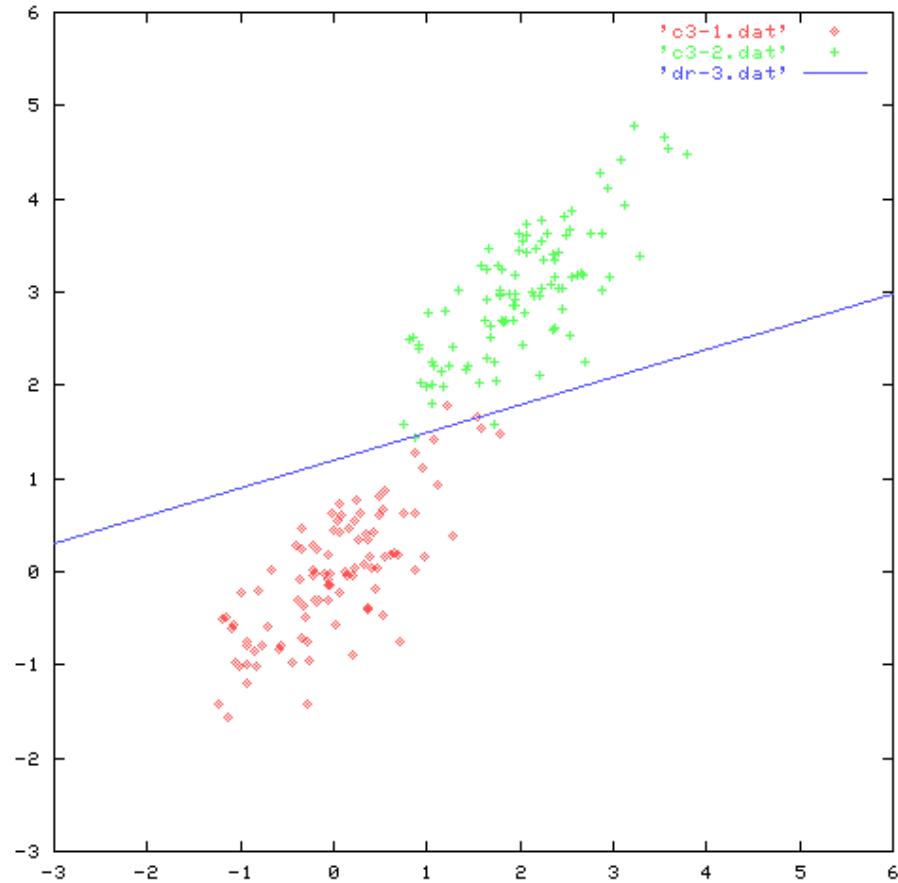


$$\Sigma = \begin{bmatrix} 0.58 & 0.44 \\ 0.44 & 0.54 \end{bmatrix}$$

$$\Sigma^{-1} = \begin{bmatrix} 4.66 & -3.84 \\ -3.84 & 5.02 \end{bmatrix}$$

Decision boundary:

$$-2.19x_0 + 7.37x_1 = 8.86$$



$$\Sigma = \begin{bmatrix} 0.52 & 0.50 \\ 0.50 & 0.50 \end{bmatrix}$$

$$\Sigma^{-1} = \begin{bmatrix} 34.78 & -34.44 \\ -34.44 & 36.11 \end{bmatrix}$$

Decision boundary:

$$-33.78x_0 + 39.44x_1 = 25.39$$

