BAYESIAN DECISION THEORY

DATA/MSML 603: Principles of Machine Learning

Dr. Kemal Davaslioglu

Agenda

- Bayesian Decision Theory
  - Bayes' Theorem & Bayesian Probability
  - Risk and Bayesian Decision Rule
  - Signal Detection & Operating Characteristics
  - Numerical Examples
  - Coding examples

Announcements

Announcements

- HW#1 will be posted (due next Friday)
  - Any assignment submitted electronically, please use this naming convention: <assignment number> <your name> msml603sec2-
  - E.g.,: HW1_JohnDoe_msml603sec2.pdf
- Clarifications on the Term Project

Term Project

- ➢ Individual projects.
- ➢ Proposals will be due around Week #4-#5.
- ➢ Report + full code needs to be delivered.
  - Report needs to be 6-8 pages.
  - You can include an appendix for additional figures/tables.
  - Any coding language or framework is acceptable.
- ➢ Final submission before Finals Week.
- ➢ Rubric will be posted soon.
- ➢ Resources to find datasets (next slide).

Term Project

Resources where you can search for datasets:

- https://datasetsearch.research.google.com/
- https://www.kaggle.com/datasets
- https://earthdata.nasa.gov/
- https://archive.ics.uci.edu/ml/index.php
- https://registry.opendata.aws/
- https://crime-data-explorer.fr.cloud.gov/downloads-and-docs
- https://data.world/
- http://opendata.cern.ch/
- https://lionbridge.ai/datasets/
- https://datahub.io/search
- https://github.com/awesomedata/awesome-public-datasets
- https://www.visualdata.io/discovery

Bayesian Decisions

Bayesian Decision Theory

The Basic Idea

- To minimize errors, choose the least risky class, i.e., the class for which the *expected loss* is smallest

Assumptions

- Problem posed in probabilistic terms, and all relevant probabilities are known.

Probability Mass vs. Probability Density Functions

Probability Mass Function, P(x)

- Probability for values of discrete random variable *x*. Each value has its own associated probability

$$P(x) \geq 0 \text{ and } \sum_{x \in \chi} P(x) = 1\,,\text{where } \chi = \{v_1, \cdots, v_m\}.$$

Probability Density, p(x)

- Probability for values of continuous random variable *x.*
- Probability returned is for an *interval* within which the value lies (intervals defined by some unit distance)

$$p(x) \geq 0, P(x \in [a, b]) = \int_a^b p(x)\mathrm{d}x, \text{ and } \int_{-\infty}^{\infty} p(x)\mathrm{d}x = 1.$$

Prior Probability

Definition ($P(w)$)

- The likelihood of a value for a random variable representing the *state of nature (true class for the current input)*, in the absence of other information \
- Informally, "what percentage of the time state X occurs"

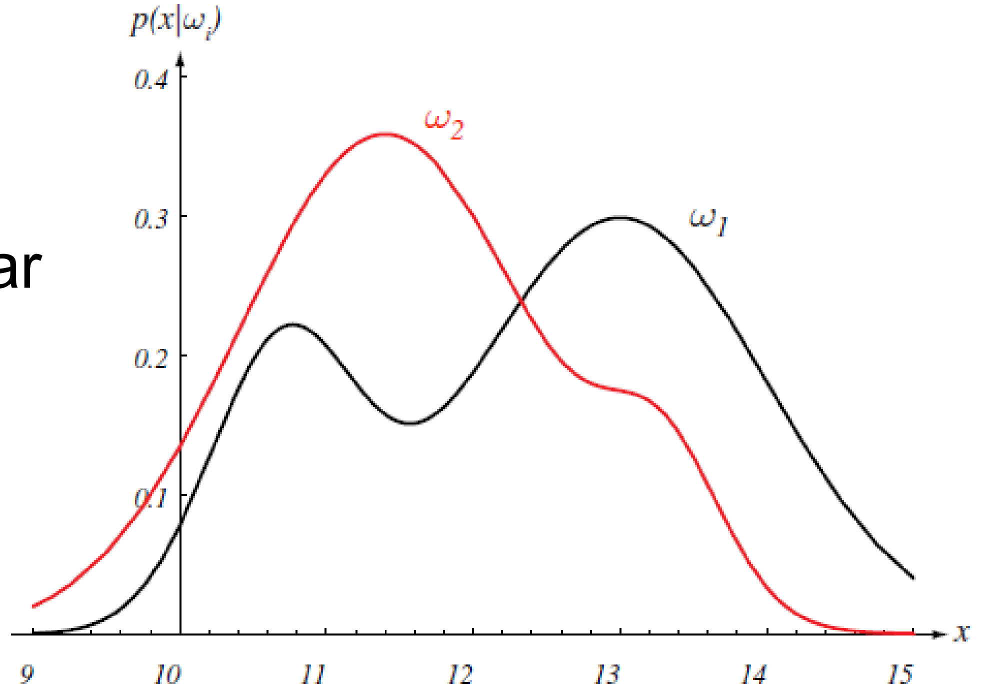
Example

- The prior probability that an instance taken from two classes is in the absence of any features is
- E.g., P(cat) = 0.7, P(dog) = 0.3.

Class-Conditional Probability Density Function (for Continuous Features)

Definition: Class Conditional Probability, $p(x|w)$

- The probability of a value for continuous random variable x, given a state of nature in w
- For each value of x, we have a different class conditional pdf for each class in w
  - Probability density of measuring a particular feature value x given the pattern is in category w_i.
  - Density functions are normalized
  - Area under the curve is 1.



Bayes Formula

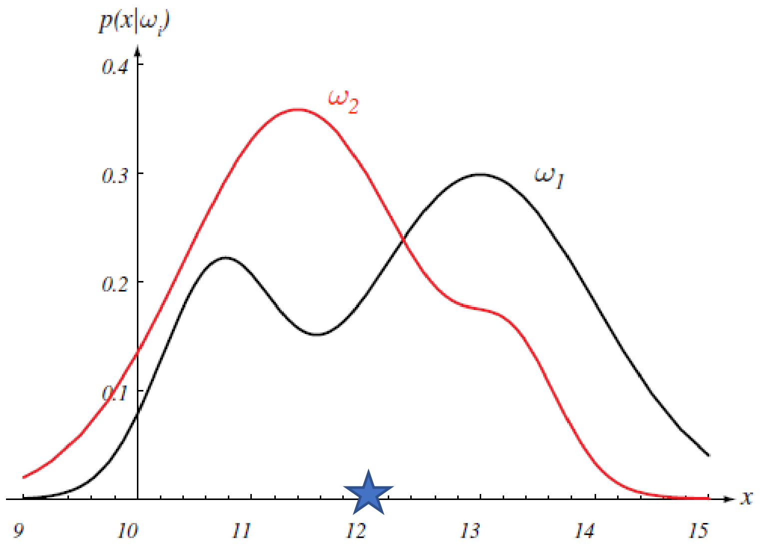
Purpose

- Convert class prior and class-conditional densities to a posterior probability for a class
- The probability of a class given the input features (*post-observation*)

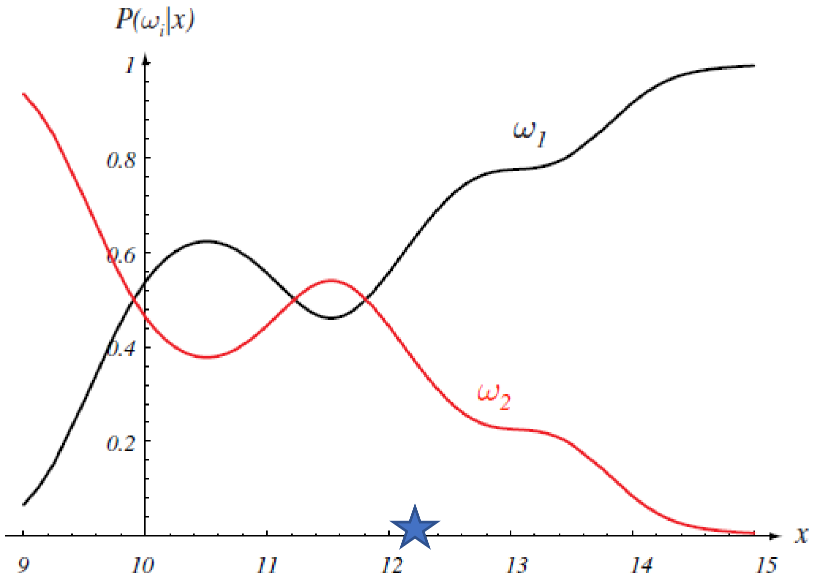
$$P(w_i|x) = \frac{p(x|w_i)P(w_i)}{p(x)} \quad \Rightarrow \quad \text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

Bayes Formula

- Converts **class prior** $p(w_i)$ and **class-conditional densities** $p(x|w_i)$ to a posterior probability for a class $p(w_i|x)$
- The probability of a class given the input features (*post-observation*)



$$p(w_1) = \frac{2}{3}$$
$$p(w_2) = \frac{1}{3}$$



Choosing the Most Likely Class

What happens if we do the following?

- Decide w_1 if $P(w_1|x) > P(w_2|x)$, otherwise decide w_2.
- We minimize the average probability of error.
- Consider the two-class case from previous slide:

$$P(error|x) = \begin{cases} P(w_1|x) \text{ if we choose } w_2 \\ P(w_2|x) \text{ if we choose } w_1 \end{cases}$$

- Average error can be formulated as

$$P(error) = \int_{-\infty}^{\infty} P(error|x)p(x)dx$$

Decision Functions and Overall Risk

Decision Function

- $\alpha(x)$: Takes on the value of exactly one action for each input vector x

Overall Risk

- The expected (average) loss associated with a decision rule

$$R = \int R(\alpha_i|x)p(x)dx$$

Bayes Decision Rule

Main Idea

- Minimize the overall risk, by choosing the action with the least conditional risk for input vector x

Bayes Risk (R*)

- The resulting overall risk produced using this procedure.
- This is the best performance that can be achieved given available information.

Bayes Decision Rule: Two Category Case

Bayes Decision Rule

- For each input, select class with least conditional risk, i.e. choose class one if:

$$R(\alpha_1|x) \leq R(\alpha_2|x)$$

- $\lambda_{ij} = \lambda(\alpha_i|w_j)$ ⇒ Select class i when sample belongs to class j
- $R(\alpha_1|x) = \lambda_{11}P(w_1|x) + \lambda_{12}P(w_2|x)$
- $R(\alpha_2|x) = \lambda_{21}P(w_1|x) + \lambda_{22}P(w_2|x)$

Alternate Equivalent Expressions of Bayes Decision Rule (“Choose Class One If...”)

Posterior Class Probabilities

$$(\lambda_{21} - \lambda_{11})P(w_1|x) > (\lambda_{12} - \lambda_{22})P(w_2|x)$$

Class Priors and Conditional Densities

- Produced by applying Bayes Formula to the above, multiplying both sides by p(x)

Likelihood ratio

$$\frac{p(x|w_1)}{p(x|w_2)} = \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \cdot \frac{P(w_2)}{P(w_1)}$$

The Zero-One Loss

Definition

- Assigns no loss to correct decision and all errors are equally costly

$$\lambda(\alpha_i|w_j) = \begin{cases} 0 & if\ i = j \\ 1 & if\ i \neq j \end{cases}, \text{for } i, j = 1, \cdots, c$$

Conditional Risk for Zero-One Loss

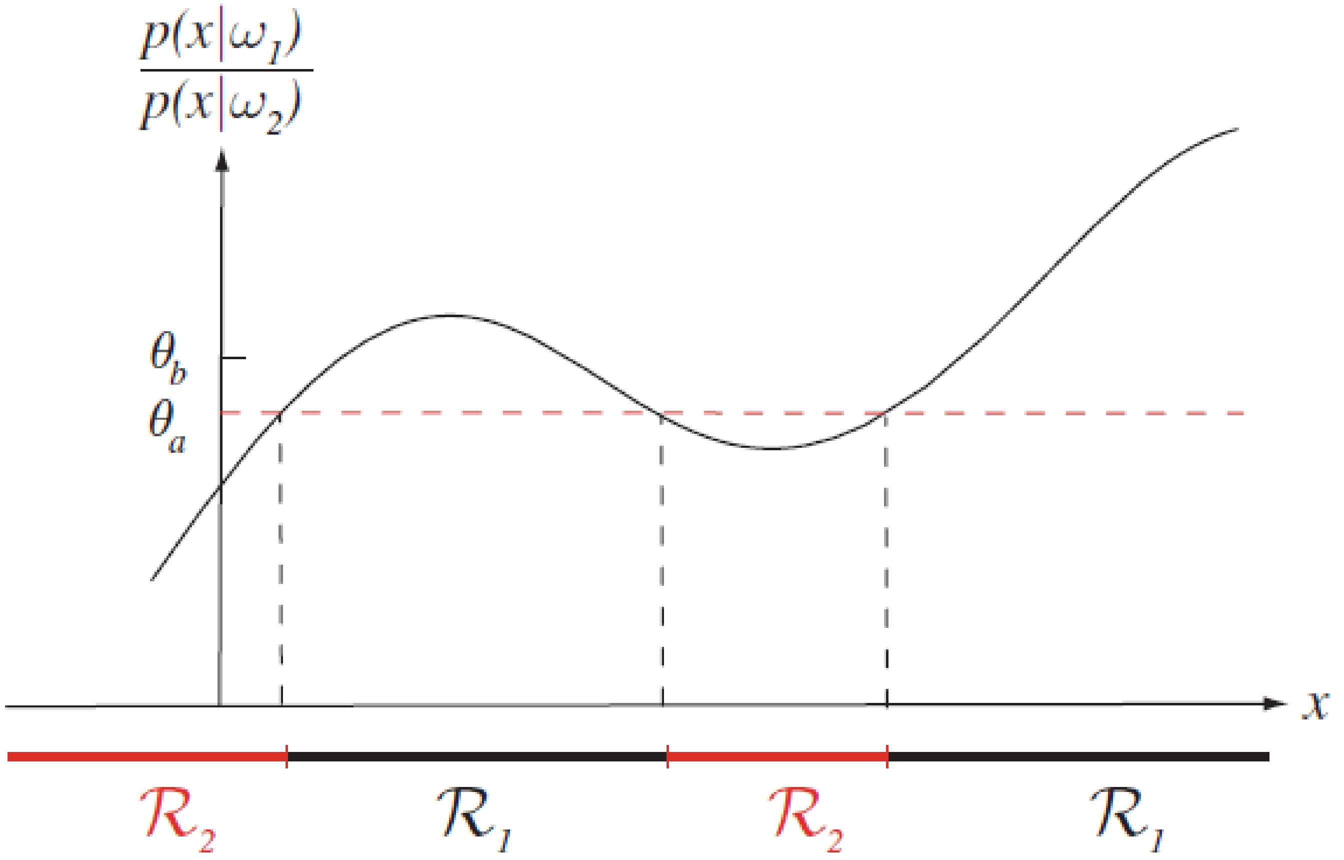
$$R(\alpha_i|x) = \sum_{j=1}^{c} \lambda(\alpha_i|w_j) P(w_j|x) = \sum_{j \neq i} P(w_j|x) = 1 - P(w_i|x)$$

Bayes Decision Rule (min error rate)

- Decide w_i if $P(w_i|x) > P(w_j|x)$ for all $j \neq i$.

Example: Likelihood Ratio

Likelihood ration can range from 0 to infinity.



Bayes Classifiers

Recall the canonical model

Decide class i if $g_i(x) > g_j(x)$ for all $j \neq i$

For Bayesian Classifiers

- General Discriminant definition $g_i(x) = -R(\alpha_i|x)$
- Discriminant definition for zero-one loss $g_i(x) = P(w_i|x)$

Equivalent Discriminants for 0-1 Loss (Min Error Rate)

Trade-off simplicity of understanding vs. computation

- We can express the discriminant definition as

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

$$g_i(x) = p(x|w_i)p(w_i)$$

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

Equivalent Discriminants for 0-1 Loss (Min Error Rate)

For two-categories

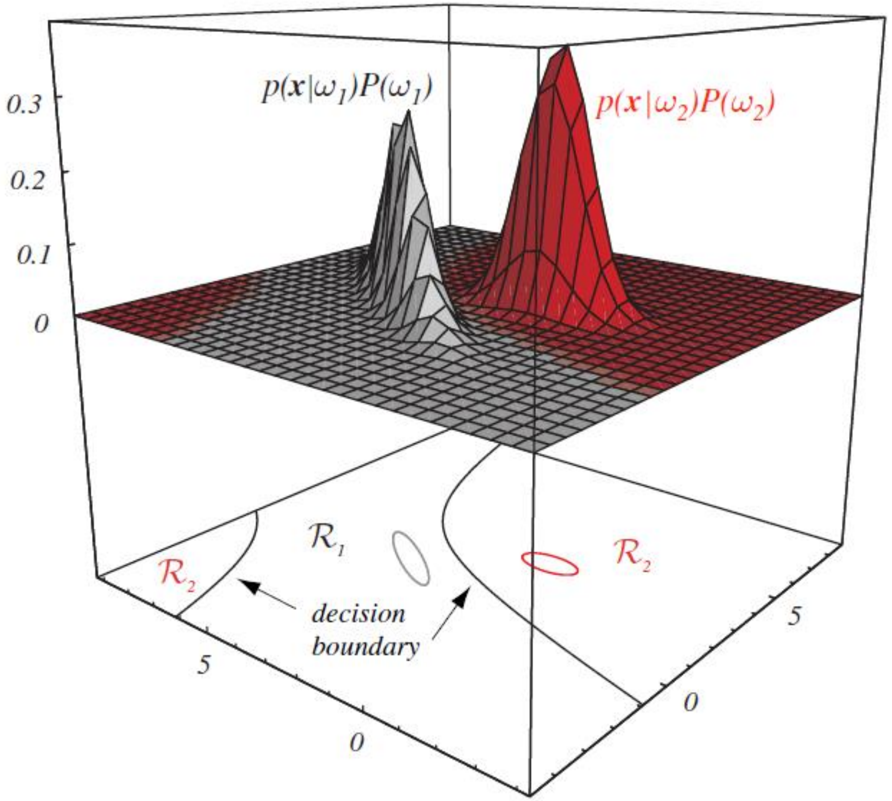
- We can use a single discriminant function, with decision rule:
  - Choose class 1 if the discriminant returns a value > 0

$$g(x) = p(w_1|x) - p(w_2|x)$$

$$g(x) = -\ln\frac{p(x|w_1)}{p(x|w_2)} + \ln\frac{P(w_1)}{P(w_2)}$$

Decision regions for binary classifier

- In this two-dimensional two-category classifier, the probability densities are Gaussian, the decision boundary consists of two hyperbolas, and thus the decision region $\mathcal{R}_2$ is not simply connected.
- The ellipses mark where the density is 1/e times that at the peak of the distribution.



The (Univariate) Normal Distribution

Why are Gaussians so Useful?

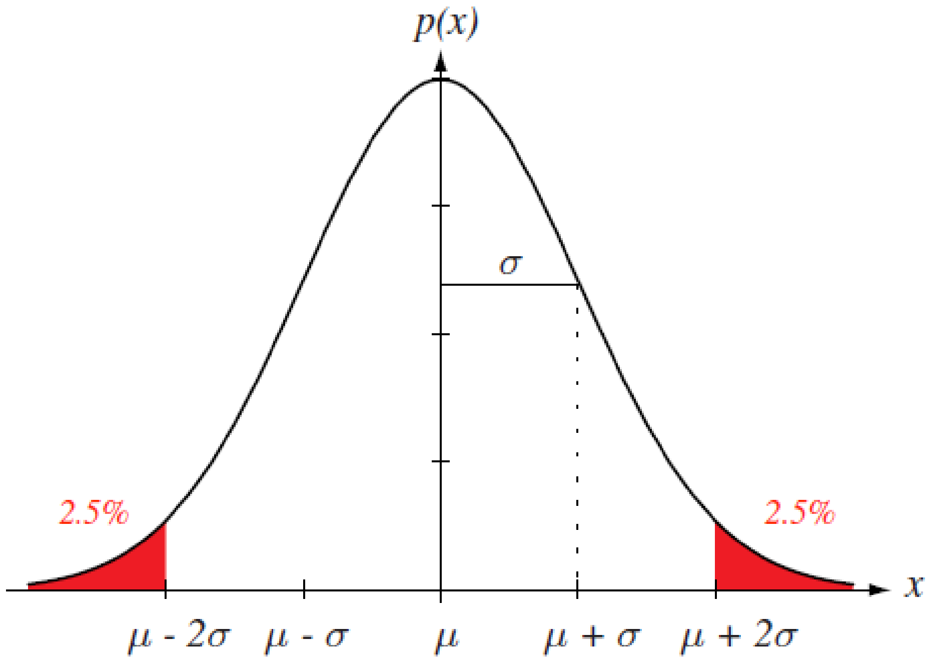
- They represent many probability distributions in nature quite accurately.
- In our case, when patterns can be represented as random variations of an ideal prototype (represented by the mean feature vector).
- Examples: Height or weight of a population

Univariate Normal Distribution

- A univariate normal distribution has roughly 95% of its area in the range

$$|x - \mu| \leq 2\sigma.$$

- The peak of the distribution has value $p(\mu) = 1/\sqrt{2\pi\sigma}$.



Formal Definition

Definition for Univariate Normal

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

Peak of the Distribution (the mean)

Has value: $1/\sqrt{2\pi\sigma}$.

Definition of mean and variance

$$\mu = \int_{-\infty}^{\infty} x \, p(x) \, dx$$

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 p(x) \, dx$$

Multivariate Normal Density

Informal Definition

A normal distribution over two or more variables (d variables/dimensions).

Formal Definition

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^t \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right]$$

$$\mu = \int_{-\infty}^{\infty} \mathbf{x}\, p(\mathbf{x})\, d\mathbf{x}$$

$$\mathbf{\Sigma} = \int (\mathbf{x}-\mu)(\mathbf{x}-\mu)^t p(\mathbf{x}) d\mathbf{x}$$

The Covariance Matrix Σ

Assumption

- Assume the covariance matrix is positive definite, so the determinant of the matrix is always positive.

Matrix Elements

- **Main diagonal:** Variances for each individual variable.
- **Off-diagonal elements:** Covariances of each variable pairing i & j (note: values are repeated, as matrix is symmetric).

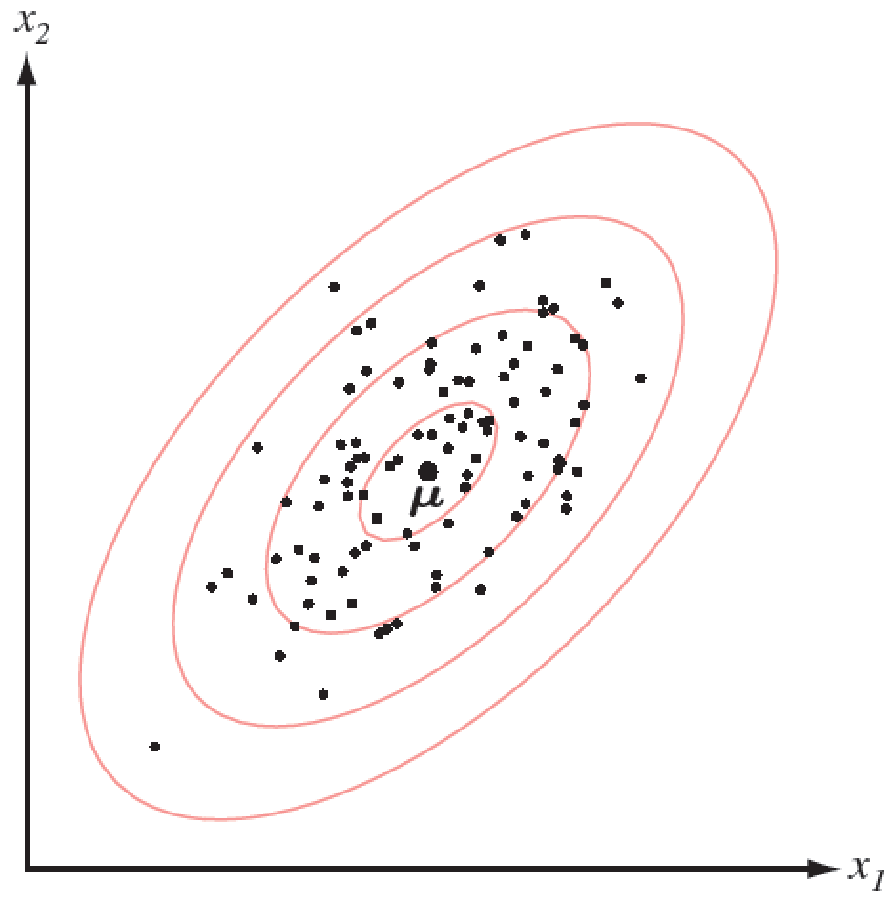
Independence and Correlation

For multivariate normal covariance matrix

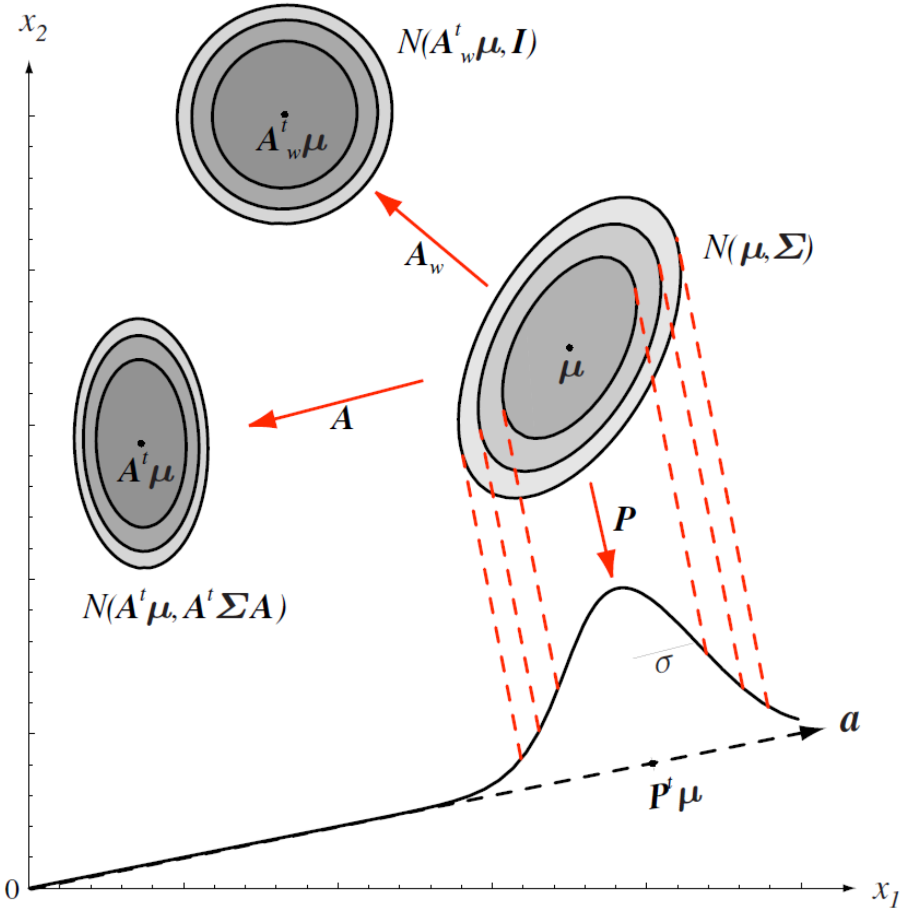
- Off-diagonal entries with a value of 0 indicate uncorrelated variables, that are statistically independent (variables likely do not influence one another)
- Covariance positive if two variables increase together (positive correlation), negative if one variable decreases when the other increases (negative correlation)

A Two-Dimensional Gaussian Distribution, with Samples Shown

- Samples drawn from a two-dimensional Gaussian lie in a cloud centered on the mean $\boldsymbol{\mu}$.
- The ellipses show lines of equal probability density of the Gaussian.



Linear Transformations in a 2D Feature Space



Discriminant Functions ($g_i(x)$) for the Normal Density

Discriminant Functions

We will consider three special cases for

- normally distributed features and
- minimum-error-rate classification (0-1 loss).

Recall

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

- if $p(x|w_i) \sim \mathcal{N}(\mu_i, \Sigma_i)$ then approximate $p(x|w_i)$ using

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^t\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right]$$

Minimum Error-Rate Discriminant Function for Multivariate Gaussian Feature Distributions

ln (natural log) of

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^t \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right]$$

Gives a general form for our discriminant functions

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mu_\mathbf{i})^t \mathbf{\Sigma_i}^{-1}(\mathbf{x}-\mu_\mathbf{i}) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\mathbf{\Sigma_i}| + \ln P(\omega_i)$$

Special Cases for Binary Classification

Purpose

Overview of commonly assumed cases for feature likelihood densities $p(x|w_i)$

Goal: Eliminate common additive constants in discriminant functions.

- These do not affect the classification decision (i.e. define $g_i(x)$ providing "just the differences")
- Look at resulting decision surfaces (defined by $g_i(x) = g_j(x)$)

Three Special Cases

1. Statistically independent features, identically distributed Gaussians for each class
2. Identical covariances for each class
3. Arbitrary covariances

Case I: $\Sigma_i = \sigma^2 I$

Recall

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mu_\mathbf{i})^t \mathbf{\Sigma_i}^{-1}(\mathbf{x}-\mu_\mathbf{i}) \boxed{-\frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\mathbf{\Sigma_i}|} + \ln P(\omega_i)$$

Remove

- Items in red: same across classes (“unimportant additive constants”)

Inverse of Covariance Matrix $\Sigma^{-1} = (\frac{1}{\sigma^2}) I$

- Only effect is to scale vector product by

Discriminant function

$$g_i(x) = -\frac{(\mathbf{x}-\mu_\mathbf{i})^t(\mathbf{x}-\mu_\mathbf{i})}{2\sigma^2} + \ln P(\omega_i)$$

$$g_i(x) = -\frac{1}{2\sigma^2}[\boxed{\mathbf{x}^t\mathbf{x}} - 2\mu_\mathbf{i}^\mathbf{t}\mathbf{x} + \mu_\mathbf{i}{}^t\mu_\mathbf{i}] + \ln P(\omega_i)$$

SCIENCE ACADEMY

Case I: $\Sigma_i = \sigma^2 \boldsymbol{I}$

Linear Discriminant Function

- Produced by factoring the previous form

$$g_i(x) = \boxed{\mathbf{w_i^t}}\mathbf{x} + \boxed{\omega_{i0}}$$

$$g_i(x) = \boxed{\frac{1}{\sigma^2}\mu_\mathbf{i}^t}\mathbf{x}\boxed{-\frac{1}{2\sigma^2}\mu_i^t\mu_i + \ln P(\omega_i)}$$

Threshold or Bias for Class i: w_{i0}

- Change in prior translates decision boundary

Case I: $\Sigma_i = \sigma^2 \boldsymbol{I}$

Decision Boundary: $g_i(x) = g_j(x)$

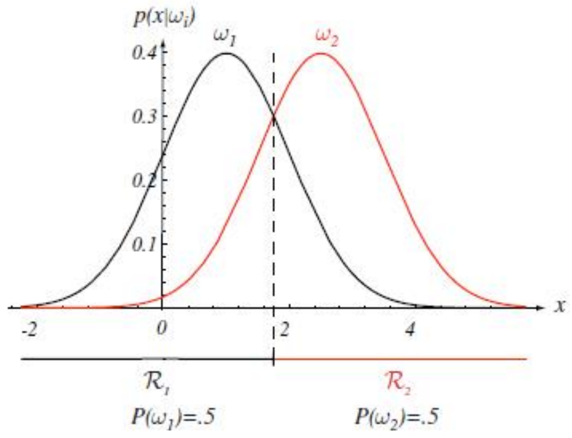
$$\boxed{\mathbf{w}}^t(\mathbf{x} - \boxed{\mathbf{x}_0}) = 0$$

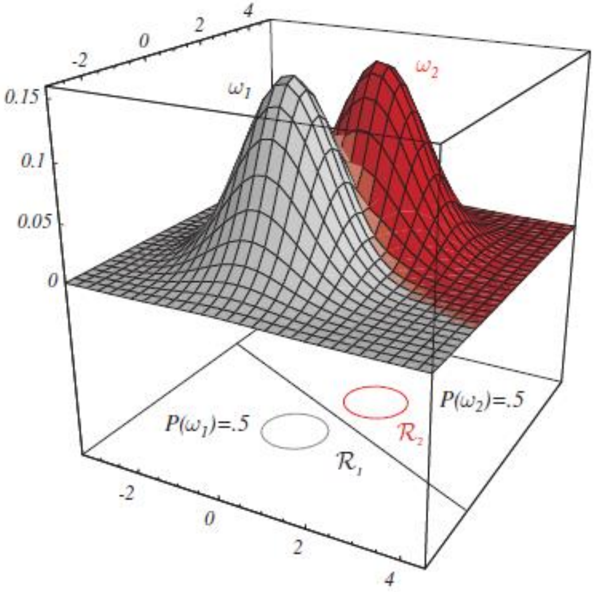
$$\boxed{(\mu_i - \mu_j)}^t \left(x - \left(\boxed{\frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2}{(\mu_i - \mu_j)^t(\mu_i - \mu_j)} \ln \frac{P(\omega_i)}{P(\omega_j)}(\mu_i - \mu_j)}\right)\right)$$

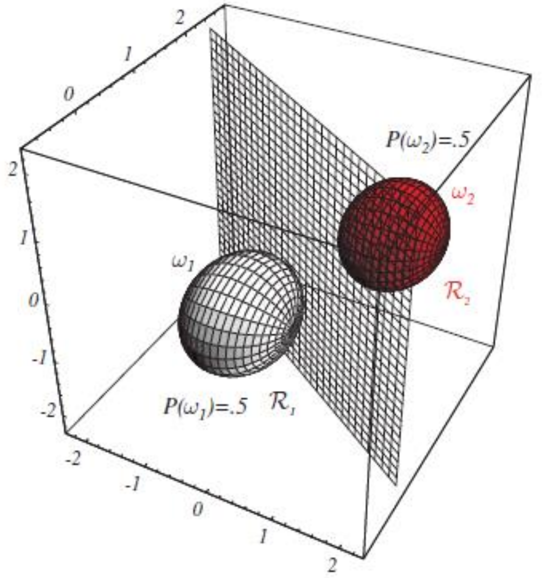
- Decision boundary goes through x_0 along line between means, orthogonal to this line.
- If priors equal, x_0 between means (minimum distance classifier), otherwise x_0 shifted.
- If variance small relative to distance between means, priors have limited effect on boundary location.

Case I: Statistically Independent Features with Identical Variances

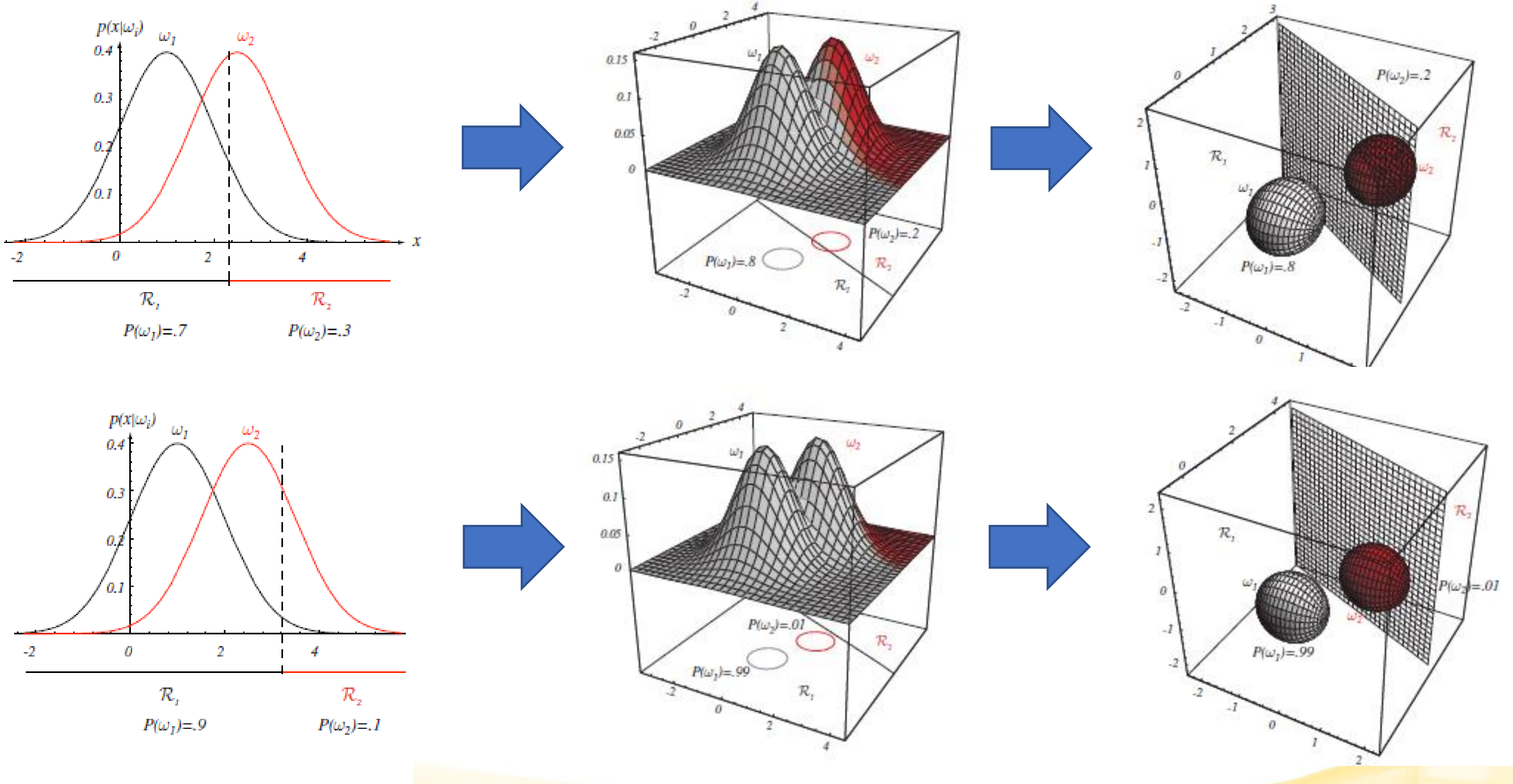
- If the covariance matrices for two distributions are equal and proportional to the identity matrix, then the distributions are spherical in d dimensions, and the boundary is a generalized hyperplane of $d\ -1$ dimensions, perpendicular to the line separating the means.
- In these one-, two-, and three-dimensional examples, we indicate $p(x|w_i)$ and the boundaries for the case $p(w_1) = p(w_2)$. In the three-dimensional case, the grid plane separates $\mathcal{R}_1 = \mathcal{R}_2$.







Translation of Decision Boundaries Through Changing Priors



Case II: Identical Covariances ($\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$)

Recall

$$g_i(\mathbf{x}) = -\frac{1}{2}\boxed{(\mathbf{x}-\mu_\mathrm{i})^t \boldsymbol{\Sigma}_\mathrm{i}^{-1}(\mathbf{x}-\mu_\mathrm{i})}\boxed{-\frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}_\mathrm{i}|} + \ln P(\omega_i)$$

Steps Remove Terms in red; as in Case I these can be ignored (same across classes)

Squared Mahalanobis Distance (shown in yellow)

Distance from x to mean for class i, taking covariance into account; defines contours of fixed density

Case II: Identical Covariances ($\Sigma_i = \Sigma$)

Expansion of squared Mahalanobis distance

$$
\begin{aligned}
(\mathbf{x}-\mu_i)^t \Sigma^{-1} (\mathbf{x}-\mu_i) \\
&= \mathbf{x}^t \Sigma^{-1} \mathbf{x} - \mathbf{x}^t \Sigma^{-1} \mu_i - \mu_i^t \Sigma^{-1} \mathbf{x} + \mu_i^t \Sigma^{-1} \mu_i \\
&= \boxed{\mathbf{x}^t \Sigma^{-1} \mathbf{x}} - 2(\Sigma^{-1} \mu_i)^t \mathbf{x} + \mu_i^t \Sigma^{-1} \mu_i
\end{aligned}
$$

the last step comes from symmetry of the covariance matrix and thus its inverse

$$\Sigma^t = \Sigma, (\Sigma^{-1})^t = \Sigma^{-1}$$

Once again, term above in red is an additive constant independent of class, and can be removed

Case II: Identical Covariances ($\Sigma_i = \Sigma$)

Linear Discriminant Function

$$g_i(x) = \boxed{\mathbf{w}_i^t}\mathbf{x} + \boxed{\omega_{i0}}$$

$$g_i(\mathbf{x}) = \boxed{(\Sigma^{-1}\mu_i)^t}\mathbf{x} \boxed{-\frac{1}{2}\mu_i^t\Sigma^{-1}\mu_i + \ln P(\omega_i)}$$

Decision Boundary

$$g_i(\boldsymbol{x}) = g_j(\boldsymbol{x})$$

$$\boxed{\mathbf{w}}^t(\mathbf{x} - \boxed{\mathbf{x}_0}) = 0$$

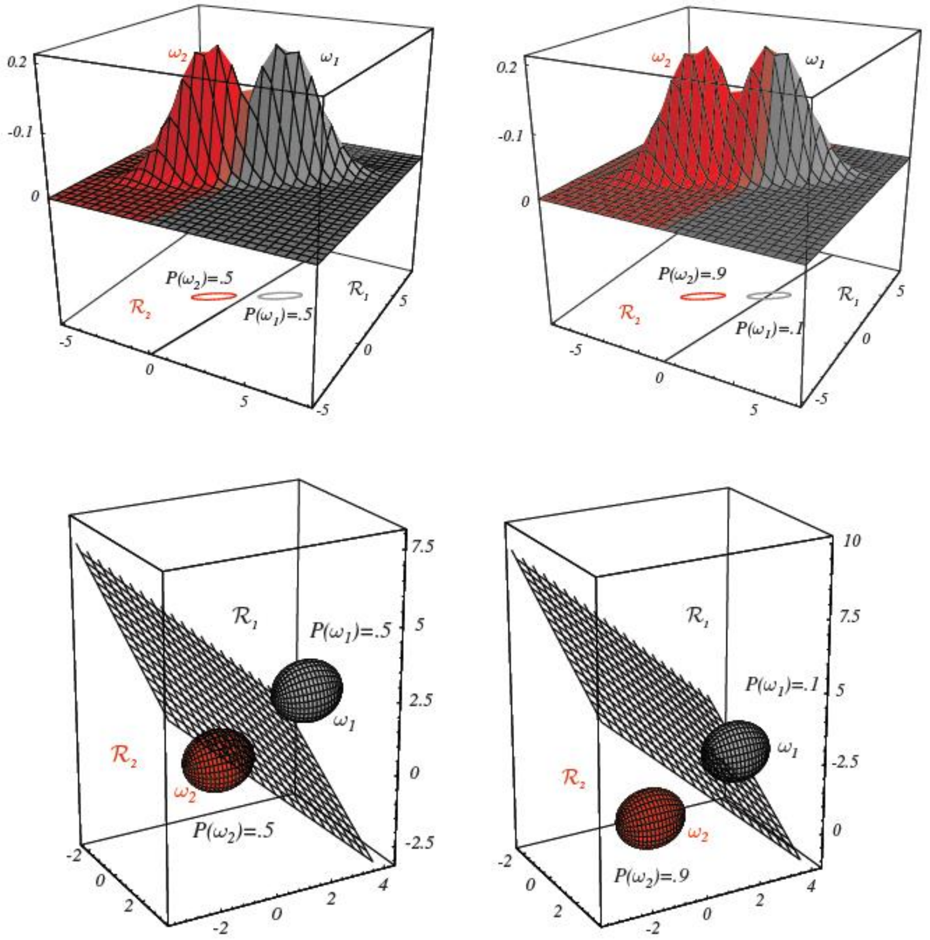
$$\boxed{(\boldsymbol{\Sigma}^{-1}(\mu_i - \mu_j))}^t(\mathbf{x} - \boxed{\left(\frac{1}{2}(\mu_i + \mu_j) - \frac{\ln[P(\omega_i)/P\omega_j)}{(\mu_i - \mu_j)\boldsymbol{\Sigma}^{-1}(\mu_i - \mu_j)}(\mu_i - \mu_j)\right)}) = 0$$

SCIENCE ACADEMY

Case II: Identical Covariances ($\Sigma_i = \Sigma$)

Notes on Decision Boundary

- As for Case I, passes through point x0 lying on the line between the two class means. Again, x0 in the middle if priors identical.
- Hyperplane defined by boundary generally not orthogonal to the line between the two means.



Case III: Arbitrary Covariance

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mu_\mathbf{i})^t\mathbf{\Sigma_i}^{-1}(\mathbf{x}-\mu_\mathbf{i}) \boxed{-\frac{d}{2}\ln 2\pi} - \frac{1}{2}\ln|\mathbf{\Sigma_i}| + \ln P(\omega_i)$$

Remove

- Can only remove the one term in red above

Discriminant Function (quadratic)

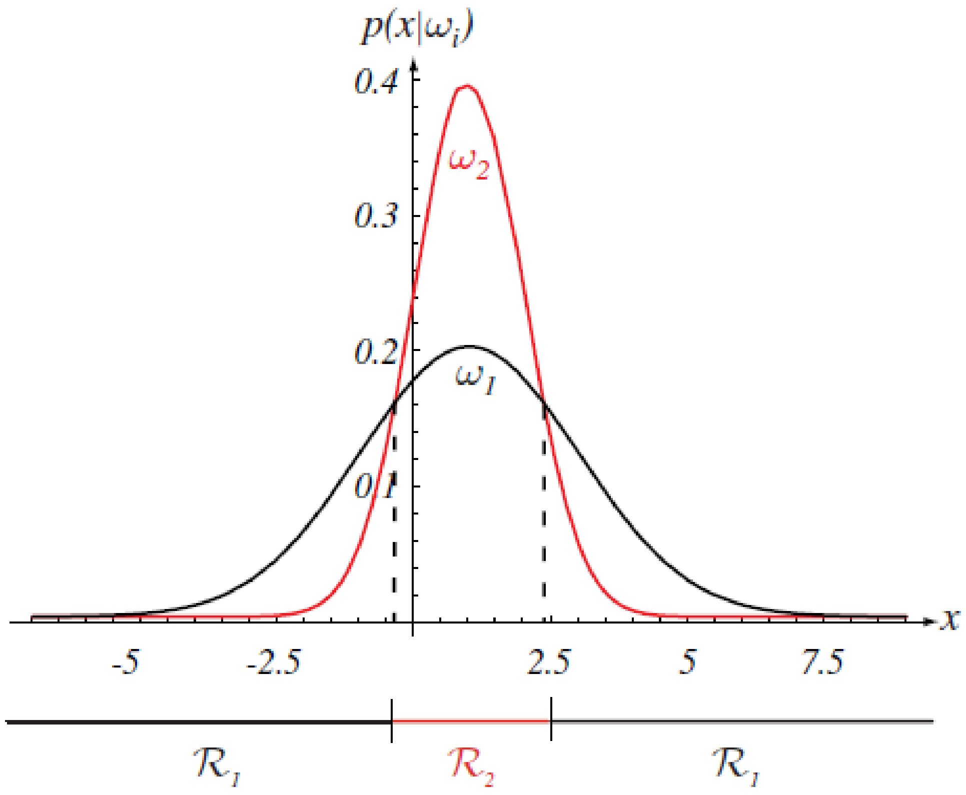
$$g_i(x) = x^t \boxed{W_i} x + \boxed{w_i^t} x + \boxed{\omega_{i0}}$$

$$g_i(x) = x^t \boxed{(-\frac{1}{2}\Sigma_i^{-1})} x + \boxed{(\Sigma_i^{-1}\mu_i)^t} x \boxed{-\frac{1}{2}\mu_i^t\Sigma_i^{-1}\mu_i - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)}$$

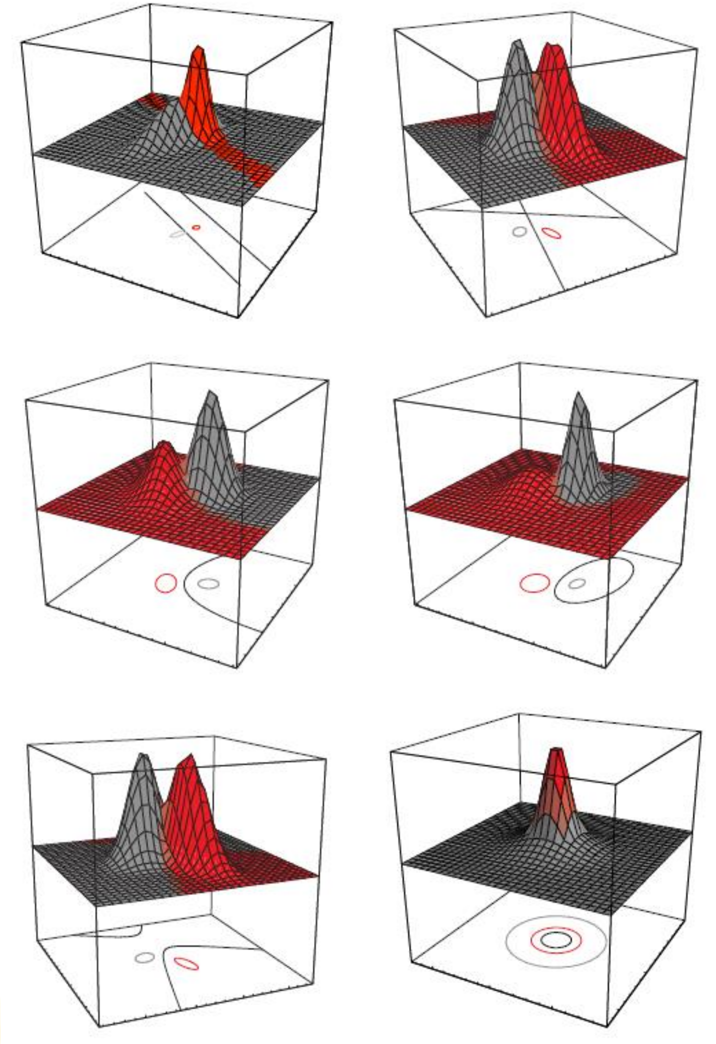
Case III: Arbitrary Covariance

Decision Boundaries

- Are hyperquadrics: can be hyperplanes, hyperplane pairs, hyperspheres, hyperellipsoids, hyperparabaloids, hyperhyperparabaloids
- Need not be simply connected, even in one dimension (next slide)



Case III: Arbitrary Covariance



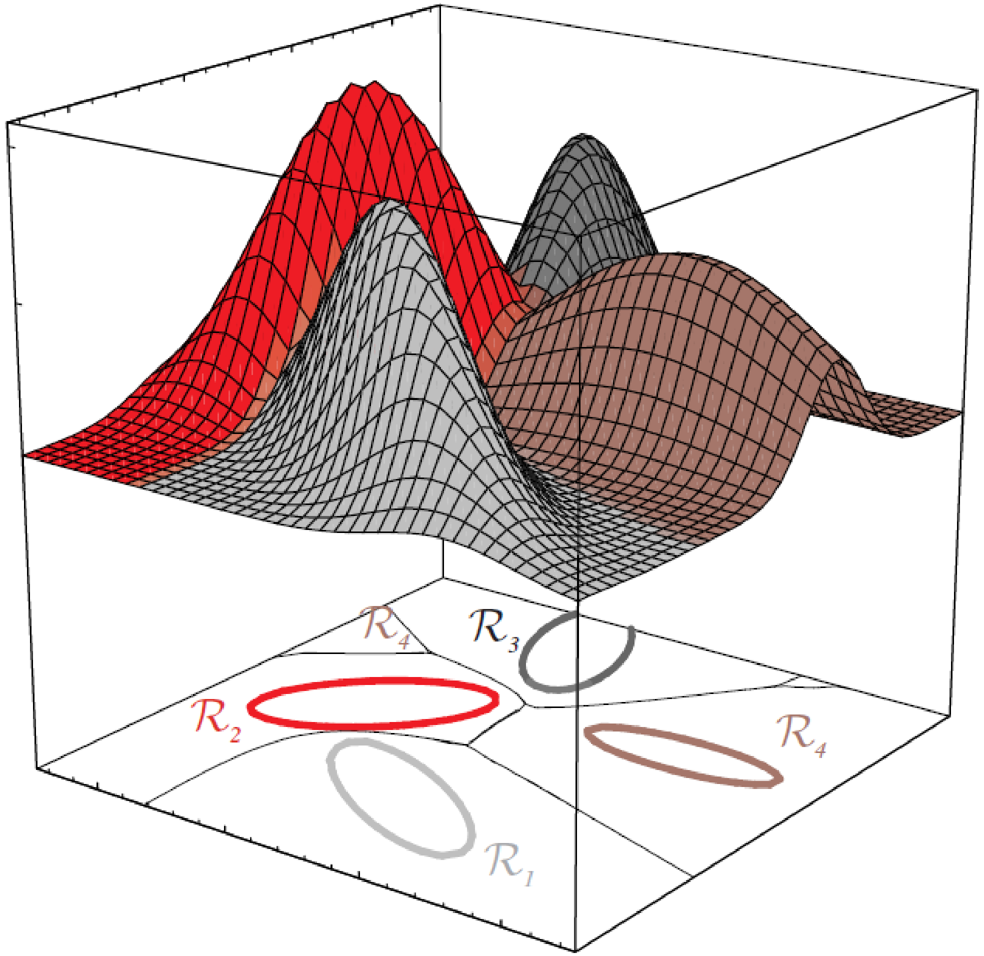
More than Two Categories

Boundary

Defined by two most likely classes for each segment

Other Distributions

Possible; underlying Bayesian Decision Theory is unmodified, however



Discrete Features

Roughly speaking...

Replace probability densities by probability mass functions. Expressions using integrals are changed to use summations, e.g.

$$\int p(x|w_j)dx \Rightarrow \sum_x P(x|w_j)$$

Bayes Formula

$$P(\omega_j|\mathbf{x}) = \frac{P(\mathbf{x}|\omega_j)P(\omega_j)}{P(\mathbf{x})}$$

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

Example: Independent Binary Features

Binary Feature Vector

$\boldsymbol{x} = \{x_1, \ldots, x_d\}$ of 0/1 -valued features, where each x_i is 0/1 with probability

$$\mathrm{p_i} = \mathrm{Pr}[\mathrm{x_i} = 1 | \mathrm{w_1}]$$

Conditional Independence

Assume that given a class, the features are independent

Likelihood Function

$$P(\boldsymbol{x}|w_1) = \prod_{i=1}^{d} p_i^{x_i}(1 - p_i)^{1-x_i}$$

Signal Detection & Operating Characteristics

SCIENCE ACADEMY

Signal Detection & Operating Characteristics

Suppose: We are interested in detecting a single weak pulse.

Model: Two Gaussian distributions where the signal (e.g., voltage) is denoted by a Gaussian with mean μ_2 if a pulse is present and mean μ_1 if not present.

$$p(x|w_i) \sim \mathcal{N}(\mu_i, \sigma^2)$$

Detection classifier Finds a threshold value $x^\star$

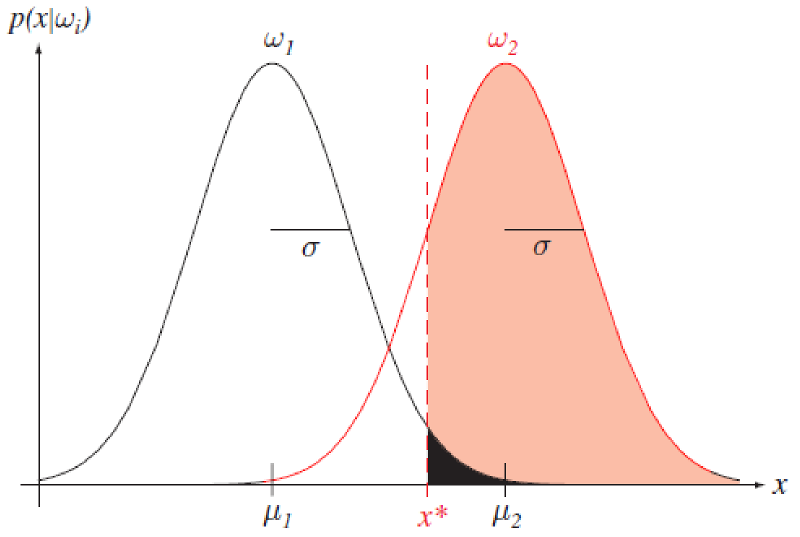


FIGURE 2.19. During any instant when no external pulse is present, the probability density for an internal signal is normal, that is, $p(x|\omega_1) \sim N(\mu_1, \sigma^2)$; when the external signal is present, the density is $p(x|\omega_2) \sim N(\mu_2, \sigma^2)$. Any decision threshold x^* will determine the probability of a hit (the pink area under the ω_2 curve, above x^*) and of a false alarm (the black area under the ω_1 curve, above x^*). From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Signal Detection & Operating Characteristics

Four probabilities

- **Hit:** Signal is above $x^\star$ given pulse present, i.e., $P(x > x^\star | x \in w_2)$
- **False alarm:** Signal is above $x^\star$ given pulse not present, i.e., $P(x > x^\star | x \in w_1)$
- **Miss:** Signal is below $x^\star$ given pulse is present, i.e., $P(x < x^\star | x \in w_2)$
- **Correct rejection:** Signal is below $x^\star$ and pulse not present, i.e., $P(x < x^\star | x \in w_1)$

Signal Detection & Operating Characteristics

Receiver Operating Characteristic (ROC)

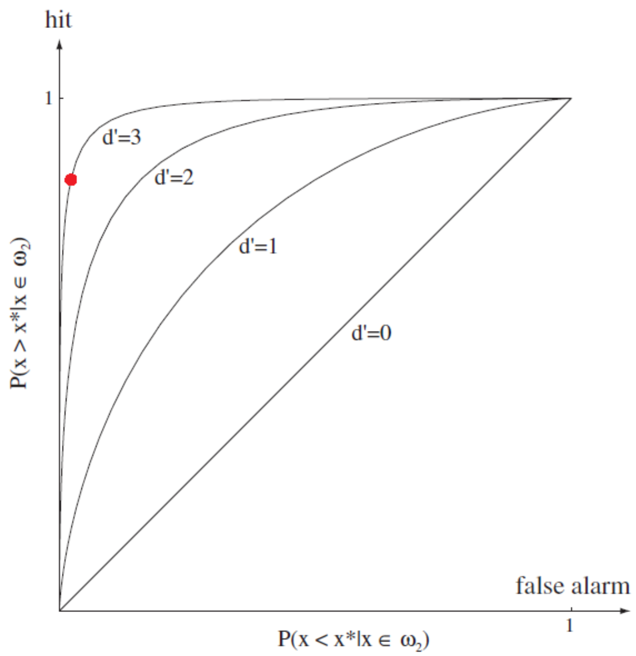
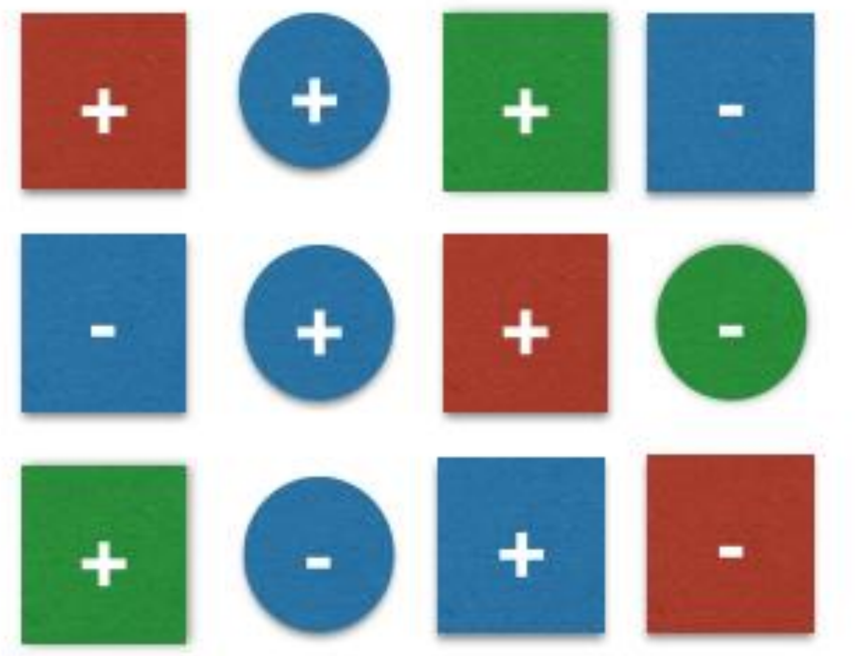


FIGURE 2.20. In a receiver operating characteristic (ROC) curve, the abscissa is the probability of false alarm, $P(x > x^*|x \in \omega_1)$, and the ordinate is the probability of hit, $P(x > x^*|x \in \omega_2)$. From the measured hit and false alarm rates (here corresponding to x^* in Fig. 2.19 and shown as the red dot), we can deduce that $d' = 3$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Numerical Examples

Example Problem

- Consider a simple toy dataset of 12 samples belonging to two different classes (+ and −).
- $w_i \in \{-, +\}$.



Features

- Color and geometrical shape:

$$x_i = [x_{i1}, x_{i2}]$$

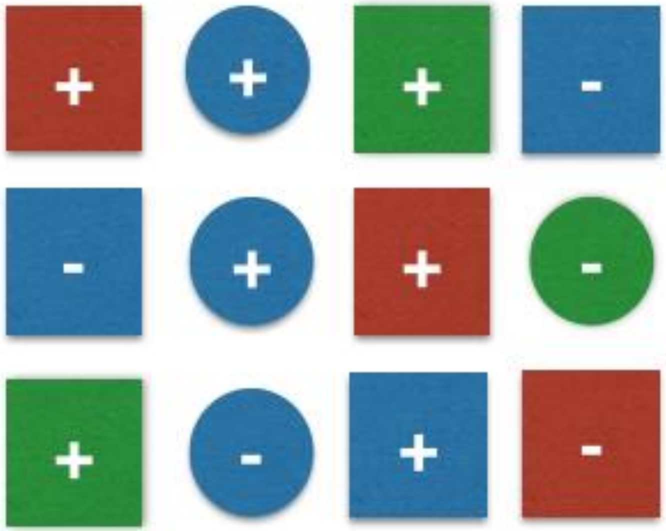
- $x_{i1} \in \{blue, green, red, yellow\}$
- $x_{i2} \in \{circle, square\}$

Problem: Classify a new sample **Blue square**

Example Problem

Prior probabilities

- $P(+) = \frac{7}{12}$ and $P(-) = \frac{5}{12}$.



Conditional probabilities

- $p(x|+) = p(blue|+) \cdot p(square|+) = \frac{3}{7} \cdot \frac{5}{7} = 0.31$
- $p(x|-) = p(blue|-) \cdot p(square|-) = \frac{3}{5} \cdot \frac{3}{5} = 0.36.$

Posterior probability

- $P(+|x) = p(x|+) \cdot P(+) = 0.31 \cdot 0.58 = 0.18$
- $P(-|x) = p(x|-) \cdot P(-) = 0.36 \cdot 0.42 = 0.15.$

Example Problem

Putting it all together

New sample can be classified by plugging the posterior probabilities

if $P(+|x) \geq P(-|x)$ Classify as $+$

else Classify as $-$

$+$

Since 0.18 > 0.15, then sample is classified as $+$.

- **Exercise: (not graded)** What would happen if prior probabilities were equal for both classes?

Challenge

What is there is a “new” value

- Consider when we have a new color attribute that is not present

in the training dataset:

A yellow square ?

- If the color yellow does not appear in our training dataset, the class-conditional probability will be 0.
- The posterior probability will also be 0:

$$P(w_1|x) = 0 \cdot 0.42 = 0 \text{ or } P(w_2|x) = 0 \cdot 0.58 = 0.$$

Additive Smoothing

- To avoid the problem of zero probabilities, an additional smoothing term can be added to the multinomial Bayes model.
- The most common variants of additive smoothing are the so-called Lidstone smoothing (< 1) and Laplace smoothing (= 1).
- $P(x_i|w_j = \frac{N_{x_i,w_j}+\alpha}{N_{w_j}+\alpha d}$ for $i = 1, \cdots, d.$
  - ➢ $\boldsymbol{N_{x_i,w_j}}$: Number of times feature x_i appears in samples from class w_j.
  - ➢ $\boldsymbol{N_{w_j}}$: Total count of all features in class w_j.
  - ➢ $\boldsymbol{\alpha}$: Parameter for additive smoothing.
  - ➢ $\boldsymbol{d}$: Dimensionality of the feature vector $x = [x_1, \cdots, x_d]$.

Coding Examples

Coding Exercises

- We will be using Python language.
- Jupyter notebooks are helpful to visualize data and classifiers.
- You can use one of the following:
  - https://colab.research.google.com/
  - Local computer (any computer works!)