# Generative Models

**DSE 210**

# Outline : Lecture 4

- **Generative Models**
  - **Review of Bayes' theorem**
  - **Generative Models**
  - **Naive Bayes(NB) and Multinomial-NB**
  - **Self practice - NB**
  - **Covariance and Correlation**
  - **Self practice - Covariance and Correlation**
  - **Uni-, bi- and multi- variate Gaussian**
  - **Discriminant Analysis**
  - **Hands-on and self practice**

# Quick review of conditional probability

Formula for conditional probability for any events *A*, *B*,

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}$$

Applied twice, this yields Bayes' rule:

$$\Pr(H|E) = \frac{\Pr(E|H)}{\Pr(E)} \Pr(H)$$

Example: Toss ten coins. What is the probability that the first is heads, given that nine of them are heads?

*H* = first coin is heads

*E* = nine of the ten coins are heads

$$\Pr(H|E) = \frac{\Pr(E|H)}{\Pr(E)} \cdot \Pr(H) = \frac{\binom{9}{8}\frac{1}{2^9}}{\binom{10}{9}\frac{1}{2^{10}}} \cdot \frac{1}{2} = \frac{9}{10}$$

# Why Bayes' Rule?

$$\Pr(H|E) = \frac{\Pr(E|H)}{\Pr(E)} \Pr(H)$$

- describes the probability of an event based on prior knowledge of conditions that might be related to the event

# Quick Quiz

- Probability of a patient having liver disease is 0.1, and the probability of an incoming patient being alcoholic is 0.05. If the probability of being alcoholic given that the person has liver disease is 0.07, then what is the probability of an incoming patient having liver disease given that he is alcoholic?

  A = Patient has liver disease
  B = Patient is an alcoholic

  $P(A) = 0.10, P(B) = 0.05, P(B|A) = 0.07$

  $P(A|B) = (0.07 * 0.1)/0.05 = 0.14$

# Disjoint and Independent Events

## Disjoint or Mutually Exclusive

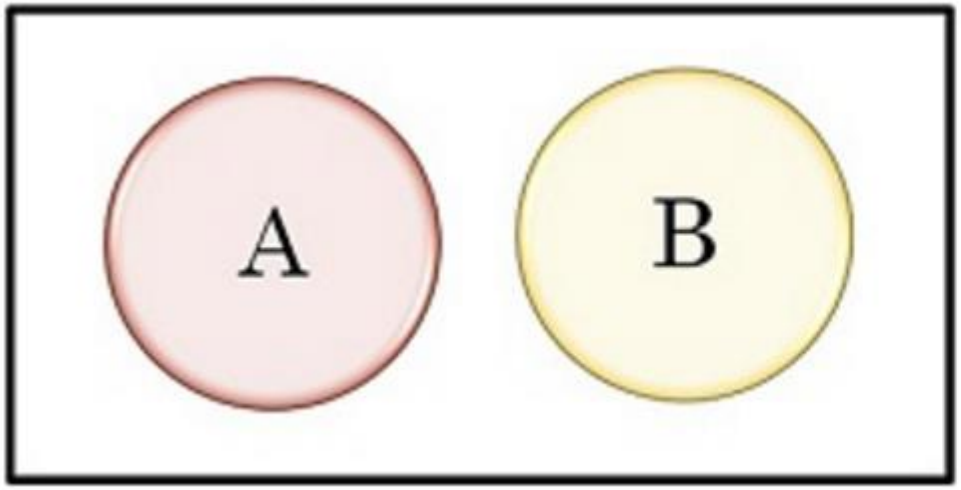
- Disjoint events cannot happen at the same time.
- e.g.: when tossing a coin, the result can either be heads or tails but cannot be both.

## Independent

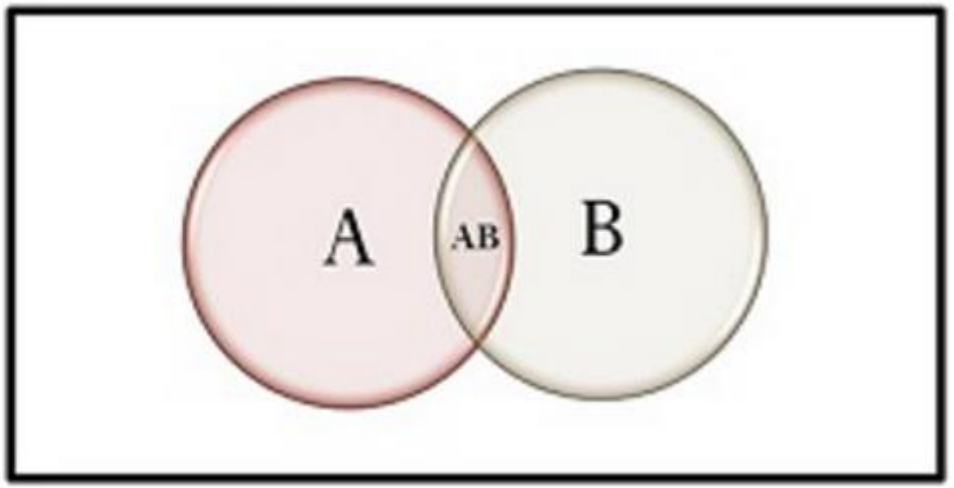
- Occurrence of one event does not influence the other(s).
- e.g.: when tossing two coins, the result of one flip does not affect the result of the other.

# Disjoint and Independent Events

Disjoint Events




Independent Events



# Summation rule

Suppose events $A_1, \ldots, A_k$ are disjoint events, one of which must occur. Then for any other event $E$ ,

$$\Pr(E) = \Pr(E \cap A_1) + \Pr(E \cap A_2) + \cdots + \Pr(E \cap A_k)$$

$$= \Pr(E \mid A_1)\Pr(A_1) + \Pr(E \mid A_2)\Pr(A_2) + \cdots + \Pr(E \mid A_k)\Pr(A_k)$$

# Generative models

An unknown underlying distribution $D$ over $X \times Y$.

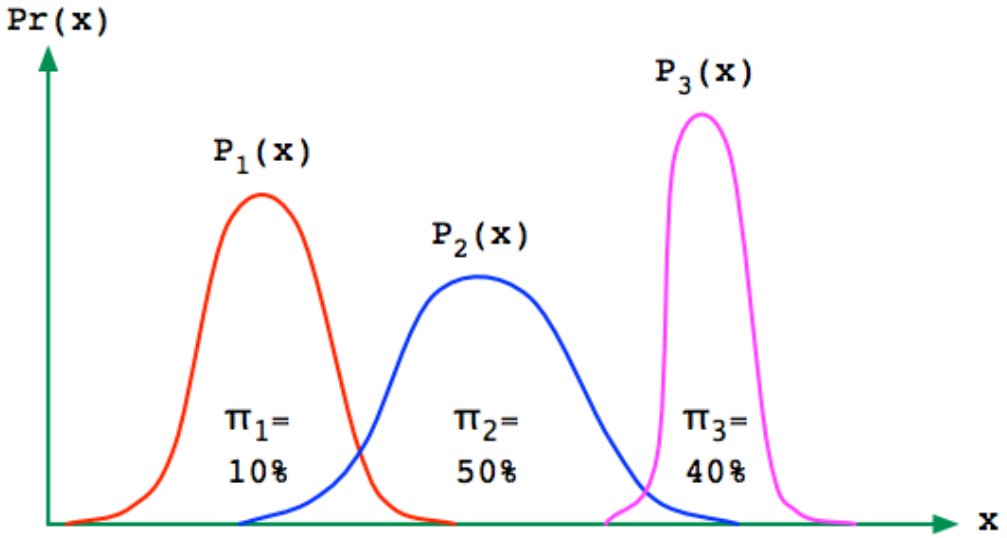
Generating a point ($x$ , $y$) in two steps:

1. When we were studying NN: first choose $x$, then choose $y$ given $x$.
2. Now: first choose $y$ , then choose $x$ given $y$.

Example:

$X$ = R

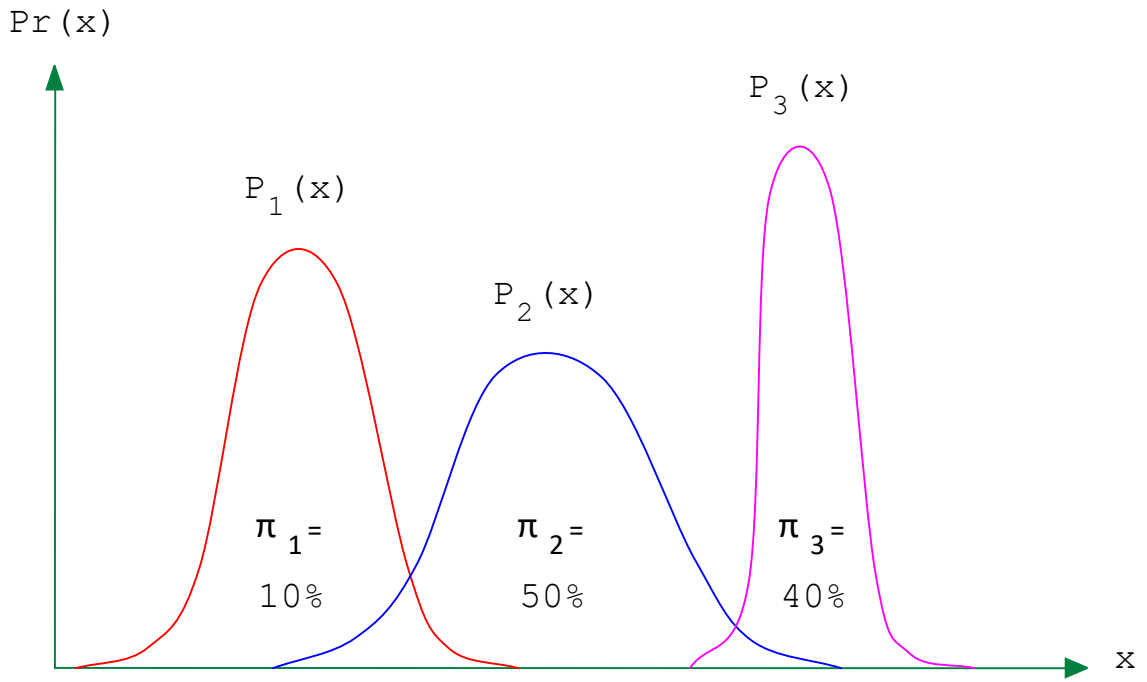
$Y$ = {1, 2, 3}




The overall density is a mixture of the individual densities,

$$\Pr(x) = \pi_1 P_1(x) + \cdots + \pi_k P_k(x).$$

# The Bayes-optimal prediction




Labels $Y = \{1, 2, \ldots, k\}$, density $\Pr(x) = \pi_1 P_1(x) + \cdots + \pi_k P_k(x)$

For any $x \in \mathcal{X}$ and any label $j$,

$$\Pr(y = j|x) = \frac{\Pr(y = j)\Pr(x|y = j)}{\Pr(x)} = \frac{\pi_j P_j(x)}{\sum_{i=1}^{k} \pi_i P_i(x)}$$

Bayes-optimal prediction: $h^*(x) = \arg\max_j \pi_j P_j(x)$.

Estimating the $\pi_j$ is easy. Estimating the $P_j$ is hard.

# Estimating class-conditional distributions

Estimating an arbitrary distribution in $R^p$ can be hard.

Solution: approximate each $P_j$ with a simple, parametric distribution.

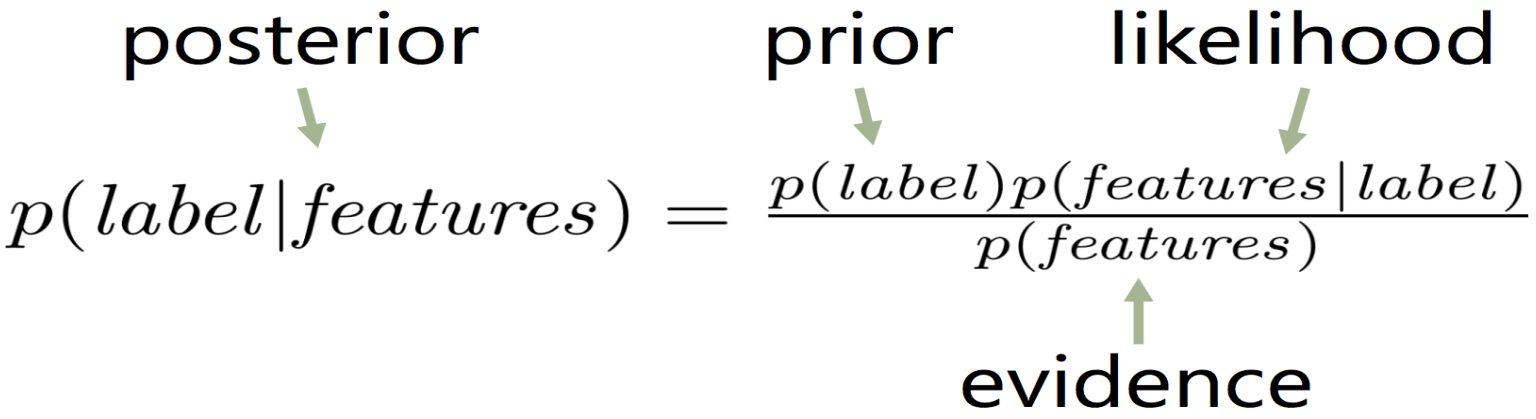
Some options:

- Product distributions.
  Assume coordinates are independent: naive Bayes.
- Multivariate Gaussians.
  Linear and quadratic discriminant analysis.
- More general graphical models.

# Naive Bayes

1. Probabilistic model (fits *P*(*label* |*data*))
2. Makes a conditional independence assumption that features are independent given the label.

$P(feature_i , feature_j | label) = P(feature_i | label) \cdot P(feature_i | label)$



# Naive Bayes

Due to the conditional independence assumption, we get

$$p(label|features) = \frac{p(label)\prod_i p(feature_i|label)}{p(features)}$$

Denominator doesn't matter because we are interested in

$$p(label|features) \quad \text{vs.} \quad p(\neg label|features)$$

both of which have same denominator

# Naive Bayes

Labels $Y = \{1, 2, \ldots, k\}$, density $\Pr(x) = \pi_1 P_1(x) + \cdots + \pi_k P_k(x)$.

Binarized MNIST:

- $k = 10$ classes
- $X = \{0, 1\}^{784}$

Assume that **within each class**, the individual pixel values are independent.

$$P_j(x) = P_{j1}(x_1) \cdot P_{j2}(x_2) \cdots P_{j,784}(x_{784}).$$

# Smoothed estimate of coin bias

Pick a class $j$ and a pixel $i$. We need to estimate

$$p_{ji} = \Pr(x_i = 1 | y = j).$$

Out of a training set of size $n$,

$n_j$ = # of instances of class $j$

$n_{ji}$ = # of instances of class $j$ with $x_i = 1$

Then the maximum-likelihood estimate of $p_{ji}$ is

$$\widehat{p}_{ji} = n_{ji}/n_j.$$

This causes problems if $n_{ji} = 0$. Instead, use "Laplace smoothing":

$$\widehat{p}_{ji} = \frac{n_{ji} + 1}{n_j + 2}.$$

# Maximum Likelihood

Given observed values $X_1 = x_1$, $X_2 = x_2$ … $X_n = x_n$.

Likelihood($\theta$) = probability of observing the given data as a function of $\theta$.

Maximum Likelihood estimate of $\theta$ = value of $\theta$ that maximises Likelihood($\theta$).

http://statweb.stanford.edu/~susan/courses/s200/lectures/lect11.pdf

# Form of the classifier

Data space $X = \{0, 1\}^p$, label space $Y = \{1, \ldots, k\}$. Estimate:

- $\{\pi_j : 1 \leq j \leq k\}$
- $\{p_{ji} : 1 \leq j \leq k, 1 \leq i \leq p\}$

Then classify point $x$ as

$$\arg\max_j \quad \pi_j \prod_{i=1}^{p} p_{ji}^{x_i} (1 - p_{ji})^{1 - x_i}.$$

To avoid underflow: take the log:

$$\arg\max_j \quad \underbrace{\log \pi_j + \sum_{i=1}^{p} (x_i \log p_{ji} + (1 - x_i) \log(1 - p_{ji}))}_{\text{of the form } w \cdot x + b}$$

A linear classifier!

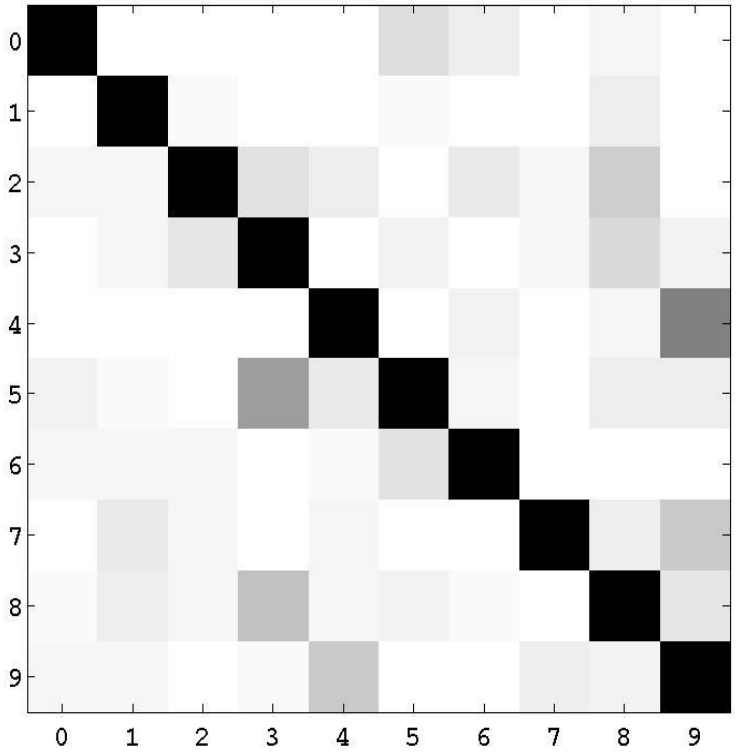
$$w_i^{(j)} = log(p_{ji}) - log(1 - p_{ji}), b^{(j)} = log(\pi_j) + \sum_{i}^{p} log(1 - p_{ji})$$

# Example: MNIST

Result of training: mean vectors for each class.

Test error rate: 15.54%.

Visualization of the "confusion matrix" →



# Other types of data

How would you handle data:

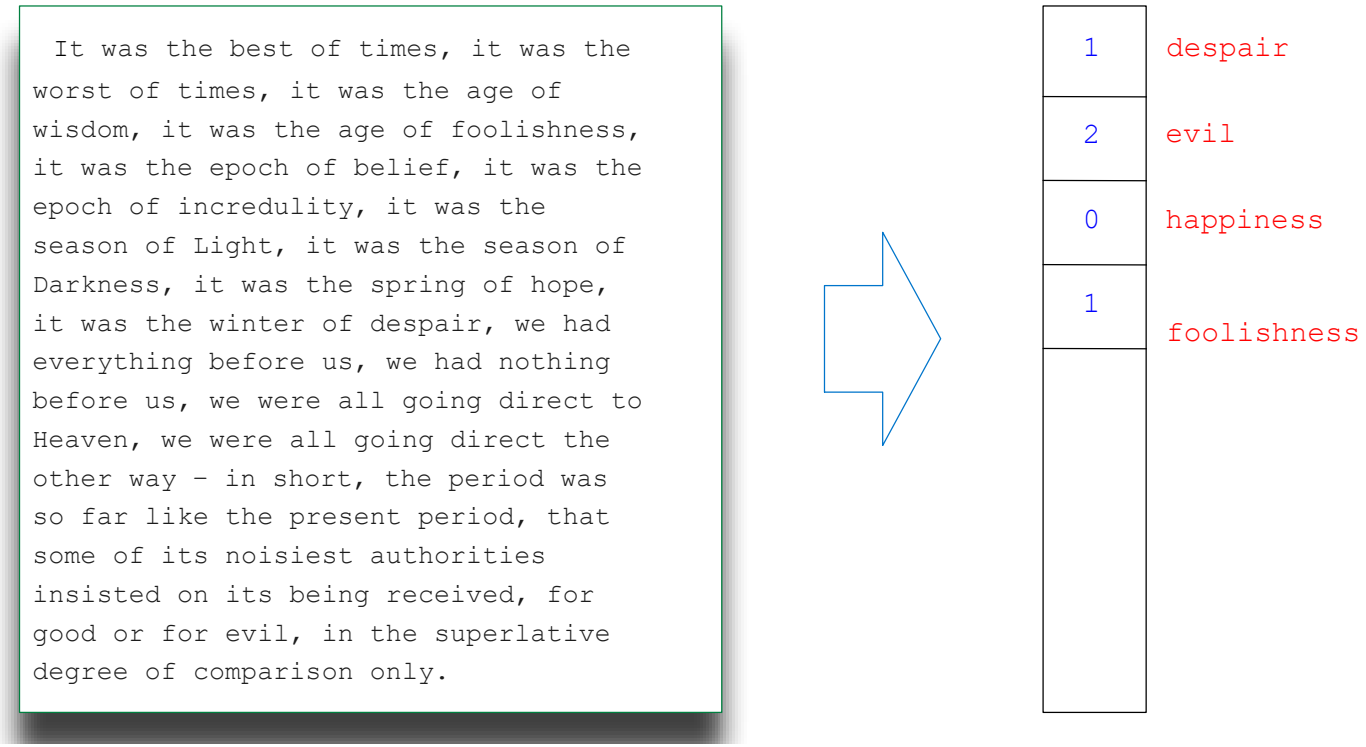
- Whose features take on more than two discrete values (such as ten possible colors)?
- Whose features are real-valued?
- Whose features are positive integers?
- Whose features are mixed: some real, some Boolean, etc?

How would you handle "missing data": situations in which data points occasionally (or regularly) have missing entries?

- At train time: **???**
- At test time: **???**

# Handling text data

## Bag-of-words: vectorial representation of text documents.




- Fix $V$ = some vocabulary.
- Treat each document as a vector of length $|V|$:

$$x = (x_1, x_2, \ldots, x_{|V|}),$$

where $x_i$ = # of times the $i$th word appears in the document.

A standard distribution over such document-vectors $x$: the **multinomial**.

# Multinomial naive Bayes

**Multinomial** distribution over a vocabulary $V$:

$$p = (p_1, \ldots, p_{|V|}), \quad \text{such that} \quad p_i \geq 0 \text{ and } \sum_i p_i = 1$$

Document $x = (x_1, \ldots, x_{|V|})$ has probability $\propto p_1^{x_1} p_2^{x_2} \cdots p_{|V|}^{x_{|V|}}$.

For naive Bayes: one multinomial distribution per class.

- Class probabilities $\pi_1, \ldots, \pi_k$
- Multinomials $p^{(1)} = (p_{11}, \ldots, p_{1|V|}), \ldots, p^{(k)} = (p_{k1}, \ldots, p_{k|V|})$

Classify document $x$ as

$$\arg\max_j \quad \pi_j \prod_{i=1}^{|V|} p_{ji}^{x_i}.$$

(As always, take log to avoid underflow: linear classifier.)

# Bernoulli vs Multinomial

Naive Bayes using Bernoulli Distribution

$$\arg\max_j \quad \pi_j \prod_{i=1}^{P} p_{ji}^{x_i}(1-p_{ji})^{1-x_i}.$$

Naive Bayes using Multinomial Distribution

$$\arg\max_j \quad \pi_j \prod_{i=1}^{|V|} p_{ji}^{x_i}.$$

# Improving performance of multinomial naive Bayes

A variety of heuristics that are standard in text retrieval, such as:

1. Compensating for burstiness.
   Problem: Once a word has appeared in a document, it has a much higher chance of appearing again.

   Solution: Instead of the number of occurrences $f$ of a word, use $\log(1 + f)$.

2. Downweighting common words.
   Problem: Common words can have a unduly large influence classification

   Solution: Weight each word $w$ by **inverse document frequency**:

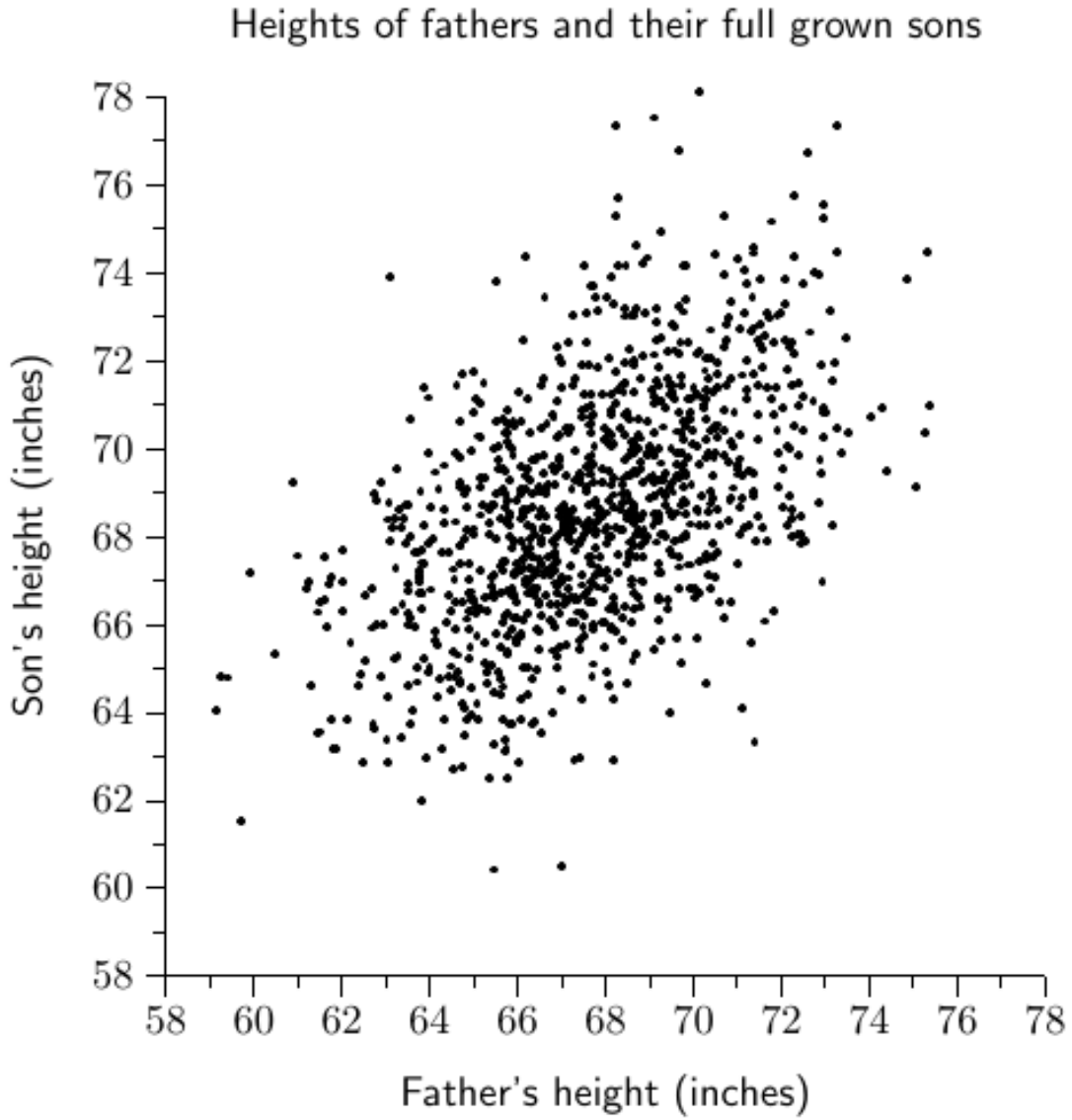
$$\log \frac{\#\ \text{docs}}{\#(\text{docs containing } w)}$$

# Self practice

Worksheet 5 - Question 18
(last page in HW4.pdf)

# Correlation and Covariance

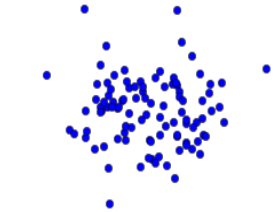
# Pearson (1903): fathers and sons



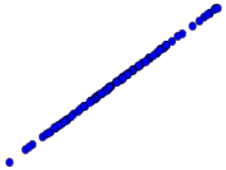

How to quantify the degree of correlation?

# Correlation pictures

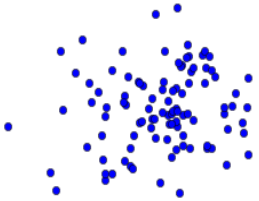
$r = 0$



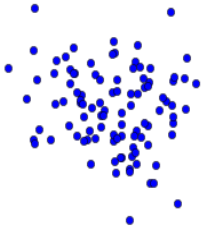
$r = 1$



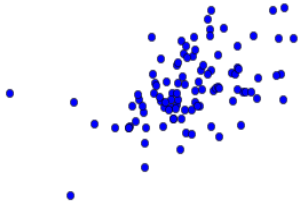
$r = 0.25$



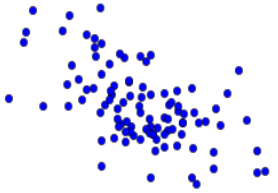
$r = -0.25$



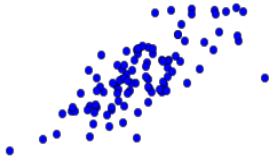
$r = 0.5$



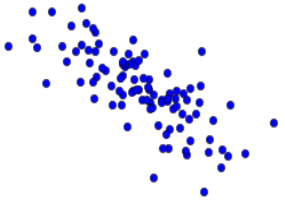
$r = -0.5$



$r = 0.75$



$r = -0.75$



# Covariance and correlation

Suppose $X$ has mean $\mu_X$ and $Y$ has mean $\mu_Y$

- Covariance

$$\text{cov}(X, Y) = \mathbf{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbf{E}[XY] - \mu_X \mu_Y$$

Maximized when $X = Y$, in which case it is $\text{var}(X)$.
In general, it is at most $\text{std}(X)\text{std}(Y)$.

- Correlation

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\text{std}(X)\ \text{std}(Y)}$$

This is always in the range $[-1, 1]$.

# Question

For variables X and Y, cov(X,Y)=0.

Are X and Y independent? No.

| x | y | P(x,y) |
|---|---|---|
| -1 | 1 | 1/3 |
| 0 | 0 | 1/3 |
| 1 | 1 | 1/3 |

$E[XY] = -1x1/3 + 0x1/3 + 1x1/3 = 0$

$E[X] = 0$

$E[Y] = 2/3$

$Cov(X,Y) = E[XY] – E[X]E[Y] = 0$

$P(X=1,Y=1) = 1/3$ $P(X=1) = 1/3, P(Y=1) = 2/3$

$P(X,Y) \neq P(X)P(Y)$

# Question

Random variables X and Y are uncorrelated.

Are X and Y independent? No.

$$\mathrm{corr}(X, Y) = \frac{\mathrm{cov}(X, Y)}{\mathrm{std}(X)\,\mathrm{std}(Y)}$$

# Question

Random variables X and Y are independent.

Can X and Y be correlated? No.

$$P(X,Y) = P(X)P(Y)$$

$$E[XY] = E[X]E[Y]$$

$$Cov(X,Y) = E[XY] - E[X]E[Y] = 0$$

# Covariance and correlation: example 1

$$\mathrm{cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}[XY] - \mu_X \mu_Y$$

$$\mathrm{corr}(X, Y) = \frac{\mathrm{cov}(X, Y)}{\mathrm{std}(X)\mathrm{std}(Y)}$$

| $x$ | $y$ | $\Pr(x, y)$ |
|---|---|---|
| $-1$ | $-1$ | $1/3$ |
| $-1$ | $1$ | $1/6$ |
| $1$ | $-1$ | $1/3$ |
| $1$ | $1$ | $1/6$ |

$$\mu_X = 0$$
$$\mu_Y = -1/3$$
$$\mathrm{var}(X) = 1$$
$$\mathrm{var}(Y) = 8/9$$
$$\mathrm{cov}(X, Y) = 0$$
$$\mathrm{corr}(X, Y) = 0$$

In this case, $X, Y$ are independent. Independent variables always have zero covariance and correlation.

# Covariance and correlation: example 2

$$\text{cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}[XY] - \mu_X \mu_Y$$

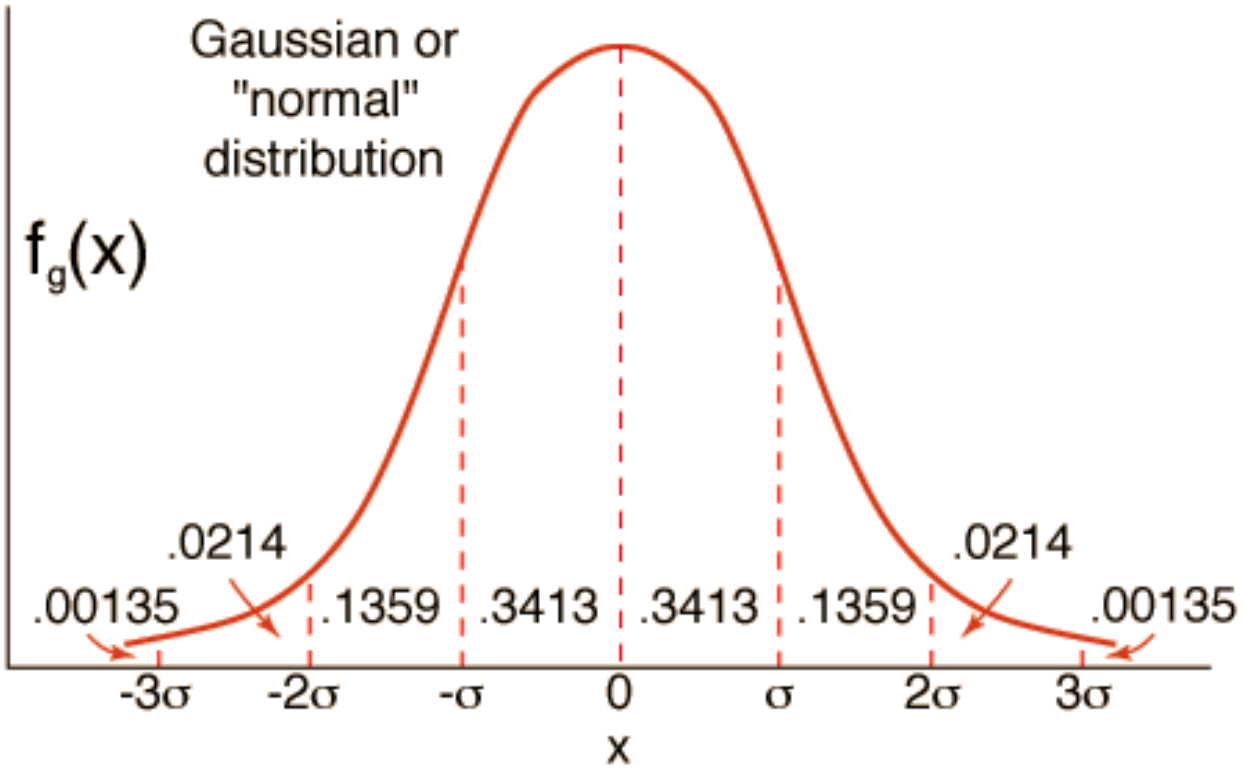
$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\text{std}(X)\text{std}(Y)}$$

| $x$ | $y$ | $\Pr(x, y)$ |
|---|---|---|
| $-1$ | $-10$ | $1/6$ |
| $-1$ | $10$ | $1/3$ |
| $1$ | $-10$ | $1/3$ |
| $1$ | $10$ | $1/6$ |

$$\mu_X = 0$$
$$\mu_Y = 0$$
$$\text{var}(X) = 1$$
$$\text{var}(Y) = 100$$
$$\text{cov}(X, Y) = -10/3$$
$$\text{corr}(X, Y) = -1/3$$

In this case, $X$ and $Y$ are negatively correlated.

# The univariate Gaussian




The Gaussian $N(\mu, \sigma^2)$ has mean $\mu$, variance $\sigma^2$, and density function

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

But what if we have **two** variables?

http://hyperphysics.phy-astr.gsu.edu/hbase/Math/gaufcn.html

# The bivariate (2-d) Gaussian

A distribution over $(x, y) \in \mathbb{R}^2$, parametrized by:

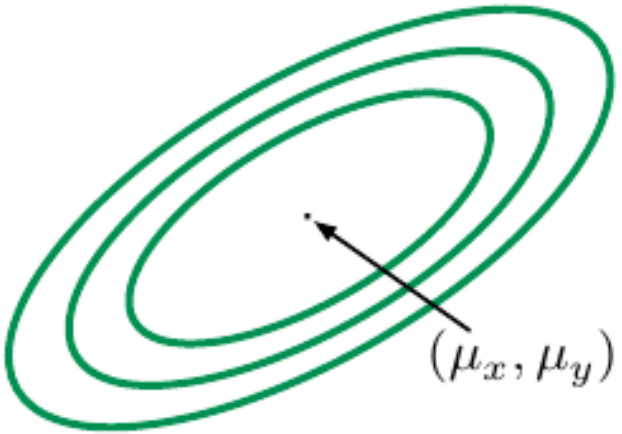
- **Mean** $(\mu_x, \mu_y) \in \mathbb{R}^2$
- **Covariance matrix**

$$\Sigma = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}$$

where $\Sigma_{xx} = \text{var}(X)$, $\Sigma_{yy} = \text{var}(Y)$, $\Sigma_{xy} = \Sigma_{yx} = \text{cov}(X, Y)$

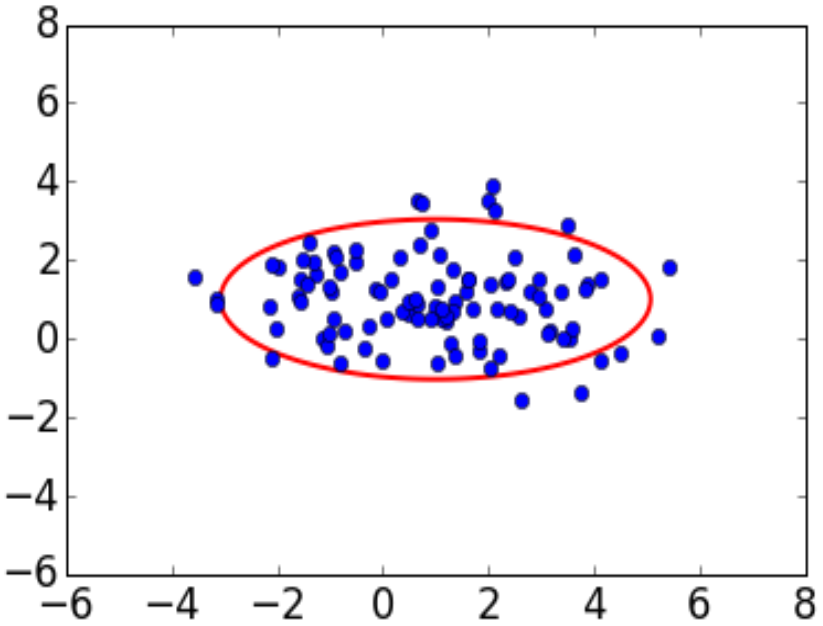
$$\text{Density} \quad p(x, y) = \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}\begin{bmatrix} x - \mu_x \\ y - \mu_y \end{bmatrix}^T \Sigma^{-1} \begin{bmatrix} x - \mu_x \\ y - \mu_y \end{bmatrix}\right)$$

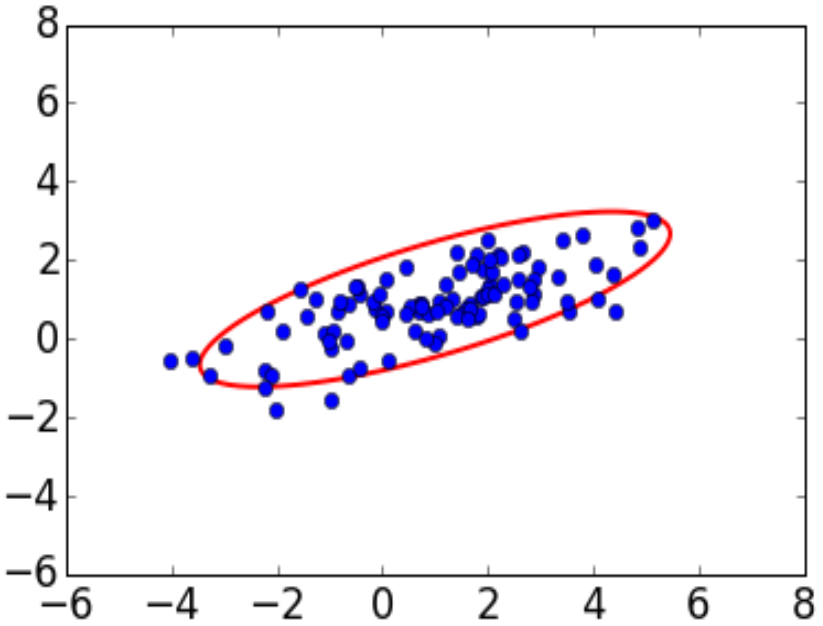
The density is highest at the mean, and falls off in ellipsoidal contours.



# Bivariate Gaussian: examples

In either case, the mean is (1, 1).

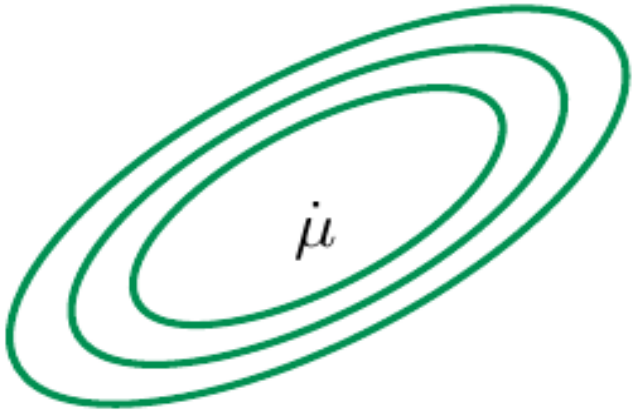



$$\Sigma = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 4 & 1.5 \\ 1.5 & 1 \end{bmatrix}$$

# The multivariate Gaussian




$N(\mu, \Sigma)$: Gaussian in $\mathbb{R}^p$

- mean: $\mu \in \mathbb{R}^p$
- covariance: $p \times p$ matrix $\Sigma$

$$\text{Density} \quad p(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

Let $X = (X_1, X_2, \ldots, X_p)$ be a random draw from $N(\mu, \Sigma)$.

- $\mu$ is the vector of coordinate-wise means:

$$\mu_1 = \mathbb{E}X_1, \ \mu_2 = \mathbb{E}X_2, \ldots, \ \mu_p = \mathbb{E}X_p.$$

- $\Sigma$ is a matrix containing all pairwise covariances:

$$\Sigma_{ij} = \Sigma_{ji} = \text{cov}(X_i, X_j) \quad \text{if } i \neq j$$
$$\Sigma_{ii} = \text{var}(X_i)$$

- In matrix/vector form: $\mu = \mathbb{E}X$ and $\Sigma = \mathbb{E}[(X-\mu)(X-\mu)^T]$.

# Special case: spherical Gaussian

The $X_i$ are independent and all have the same variance $\sigma^2$. Thus

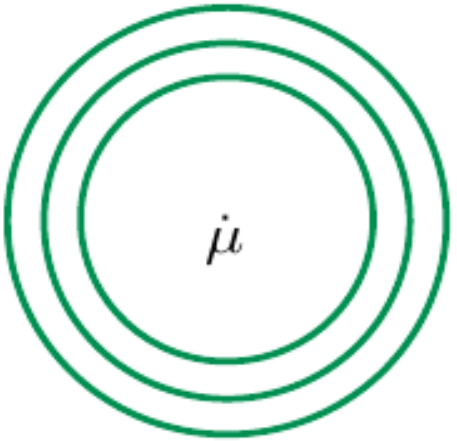
$$\Sigma = \sigma^2 I_p = \text{diag}(\sigma^2, \sigma^2, \ldots, \sigma^2)$$

(off-diagonal elements zero, diagonal elements $\sigma^2$).

Each $X_i$ is an independent univariate Gaussian $N(\mu_i, \sigma^2)$:

$$\Pr(x) = \prod_{i=1}^{p} \left( \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i-\mu_i)^2/2\sigma^2} \right) = \frac{1}{(2\pi)^{p/2}\sigma^p} \exp\left( -\frac{\|x-\mu\|^2}{2\sigma^2} \right)$$

Density at a point depends only on its distance from $\mu$:



# Special case: diagonal Gaussian

The $X_i$ are independent, with variances $\sigma_i^2$. Thus

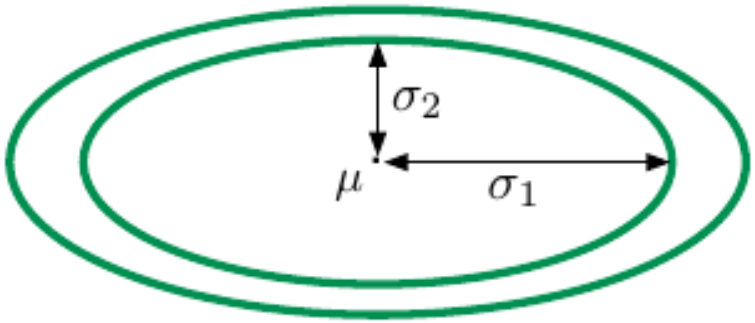
$$\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$$

(all off-diagonal elements zero).

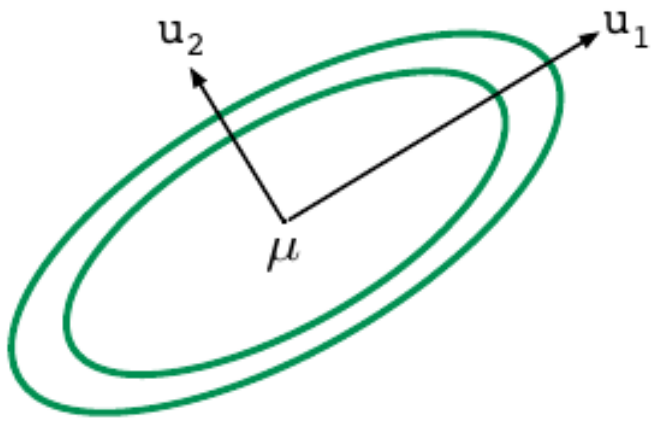
Each $X_i$ is an independent univariate Gaussian $N(\mu_i, \sigma_i^2)$:

$$p(x) = \frac{1}{(2\pi)^{p/2}\sigma_1 \cdots \sigma_p} \exp\left(-\sum_{i=1}^{p} \frac{(x_i - \mu_i)^2}{2\sigma_i^2}\right)$$

Contours of equal density are axis-aligned ellipsoids centered at $\mu$:



# The general Gaussian $N(\mu, \Sigma)$ in $\mathbb{R}^p$




Eigendecomposition of $\Sigma$ yields:

- **Eigenvalues**
  $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$
- Corresponding **eigenvectors**
  $u_1, \ldots, u_p$

Recall density: $$p(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}\underbrace{(x-\mu)^T\Sigma^{-1}(x-\mu)}_{\text{What is this?}}\right)$$

If we write $S = \Sigma^{-1}$ then $S$ is a $p \times p$ matrix and

$$(x-\mu)^T\Sigma^{-1}(x-\mu) = \sum_{i,j} S_{ij}(x_i-\mu_i)(x_j-\mu_j),$$

a **quadratic function** of $x$.

# Binary classification with Gaussian generative model

Estimate class probabilities $\pi_1, \pi_2$ and fit a Gaussian to each class:

$$P_1 = N(\mu_1, \Sigma_1), P_2 = N(\mu_2, \Sigma_2)$$

E.g. If data points $x^{(1)}, \ldots, x^{(m)} \in \mathbb{R}^p$ are class 1:

$$\mu_1 = \frac{1}{m}\left(x^{(1)} + \cdots + x^{(m)}\right) \quad \text{and} \quad \Sigma_1 = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)} - \mu_1)(x^{(i)} - \mu_1)^T$$

Given a new point $x$, predict class 1 iff:

$$\pi_1 P_1(x) > \pi_2 P_2(x) \quad \Leftrightarrow \quad x^T M x + 2w^T x \geq \theta,$$

where:

$$M = \frac{1}{2}(\Sigma_2^{-1} - \Sigma_1^{-1})$$
$$w = \Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2$$

and $\theta$ is a constant depending on the various parameters.

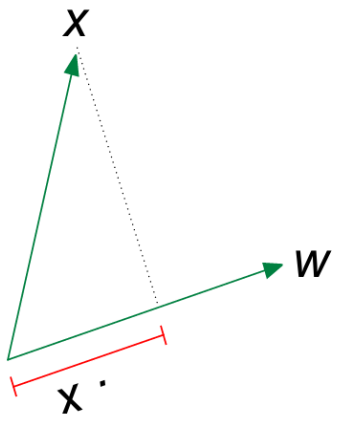
$\Sigma_1 = \Sigma_2$: linear decision boundary. Otherwise, quadratic boundary.

# Linear decision boundary

When $\Sigma_1 = \Sigma_2 = \Sigma$: choose class 1 iff

$$x \cdot \underbrace{\Sigma^{-1}(\mu_1 - \mu_2)}_{w} \geq \theta.$$

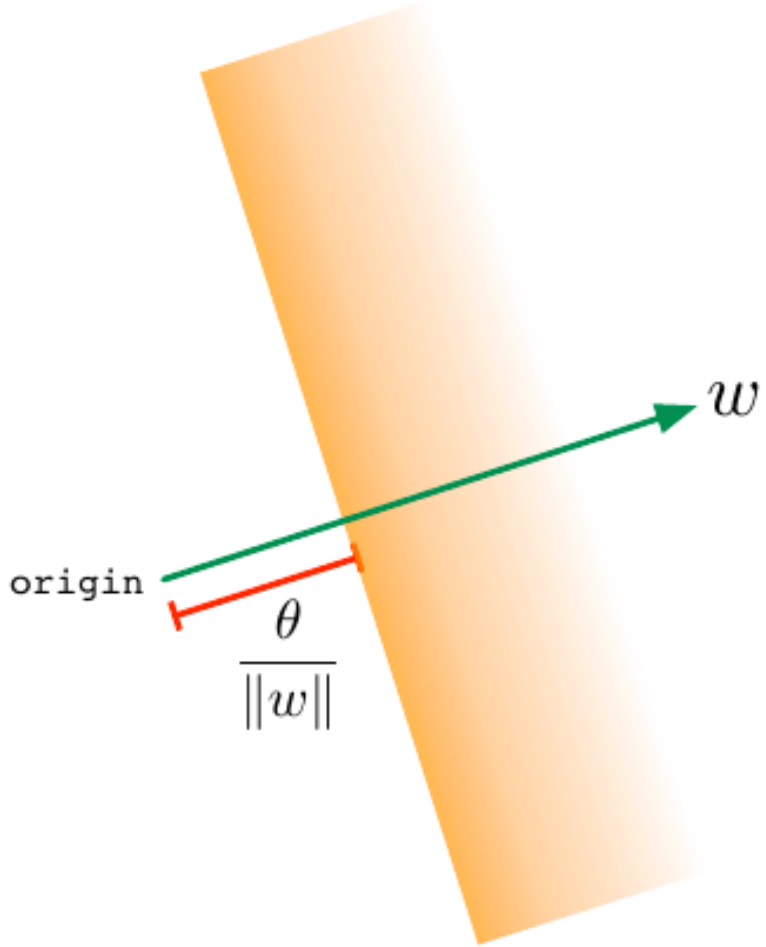
Geometric picture: Suppose *w* is a unit vector (that is, ||*w*|| = 1). Then *x* · *w* is the **projection** of vector *x* onto direction *w* .




And we can always make *w* a unit vector by dividing both sides of the inequality by $\|w\|$.

# Linear decision boundary

Let $w$ be any vector in $\mathbb{R}^p$. What is meant by decision rule $w \cdot x \geq \theta$ ?

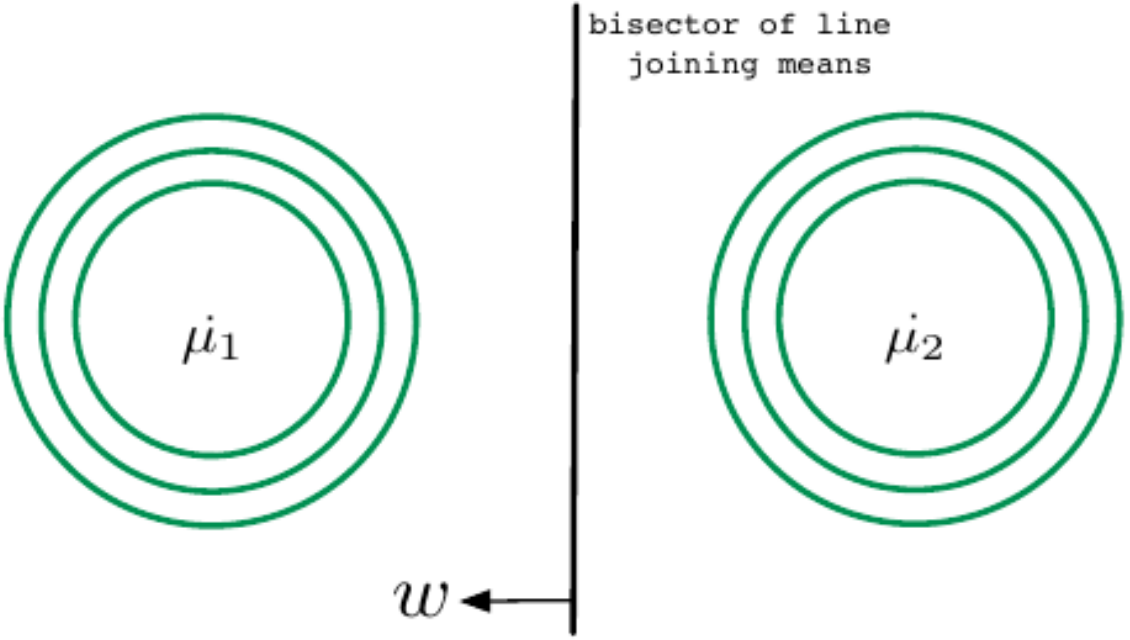


# Common covariance: $\Sigma_1 = \Sigma_2 = \Sigma$

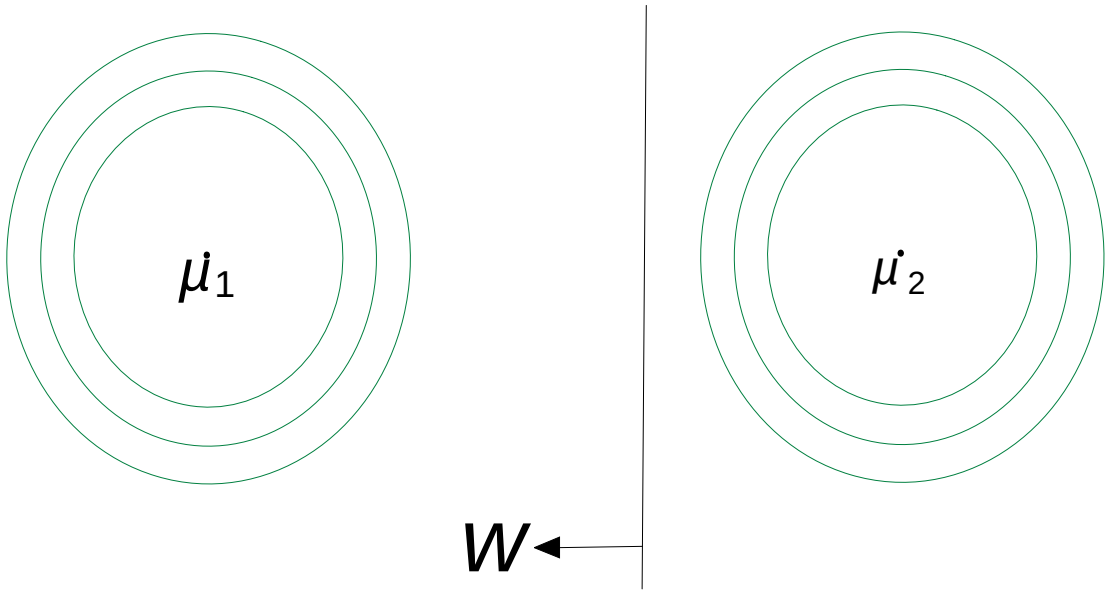
Linear decision boundary: choose class 1 iff

$$x \cdot \underbrace{\Sigma^{-1}(\mu_1 - \mu_2)}_{w} \geq \theta.$$

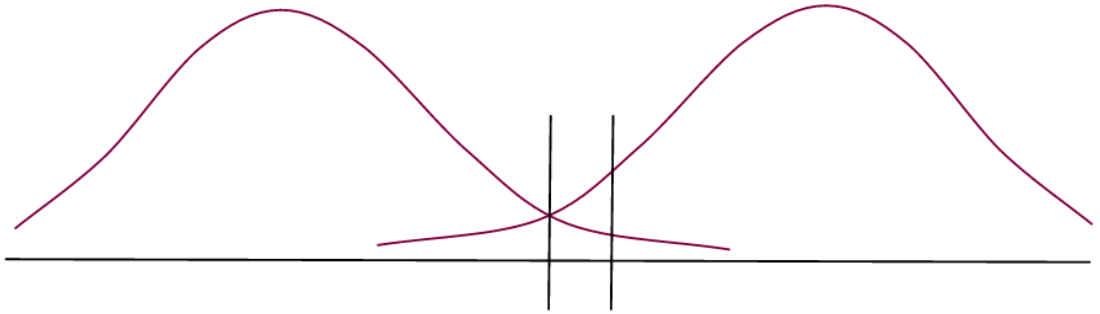
Example 1: Spherical Gaussians with $\Sigma = I_p$ and $\pi_1 = \pi_2$.



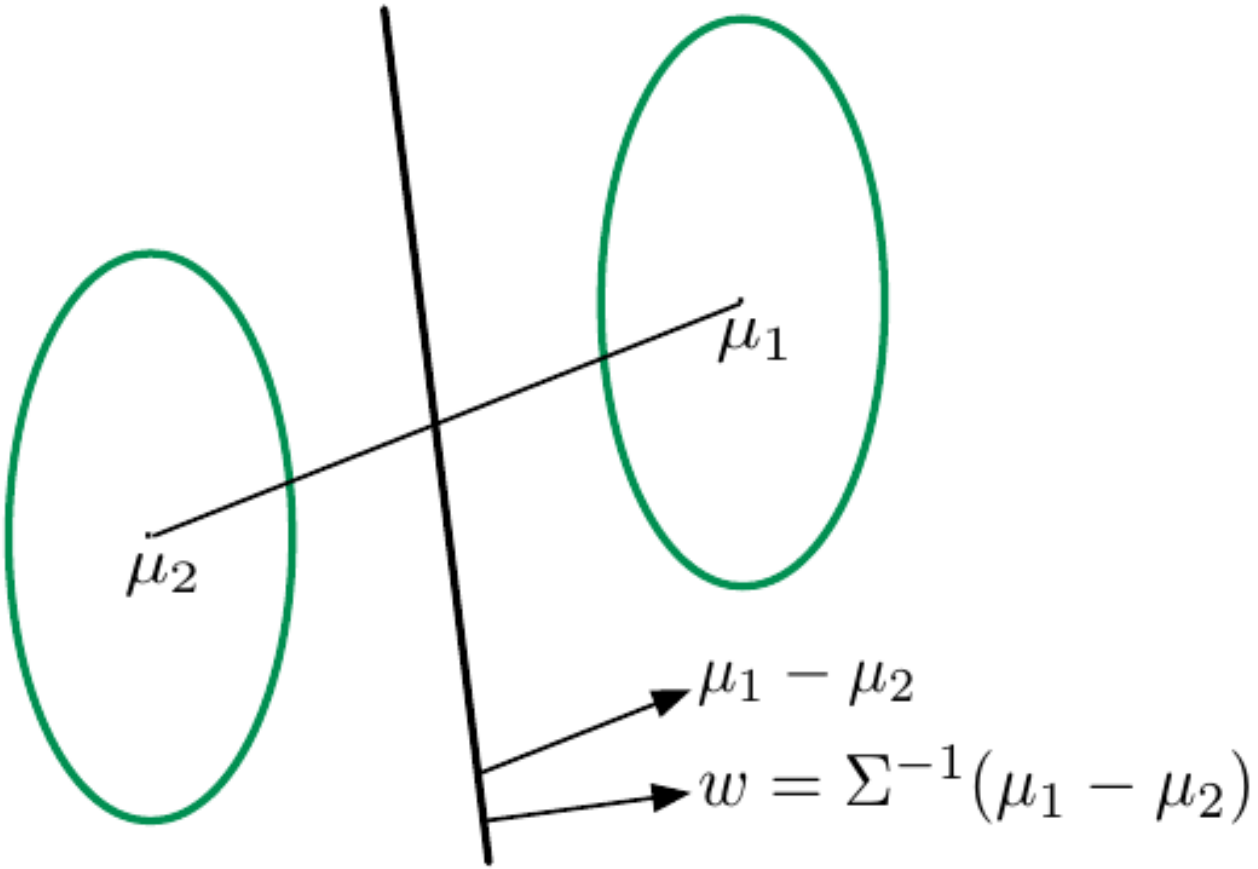
Example 2: Again spherical, but now $\pi_1 > \pi_2$ .




One-d projection onto *w* :



## Example 3: Non-spherical.




Rule: $w \cdot x \geq \theta$

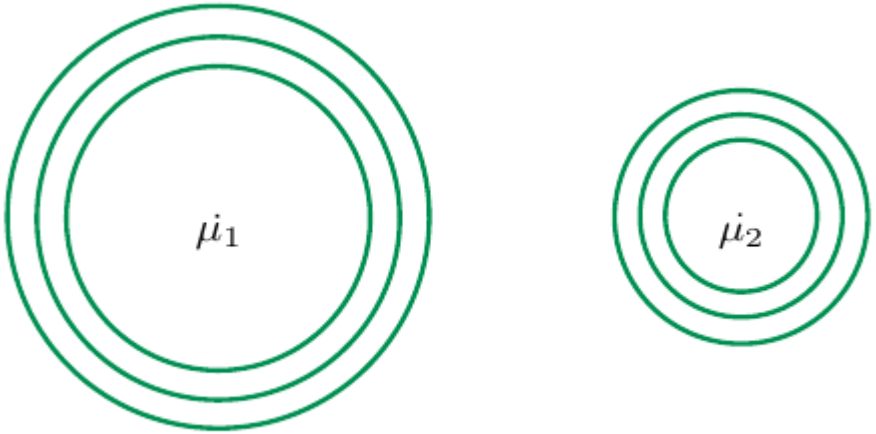
- $w, \theta$ dictated by probability model, assuming it is a perfect fit
- Common practice: choose $w$ as above, but fit $\theta$ to minimize training/validation error

# Different covariances: $\Sigma_1 \neq \Sigma_2$

Quadratic boundary: choose class 1 iff $x^T Mx + 2w^T x \geq \theta$, where:

$$M = \frac{1}{2}(\Sigma_2^{-1} - \Sigma_1^{-1})$$
$$w = \Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2$$

Example 1: $\Sigma_1 = \sigma_1^2 I_p$ and $\Sigma_2 = \sigma_2^2 I_p$ with $\sigma_1 > \sigma_2$



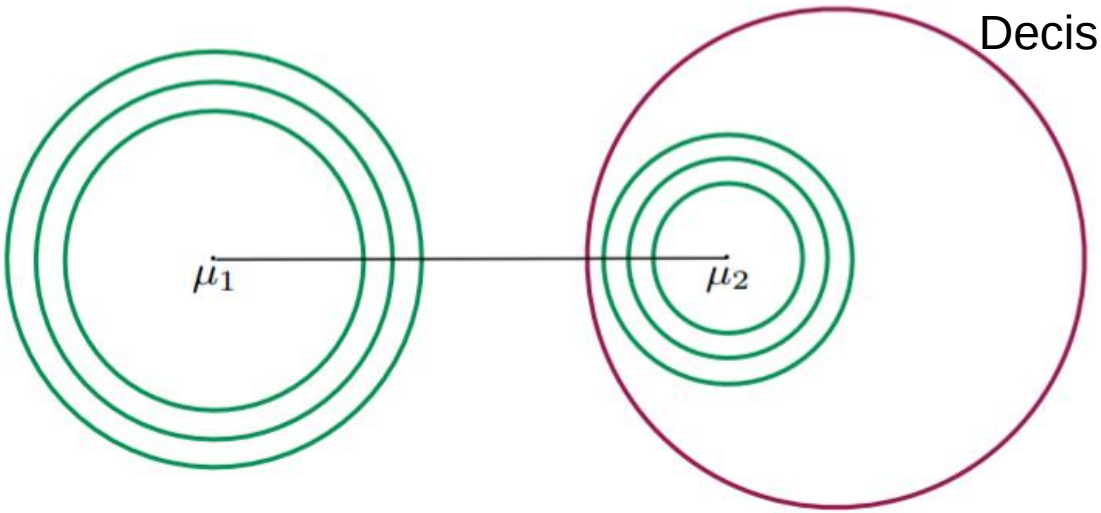
# Different covariances: $\Sigma_1 \neq \Sigma_2$

Quadratic boundary: choose class 1 iff $x^T Mx + 2w^T x \geq \theta$, where:

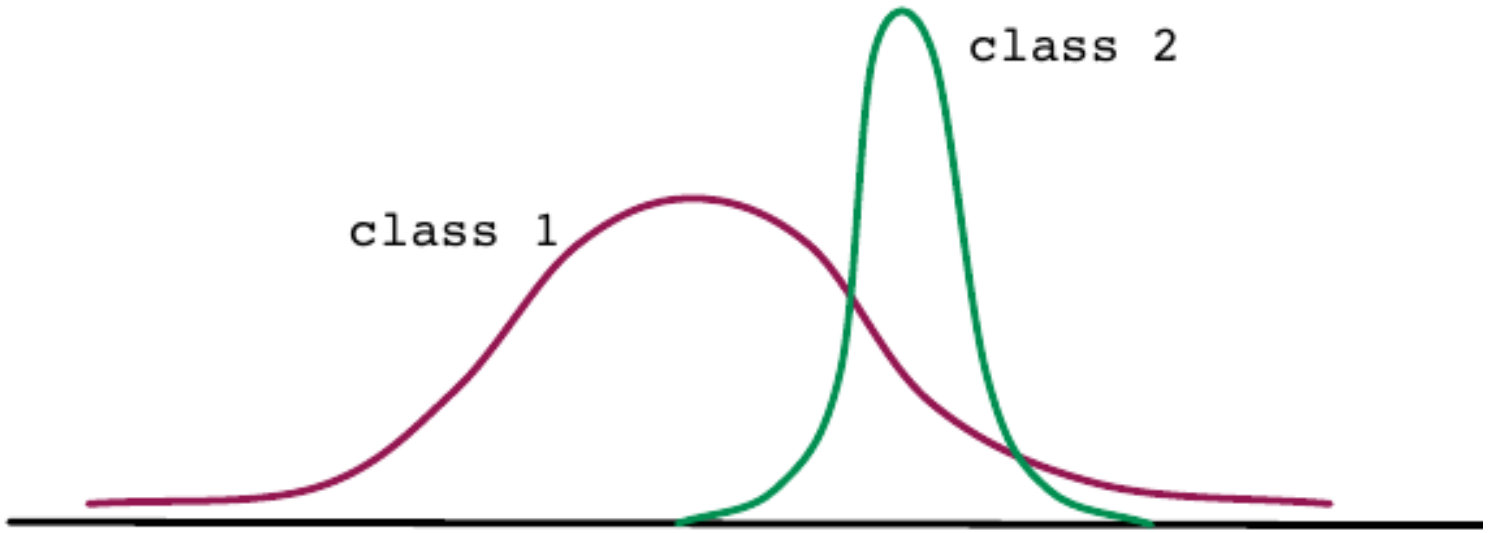
$$M = \frac{1}{2}(\Sigma_2^{-1} - \Sigma_1^{-1})$$

$$w = \Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2$$

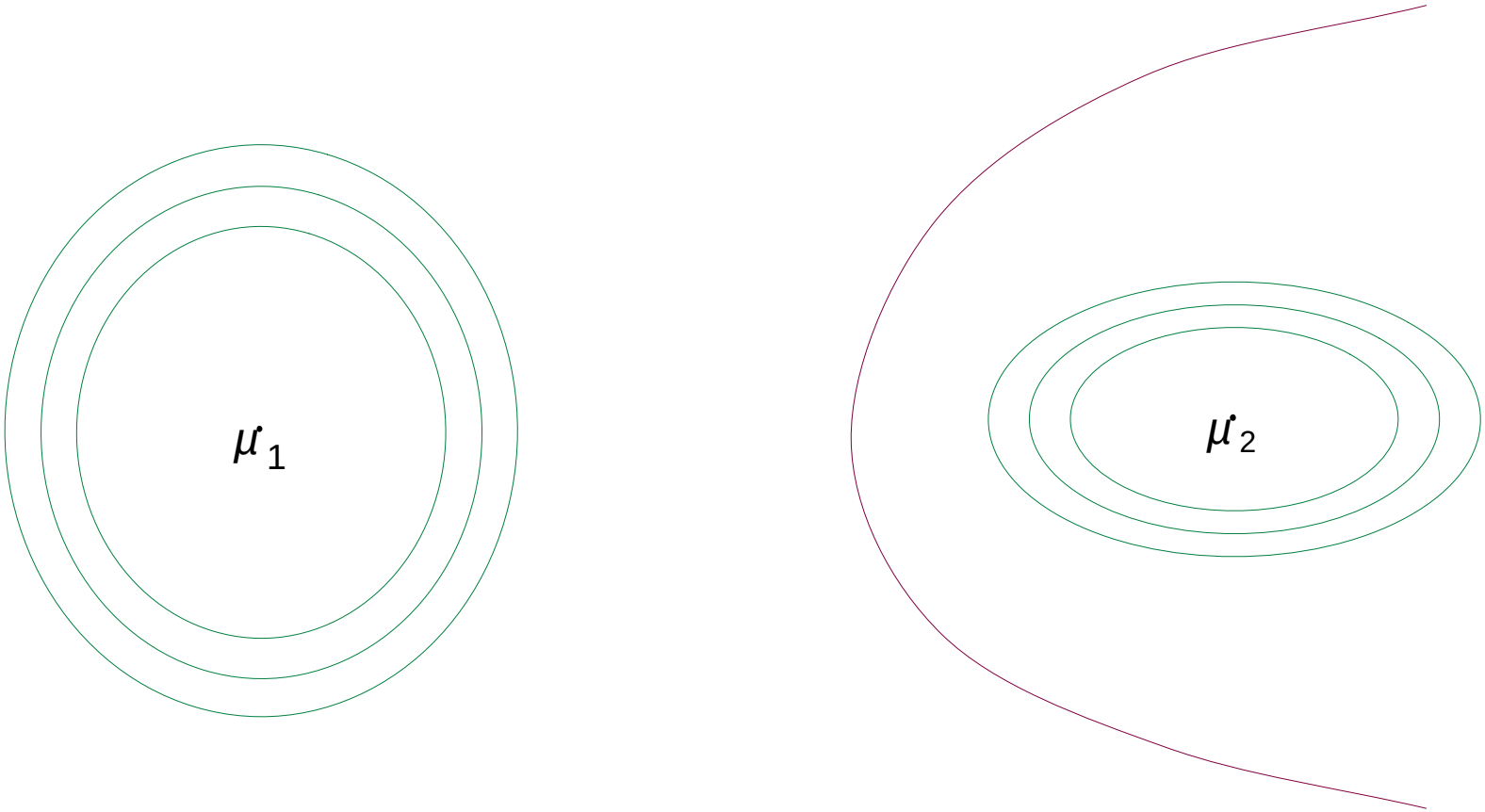
Example 1: $\Sigma_1 = \sigma_1^2 I_p$ and $\Sigma_2 = \sigma_2^2 I_p$ with $\sigma_1 > \sigma_2$



Example 2: Same thing in 1-d. $\mathcal{X} = \mathbb{R}$.



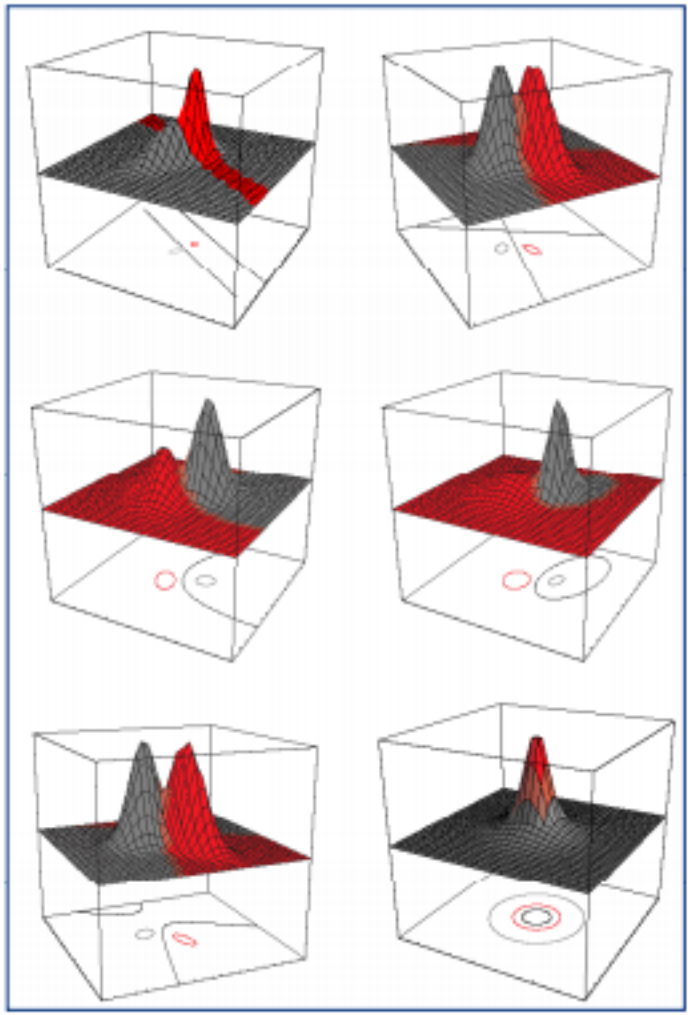
Example 3: A parabolic boundary.

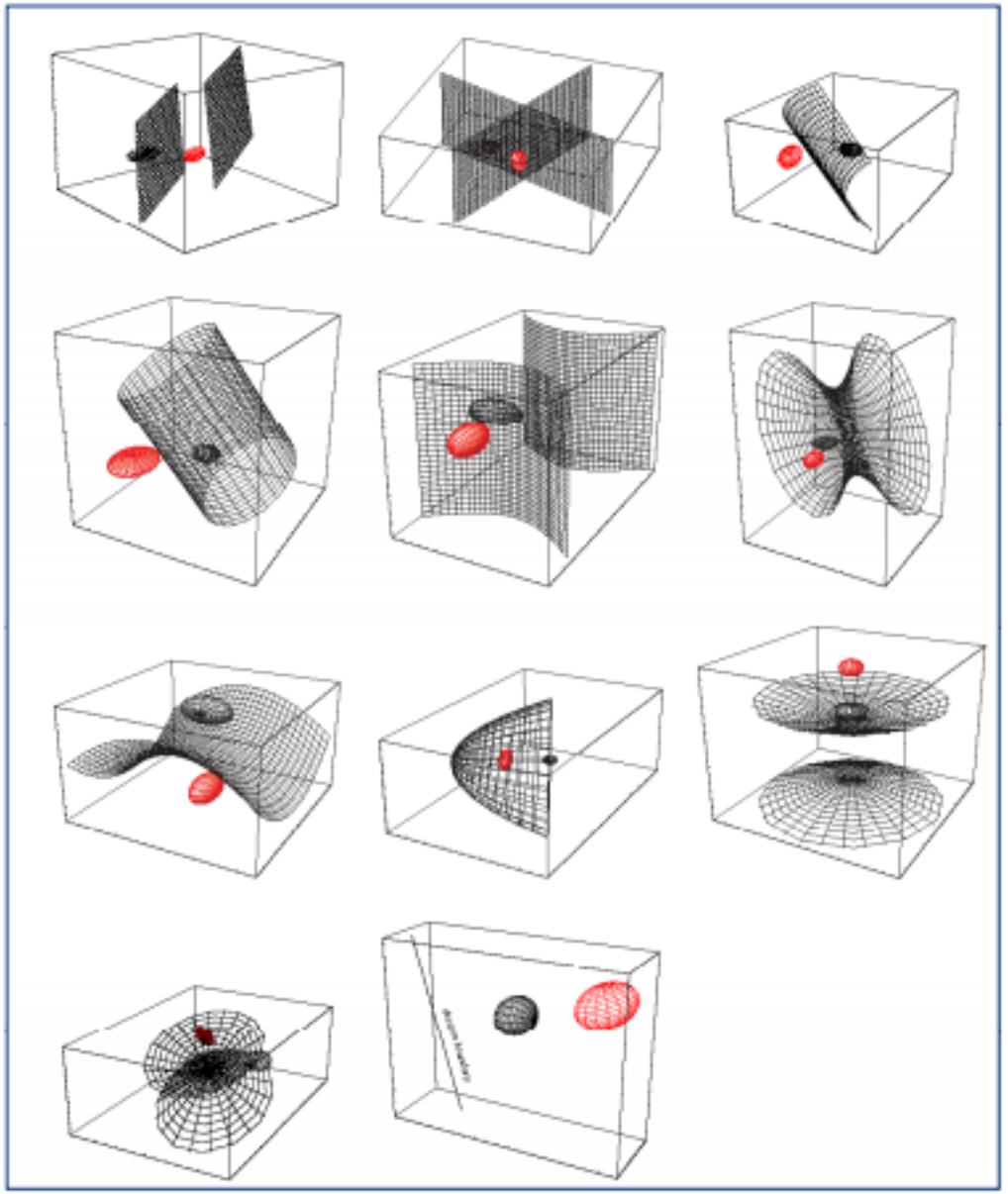



Many other possibilities!

# Geometric interpretation

- in 2 and 3D:





# Hands-on and Self practice

HW4 - Question 1 to 7

# Multiclass discriminant analysis

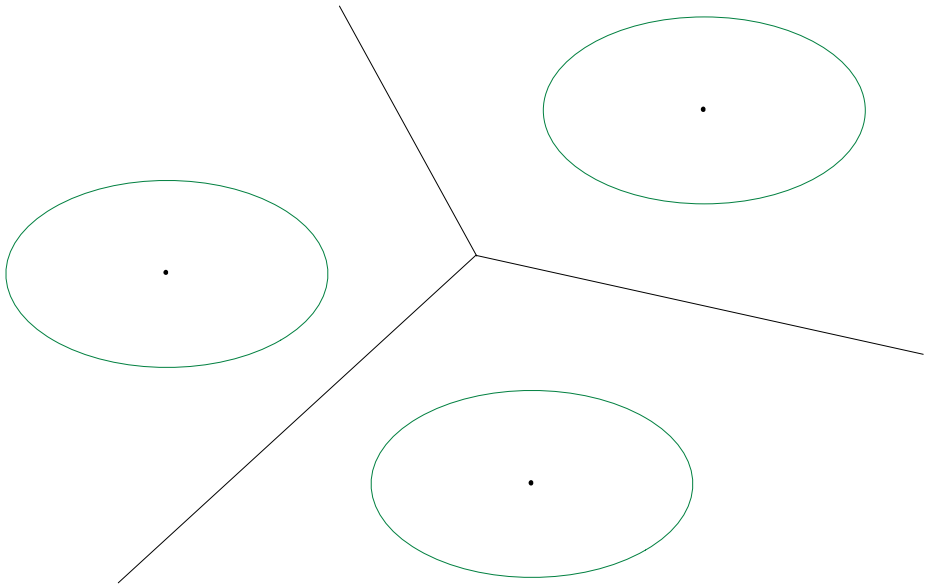
$k$ classes: weights $\pi_j$, class-conditional distributions $P_j = N(\mu_j , \Sigma_j)$

Each class has an associated **quadratic** function

$$f_j(x) = \log (\pi_j P_j (x ))$$

To classify a point $x$ , pick $\arg_j \max f_j(x)$.

If $\Sigma_1 = \cdot \cdot \cdot = \Sigma_k$, the boundaries are **linear**.

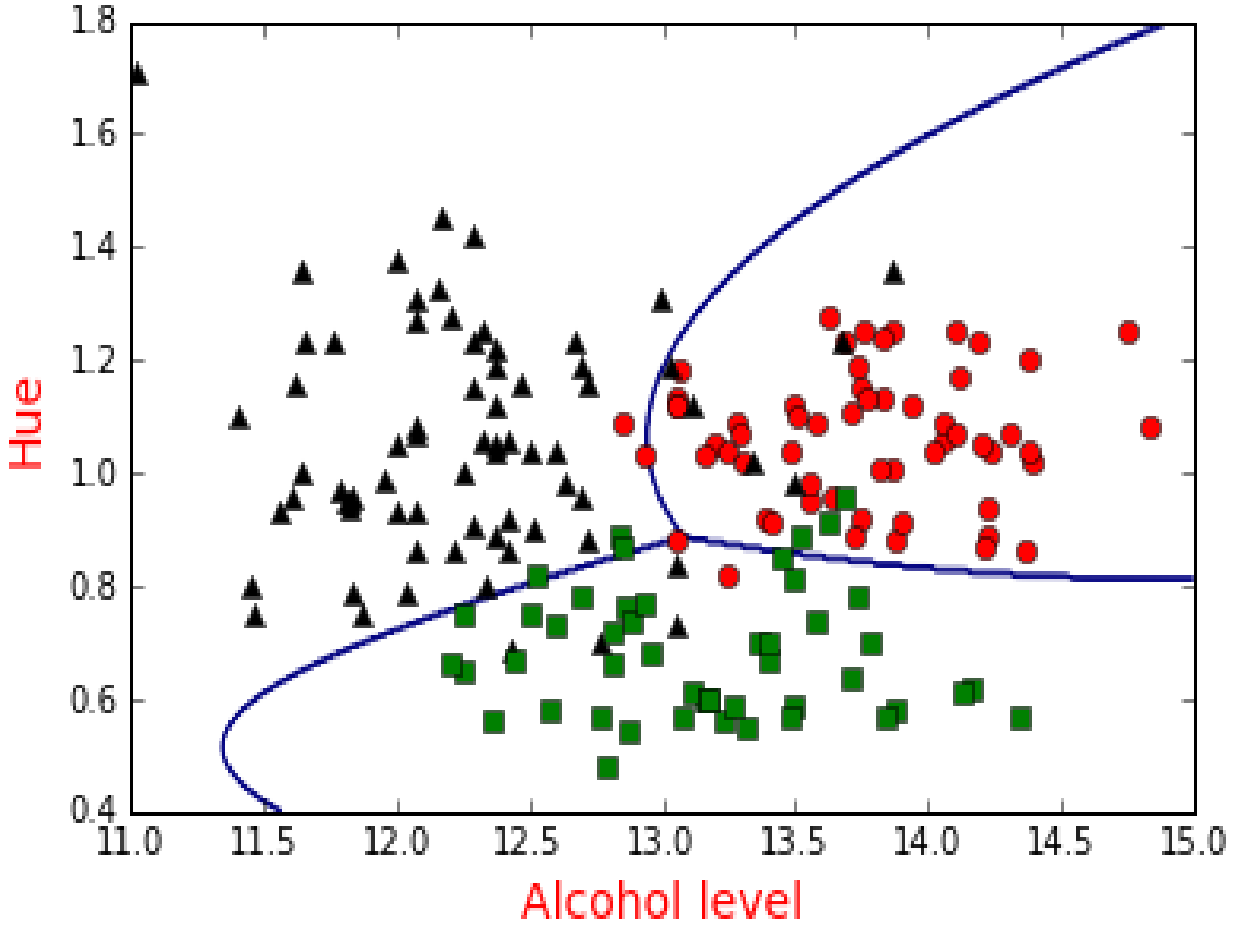


# Example: "wine" data set

Data from three wineries from the same region of Italy

- 13 attributes: hue, color intensity, flavanoids, ash content, ...
- 178 instances in all: split into 118 train, 60 test

Test error using multiclass discriminant analysis: 1/60



# Example: MNIST

To each digit, fit:

- class probability $\pi_j$
- mean $\mu_j \in \mathrm{R}^{784}$
- covariance matrix $\Sigma_j \in \mathrm{R}^{784 \times 784}$

Problem: formula for normal density uses $\Sigma_j^{-1}$, which is singular.

- Need to regularize: $\Sigma_j \rightarrow \Sigma_j + \sigma^2 I$
- This is a good idea even without the singularity issue

# Fisher's linear discriminant

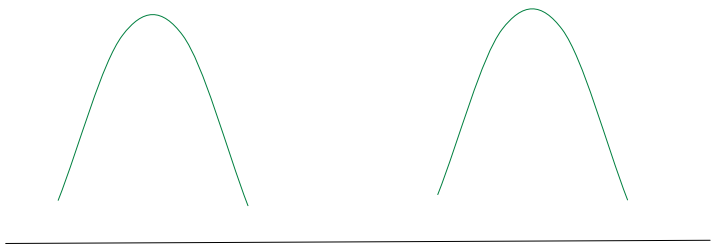
A framework for linear classification without Gaussian assumptions.

Use only first- and second-order statistics of the classes.

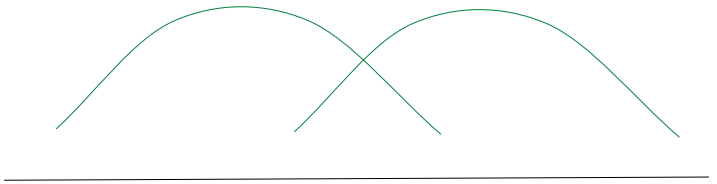
| Class 1 | Class 2 |
|---|---|
| mean $\mu_1$ | mean $\mu_2$ |
| cov $\Sigma_1$ | cov $\Sigma_2$ |
| # pts $n_1$ | # pts $n_2$ |

A linear classifier projects all data onto a direction $w$ . Choose $w$ so that:

- Projected means are well-separated, i.e. $(w \cdot \mu_1 - w \cdot \mu_2)^2$ is large.
- Projected within-class variance is small.



better than



# Fisher LDA (linear discriminant analysis)

Two classes: means $\mu_1, \mu_2$; covariances $\Sigma_1, \Sigma_2$; sample sizes $n_1, n_2$.

Project data onto direction (unit vector) $w$.

- Projected means: $w \cdot \mu_1$ and $w \cdot \mu_2$
- Projected variances: $w^T \Sigma_1 w$ and $w^T \Sigma_2 w$
- Average projected variance:

$$\frac{n_1(w^T \Sigma_1 w) + n_2(w^T \Sigma_2 w)}{n_1 + n_2} = w^T \Sigma w,$$

where $\Sigma = (n_1 \Sigma_1 + n_2 \Sigma_2)/(n_1 + n_2)$.

Find $w$ to maximize $J(w) = \dfrac{(w \cdot \mu_1 - w \cdot \mu_2)^2}{w^T \Sigma w}$

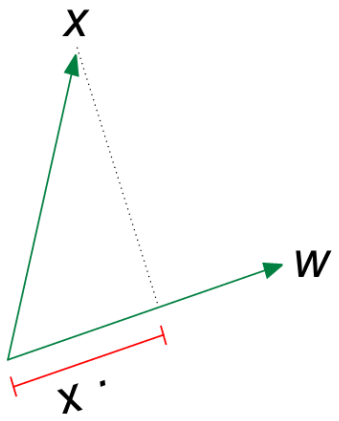
Solution: $w \propto \Sigma^{-1}(\mu_1 - \mu_2)$. Look familiar?

# Recall: Linear decision boundary

When $\Sigma_1 = \Sigma_2 = \Sigma$: choose class 1 iff

$$x \cdot \underbrace{\Sigma^{-1}(\mu_1 - \mu_2)}_{w} \geq \theta.$$

Geometric picture: Suppose *w* is a unit vector (that is, ||*w*|| = 1). Then $x \cdot w$ is the **projection** of vector *x* onto direction *w* .




And we can always make *w* a unit vector by dividing both sides of the inequality by $\|w\|$.

# Fisher LDA: proof

Goal: find $w$ to maximize $J(w) = \frac{(w \cdot \mu_1 - w \cdot \mu_2)^2}{w^T \Sigma w}$

1. Assume $\Sigma_1$, $\Sigma_2$ are full rank; else project.
2. Since $\Sigma_1$ and $\Sigma_2$ are p.d., so is their weighted average, $\Sigma$.
3. Write $u = \Sigma^{1/2} w$. Then

$$\max_w \frac{(w^T(\mu_1 - \mu_2))^2}{w^T \Sigma w} = \max_u \frac{(u^T \Sigma^{-1/2}(\mu_1 - \mu_2))^2}{u^T u}$$
$$= \max_{u:\|u\|=1} (u \cdot (\Sigma^{-1/2}(\mu_1 - \mu_2)))^2$$

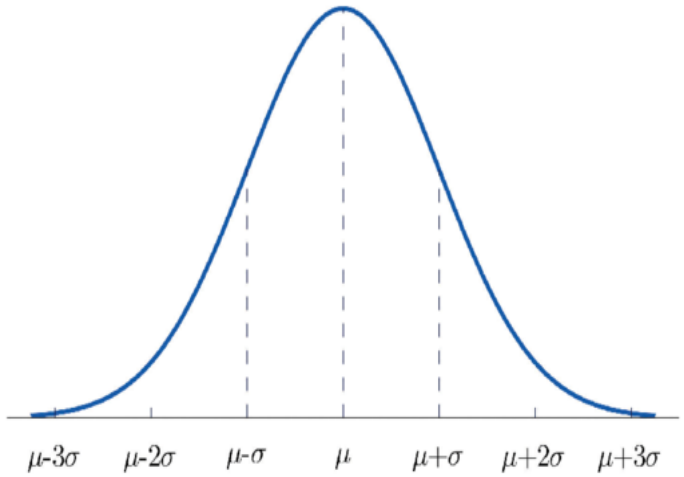
4. Solution: $u$ is the unit vector in direction $\Sigma^{-1/2}(\mu_1 - \mu_2)$.
5. Therefore: $w = \Sigma^{-1/2} u \propto \Sigma^{-1}(\mu_1 - \mu_2)$.

# Beyond Gaussians

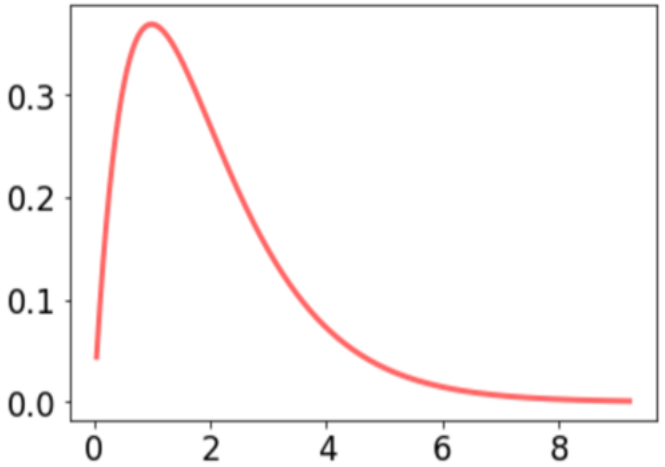
The generative methodology:

- Fit a **distribution** to each class separately
- Use Bayes' rule to classify new data

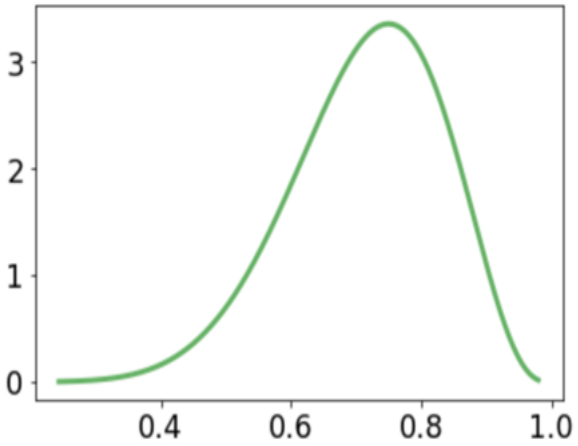
What distribution to use? Are Gaussians enough?



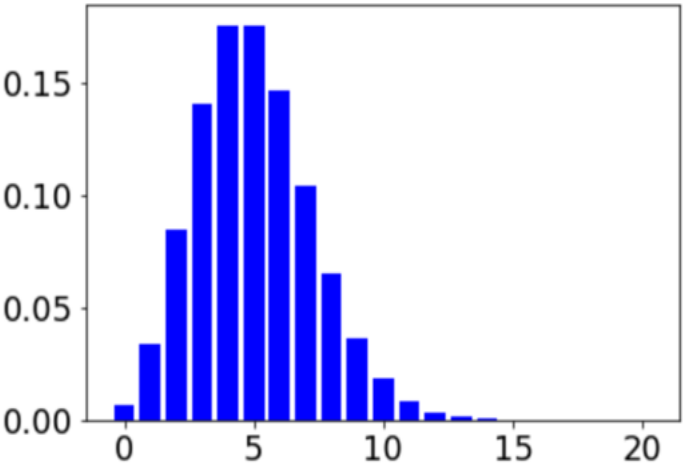
# Exponential families of distributions




**GAMMA**




**BETA**




**POISSON**

It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair, we had everything before us, we had nothing before us, we were all going direct to Heaven, we were all going direct the other way – in short, the period was so far like the present period, that some of its noisiest authorities insisted on its being received, for good or for evil, in the superlative degree of comparison only.

| | |
|---|---|
| 1 | despair |
| 2 | evil |
| 0 | happiness |
| 1 | foolishness |

**CATEGORICAL**

# Hands-on and Self practice

HW4 - Question 8 and 9

# Appendix

# Proof: Decision Rule For Different Covariances

- the optimal decision rule can be written as
  - 1) $i^*(x) = \arg\max_i P_{Y|X}(i \mid x)$
  - 2) $i^*(x) = \arg\max_i \left[ P_{X|Y}(x \mid i) P_Y(i) \right]$
  - 3) $i^*(x) = \arg\max_i \left[ \log P_{X|Y}(x \mid i) + \log P_Y(i) \right]$
- we have started to study the case of Gaussian classes

$$P_{X|Y}(x \mid i) = \frac{1}{\sqrt{(2\pi)^d \mid \Sigma_i \mid}} \exp\left\{ -\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1} (x-\mu_i) \right\}$$

# Proof:

## The Gaussian classifier

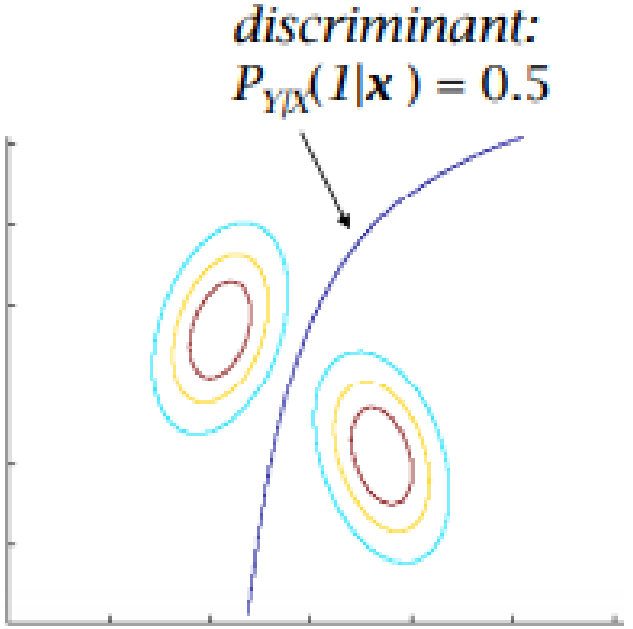
- the optimal decision rule can be written as

$$i^*(x) = \arg\min_i \left[ d_i(x, \mu_i) + \alpha_i \right]$$

with

$$d_i(x, y) = (x - y)^T \Sigma_i^{-1} (x - y)$$

$$\alpha_i = \log(2\pi)^d \left| \Sigma_i \right| - 2 \log P_Y(i)$$




- the optimal rule is to assign x to the closest class
- closest is measured with the Mahalanobis distance $d_i(x,y)$
- to which the $\alpha$ constant is added to account for the class prior

# Proof:

Let $g_i(x)$ be $\left[d_i(x,\mu_i)+\alpha_i\right]$

We can drop the term $(2\pi)^d$ since it does not affect the decision rule.

$$g_i(x)=(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)+\log\left|\Sigma_i\right|-2\log P_Y(i)$$

$$=x^T\Sigma_i^{-1}x-2x^T\Sigma_i^{-1}\mu_i+\mu_i^T\Sigma_i^{-1}\mu_i+\log\left|\Sigma_i\right|-2\log P_Y(i)$$

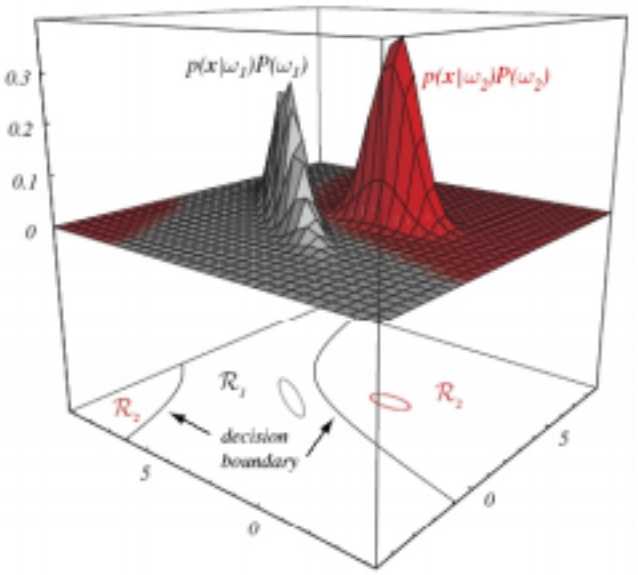
# Proof:

- and

$$g_i(x) = x^T \Sigma_i^{-1} x - 2x^T \Sigma_i^{-1} \mu_i + \mu_i^T \Sigma_i^{-1} \mu_i + \log|\Sigma_i| - 2\log P_Y(i)$$

  - which can be written as

$$g_i(x) = x^T W_i\, x + w_i^T x + w_{i0}$$
$$W_i = \Sigma_i^{-1}$$
$$w_i = -2\Sigma_i^{-1}\mu_i$$
$$w_{i0} = \mu_i^T \Sigma_i^{-1} \mu_i + \log|\Sigma_i| - 2\log P_Y(i)$$




- for 2 classes the decision boundary is hyper-quadratic
  - this could mean hyper-plane, pair of hyper-planes, hyper-spheres, hyper-elipsoids, hyper-hyperboloids, etc.

- The decision rule gets the form of:

$$x^T M x + 2w^T x \geq \theta$$