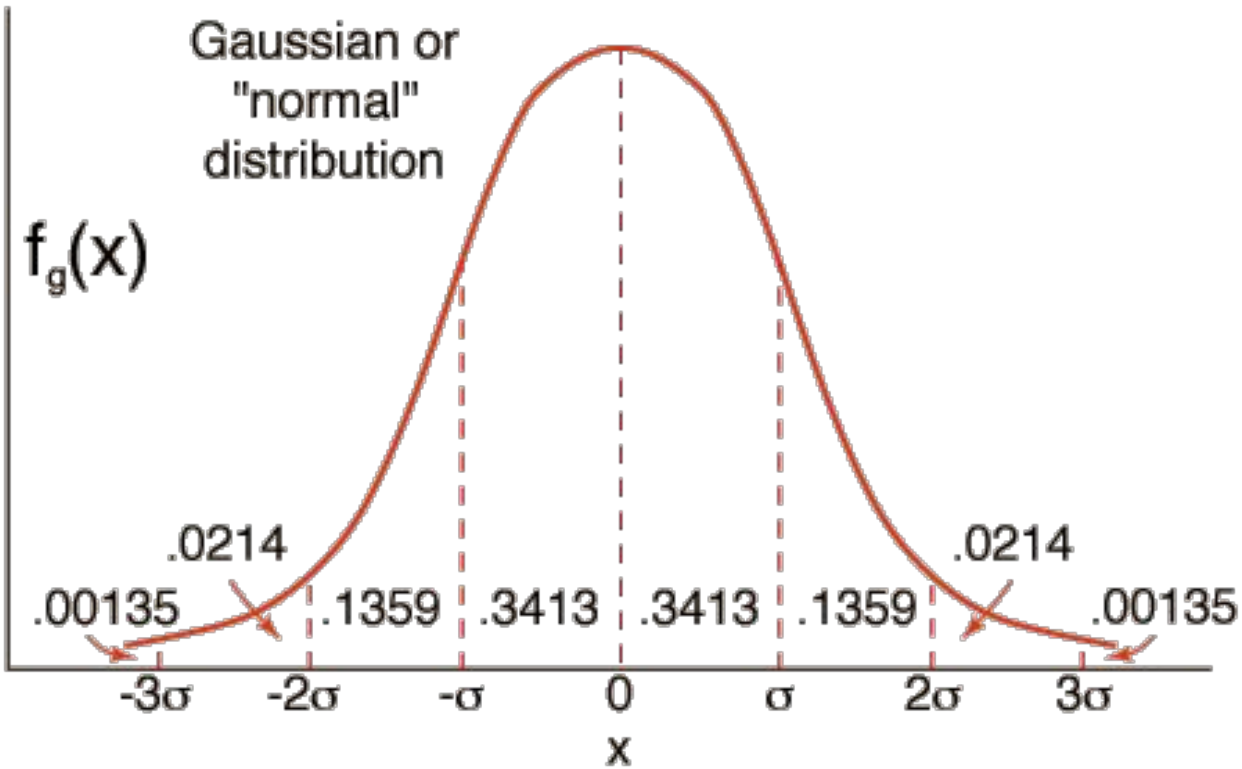
Classification with Generative Models (2)

MGTF 495

Class Outline

- Parametric Methods
  - Generative Models
    - Naive Bayes
    - Binary Features, Multinomial Features
    - *Hands-On*
    - Gaussian Generative Model
    - Fisher Linear Discriminant Analysis
    - *Hands-On*

The univariate Gaussian



The Gaussian $N(\mu, \sigma^2)$ has mean μ, variance σ^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

Bivariate distributions

Simplest option: treat each variable as independent.

Example: For a large collection of people, measure the two variables

$$H = \text{height}$$
$$W = \text{weight}$$

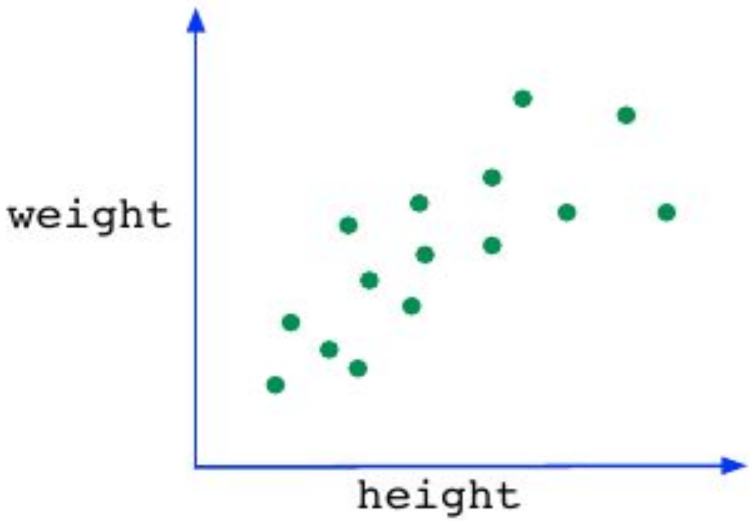
Independence would mean

$$\Pr(H = h,\ W = w) = \Pr(H = h)\Pr(W = w),$$

which would also imply $\mathbf{E}(HW) = \mathbf{E}(H)\,\mathbf{E}(W)$.

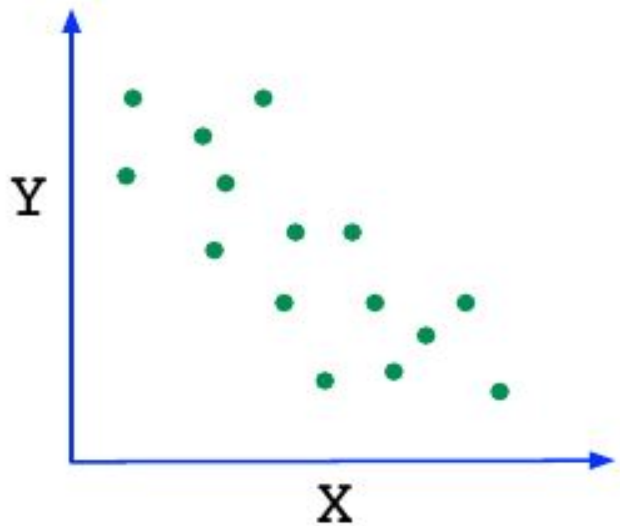
Is this an accurate approximation?
No: we'd expect height and weight to be **positively correlated**.

Types of correlation

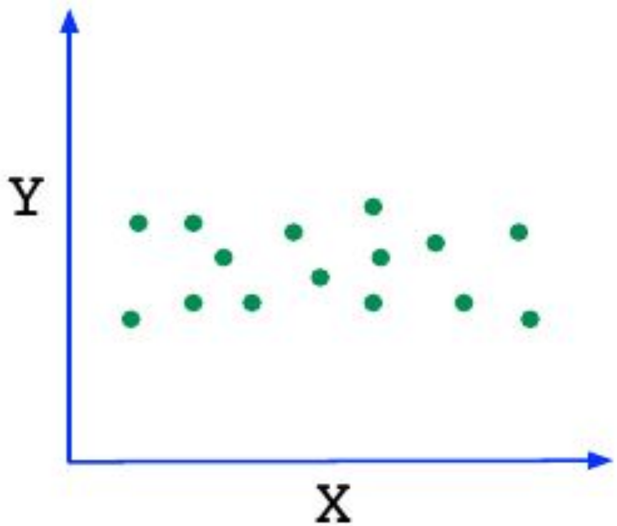


H, W positively correlated. This also implies

$$\mathbb{E}(HW) > \mathbb{E}(H)\mathbb{E}(W).$$

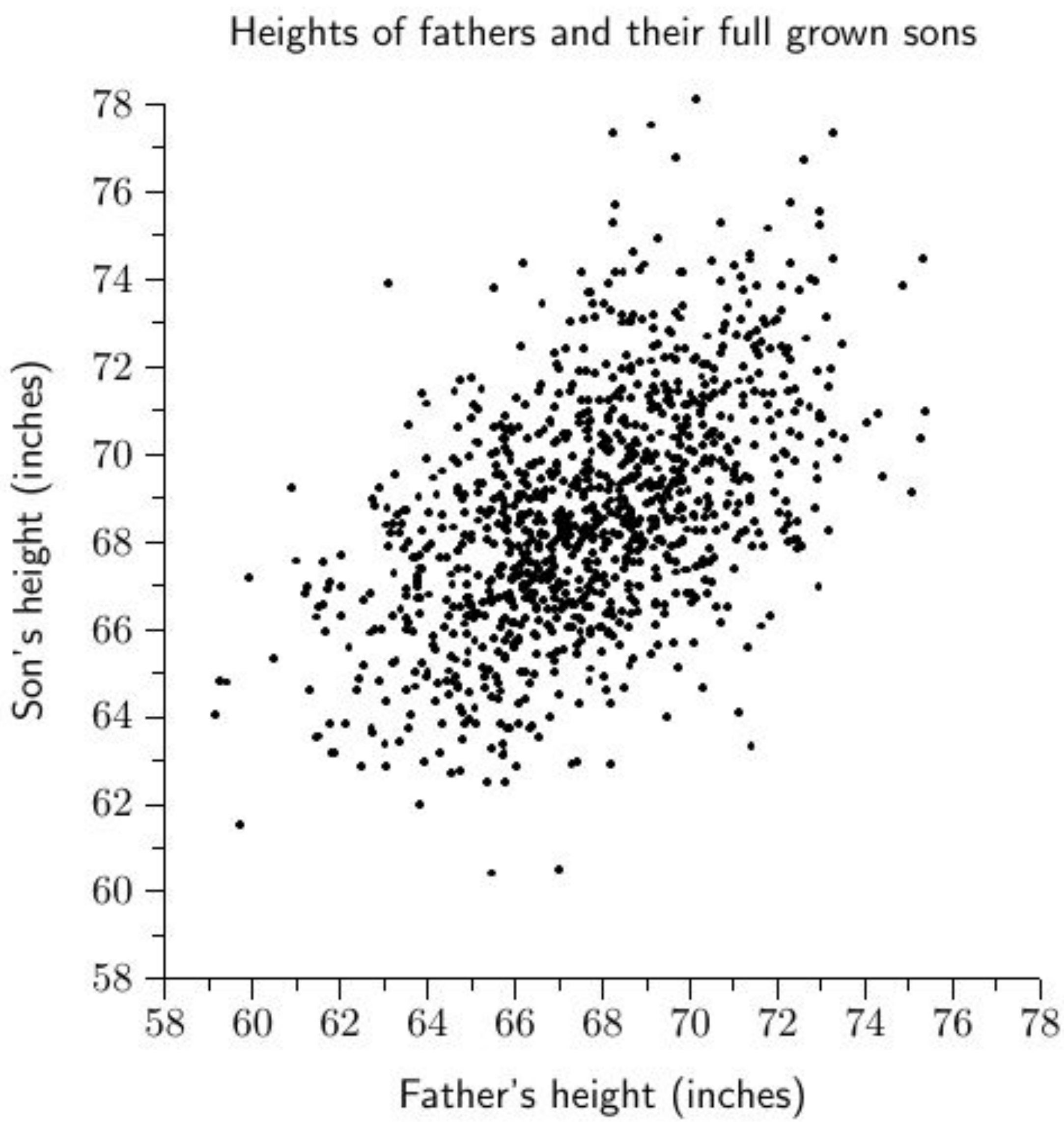


X, Y negatively correlated



X, Y uncorrelated

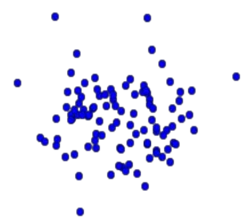
Pearson (1903): fathers and sons



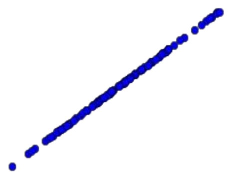
How to quantify the degree of correlation?

Correlation pictures

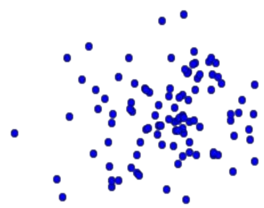
$r = 0$



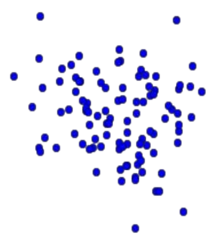
$r = 1$



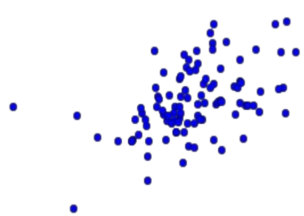
$r = 0.25$



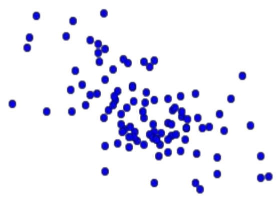
$r = -0.25$



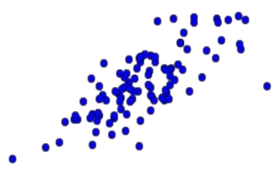
$r = 0.5$



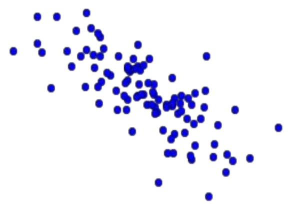
$r = -0.5$



$r = 0.75$



$r = -0.75$



Covariance and correlation

Suppose X has mean μ_x and Y has mean μ_Y

- Covariance

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

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

- Correlation

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

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

Question

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

Are X and Y independent?

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

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

$E[X] = 0$

$E[Y] = 2/3$

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

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

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

Question

Random variables X and Y are uncorrelated.

Are X and Y independent?

Question

Random variables X and Y are uncorrelated.

Are X and Y independent? No.

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

Question

Random variables X and Y are independent.

Can X and Y be correlated?

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

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

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

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

$$\begin{aligned} \mu_X &= 0 \\ \mu_Y &= -1/3 \\ \text{var}(X) &= 1 \\ \text{var}(Y) &= 8/9 \\ \text{cov}(X, Y) &= 0 \\ \text{corr}(X, Y) &= 0 \end{aligned}$$

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

Covariance and correlation: example 2

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

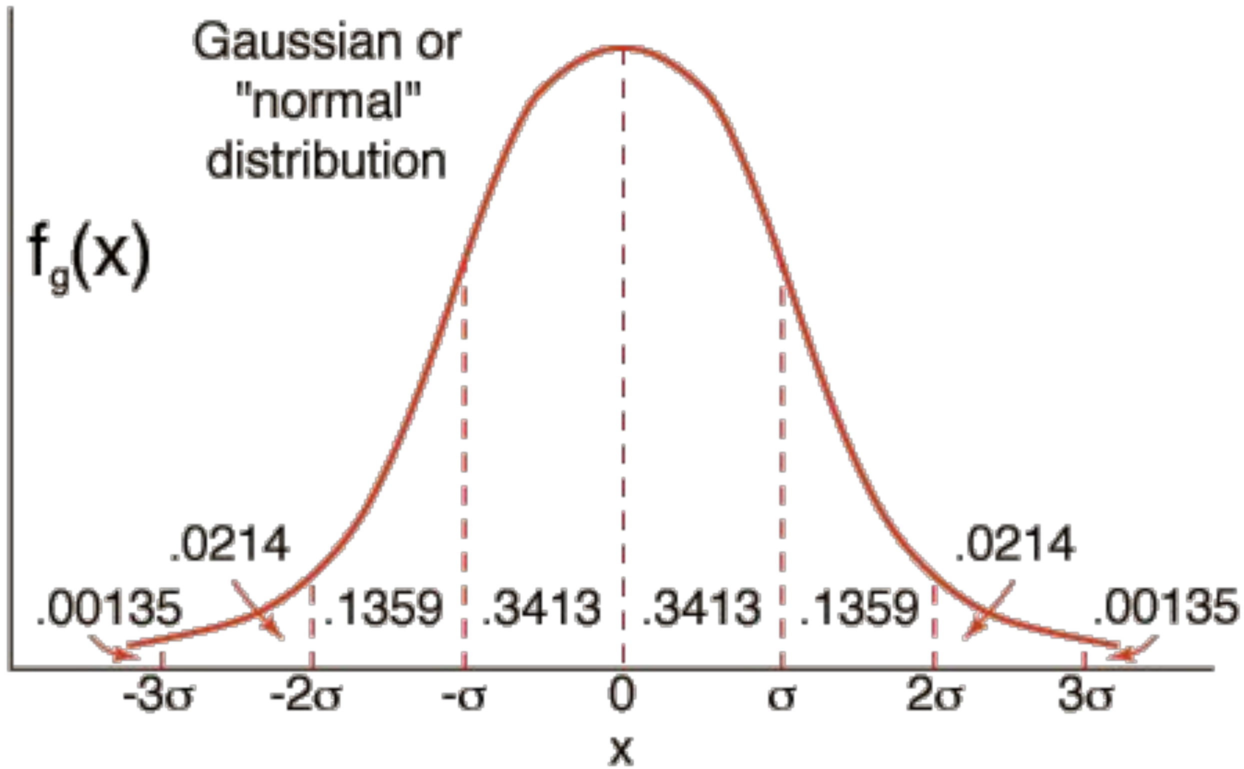
$$\mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\mathrm{std}(X)\mathrm{std}(Y)}$$

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

$$\begin{aligned}
\mu_X &= 0\\
\mu_Y &= 0\\
\mathrm{var}(X) &= 1\\
\mathrm{var}(Y) &= 100\\
\mathrm{cov}(X,Y) &= -10/3\\
\mathrm{corr}(X,Y) &= -1/3
\end{aligned}$$

In this case, X and Y are negatively correlated.

The univariate Gaussian



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

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

vector inverse

is a vector

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:

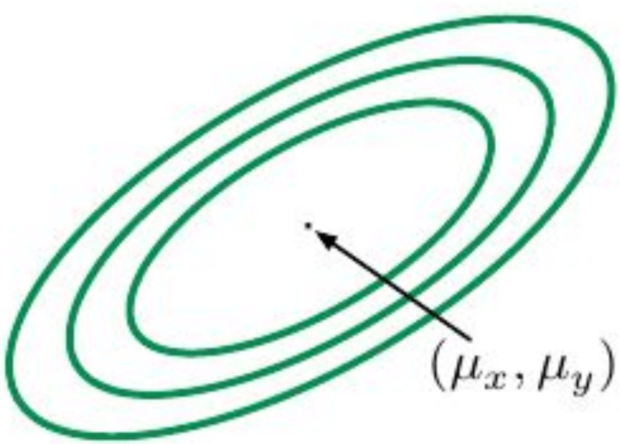
- **Mean** $(\mu_x, \mu_y) \in \mathbb{R}^2$
- **Covariance matrix**

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

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

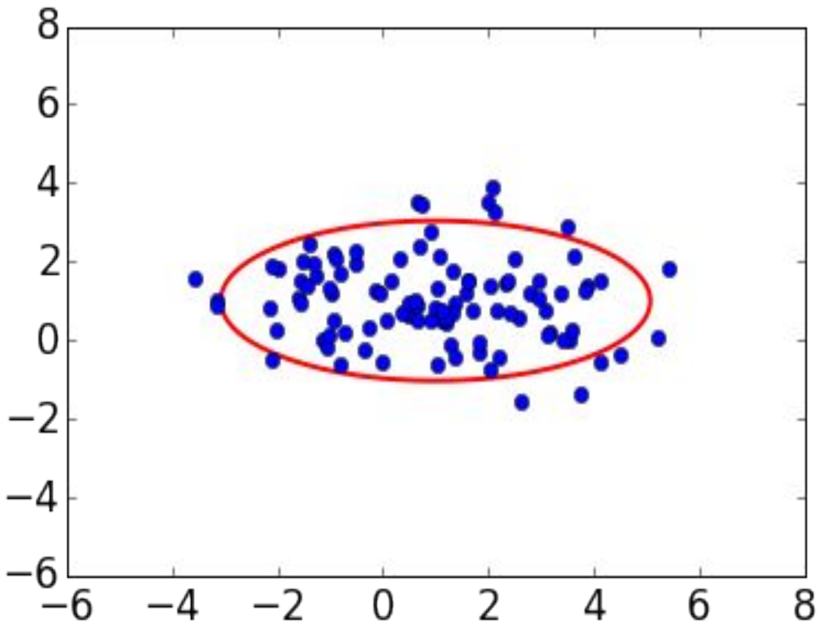
Density $\quad p(x, y) = \dfrac{1}{2\pi|\Sigma|^{1/2}} \exp\left(-\dfrac{1}{2} \begin{bmatrix} x - \mu_x \\ y - \mu_y \end{bmatrix}^T \Sigma^{-1} \begin{bmatrix} x - \mu_x \\ y - \mu_y \end{bmatrix} \right)$

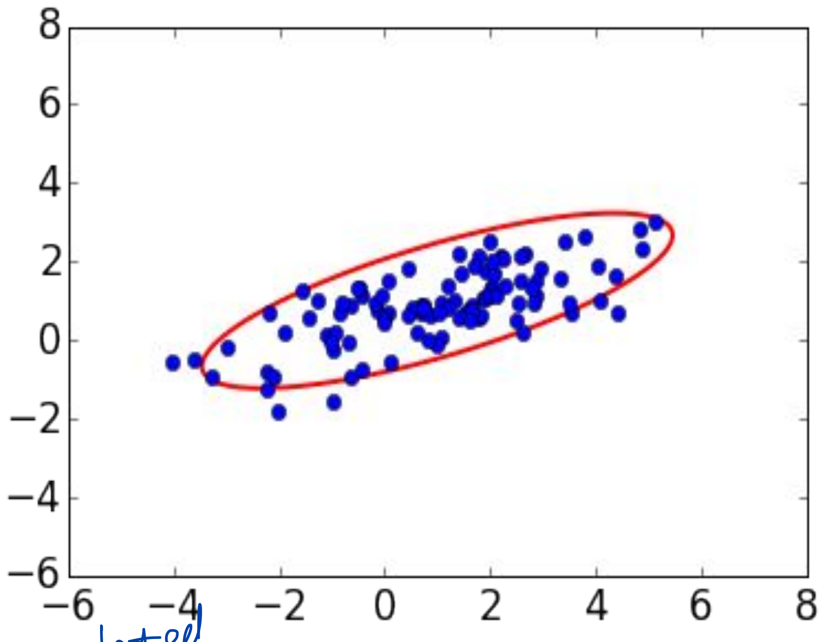
The density is highest at the mean, and falls off in ellipsoidal contours.



Bivariate Gaussian: examples

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



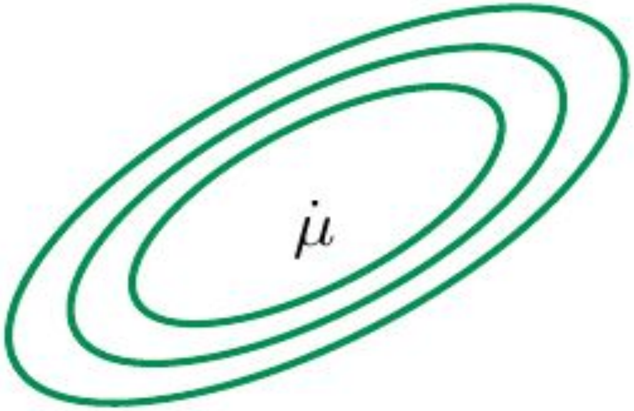


cov = 0 uncorrelated

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

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

The multivariate Gaussian



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

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

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

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

- μ is the vector of coordinate-wise means:

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

- Σ is a matrix containing all pairwise covariances:

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

- 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 σ^2. Thus

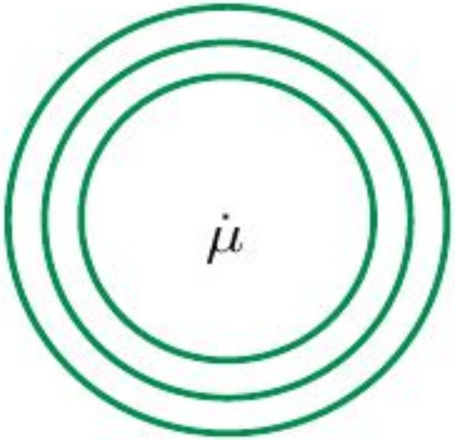
$$\Sigma = \sigma^2 I_p = \mathrm{diag}(\sigma^2, \sigma^2, \ldots, \sigma^2)$$ uncorrelated
share the same variance

(off-diagonal elements zero, diagonal elements σ^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 μ:



Special case: diagonal Gaussian

The X_i are independent, with variances σ_i^2. Thus

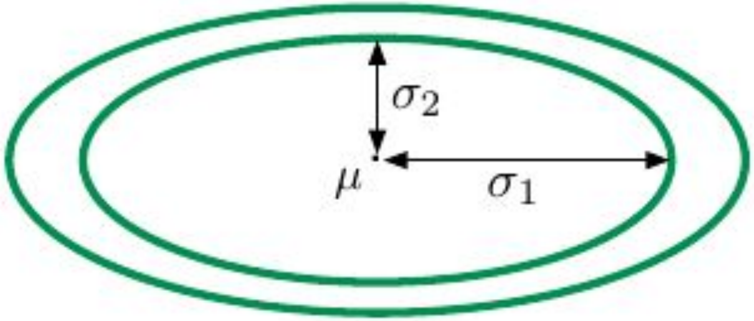
$$\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$$

(all off-diagonal elements zero).

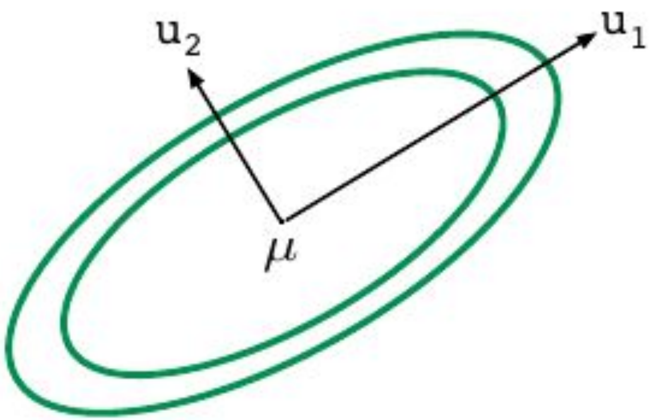
Each X_i is an independent univariate Gaussian $N(\mu_i, \sigma_i^2)$:

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

Contours of equal density are axis-aligned ellipsoids centered at μ:



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



Eigendecomposition of Σ 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 π_1, π_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 $x^{(1)}, \dots, x^{(m)} \in \mathbb{R}^p$ are class 1:

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

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 θ is a constant depending on the various parameters.

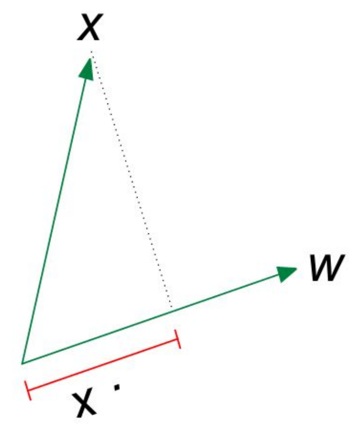
$\Sigma_1 = \Sigma_2$: linear decision boundary. Otherwise, quadratic boundary.

Linear decision boundary

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

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

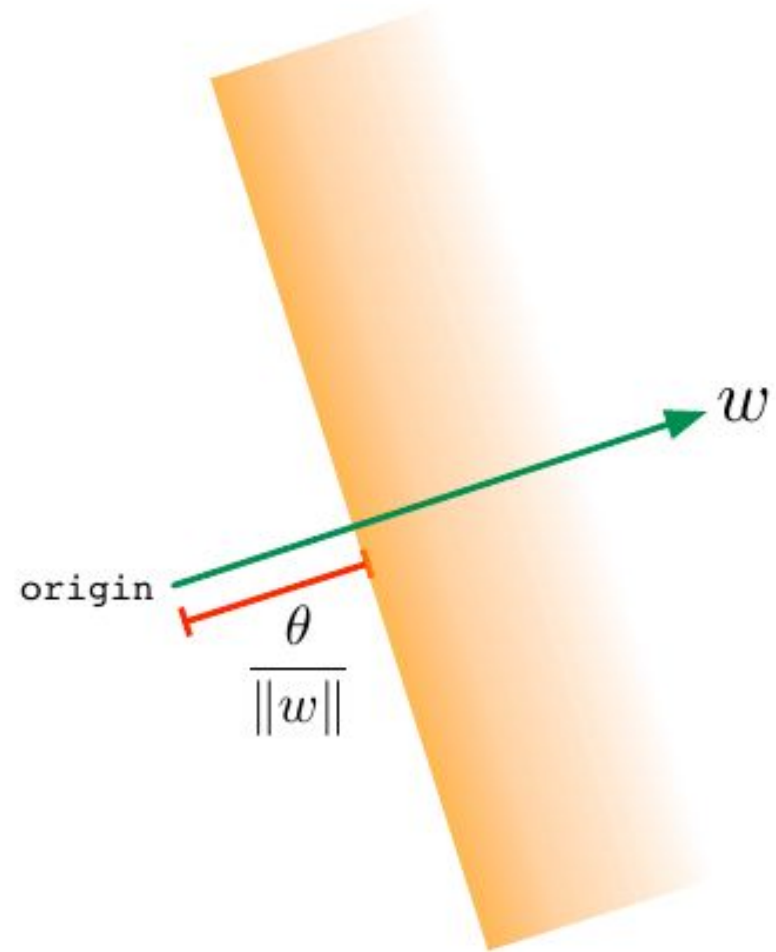
Geometric picture: Suppose w is a unit vector (that is, $||w|| = 1$). Then $x \cdot w$ is the **projection** of vector x onto direction w .



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

Linear decision boundary

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

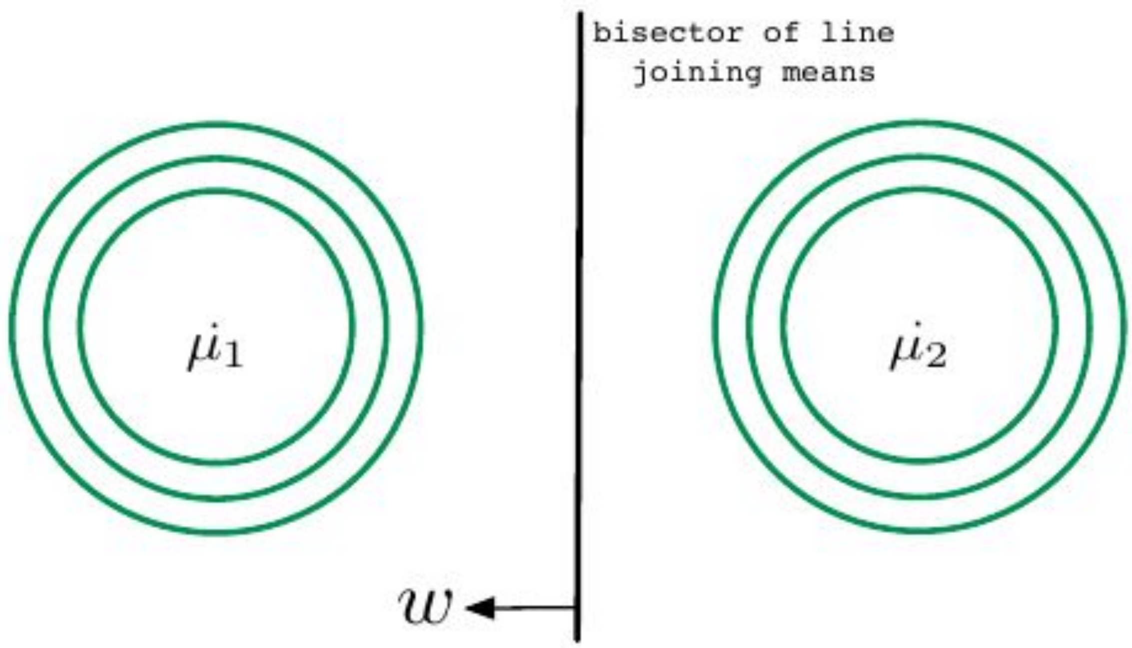


Common covariance: $\Sigma_1 = \Sigma_2 = \Sigma$

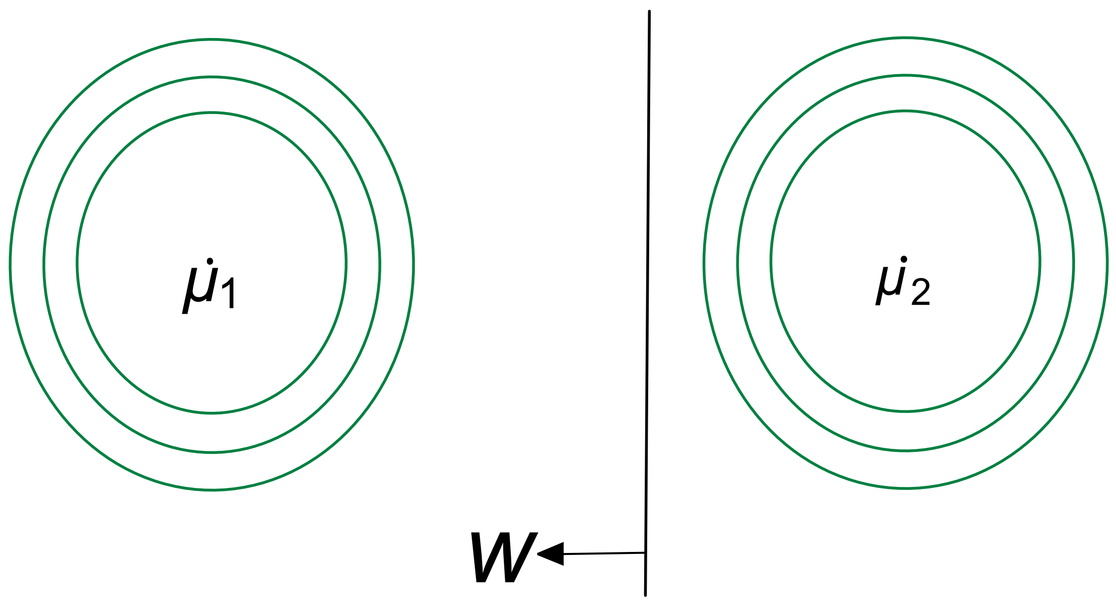
Linear decision boundary: choose class 1 iff

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

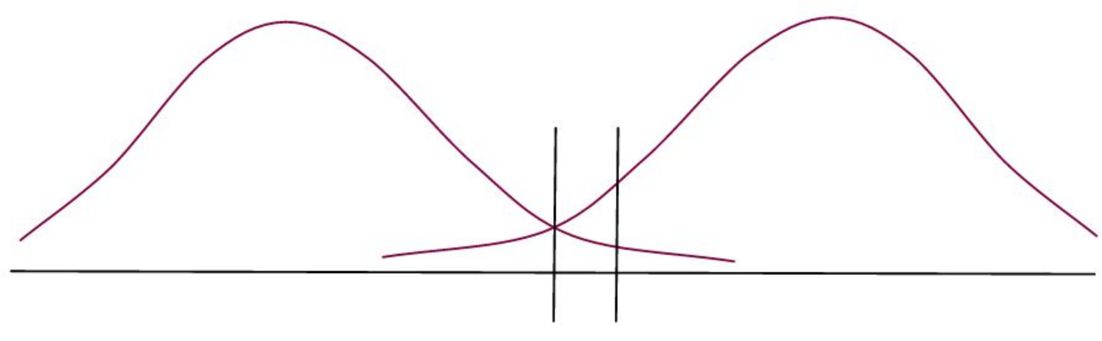
Example 1: Spherical Gaussians with $\Sigma = I_p$ and $\pi_1 = \pi_2$.



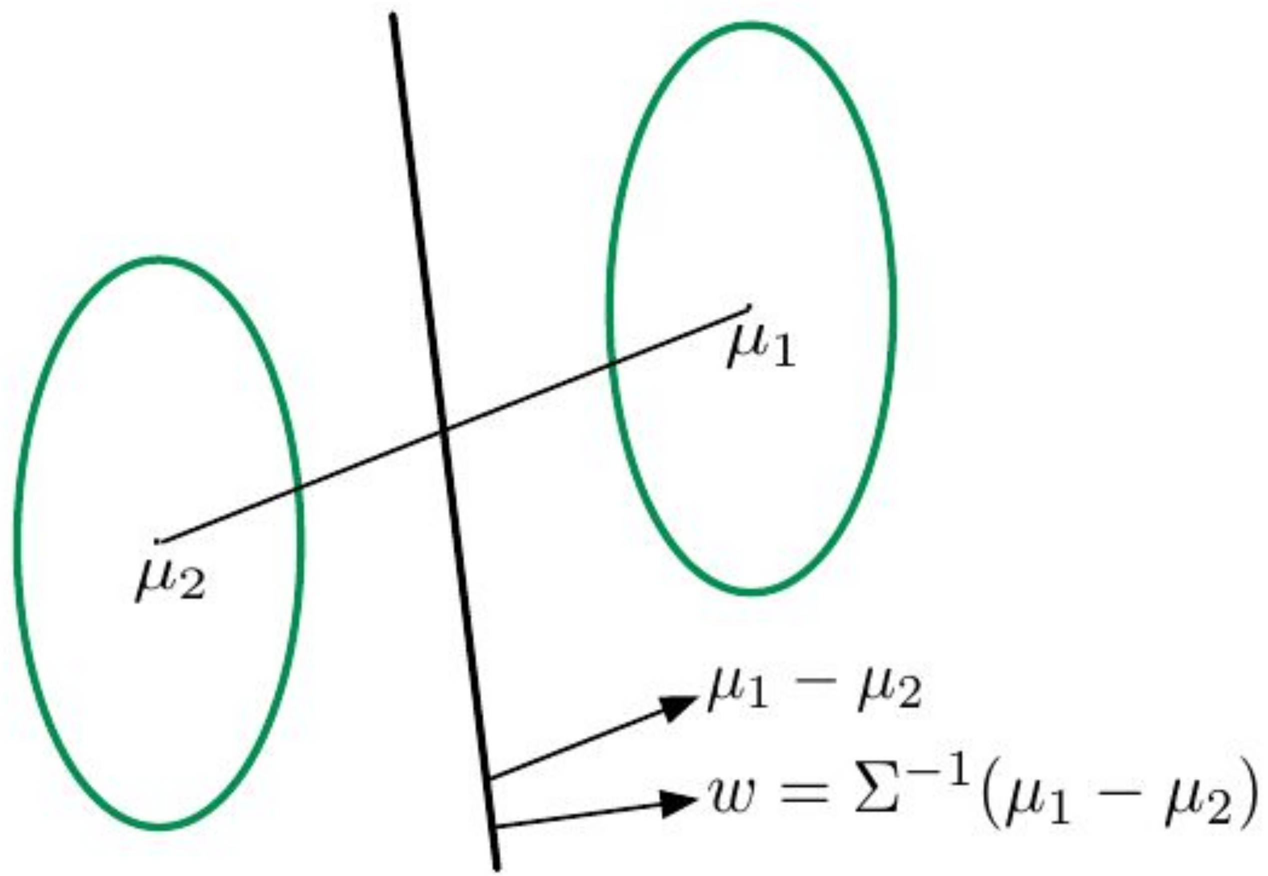
Example 2: Again spherical, but now $\pi_1 > \pi_2$.



One-d projection onto w :



Example 3: Non-spherical.



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

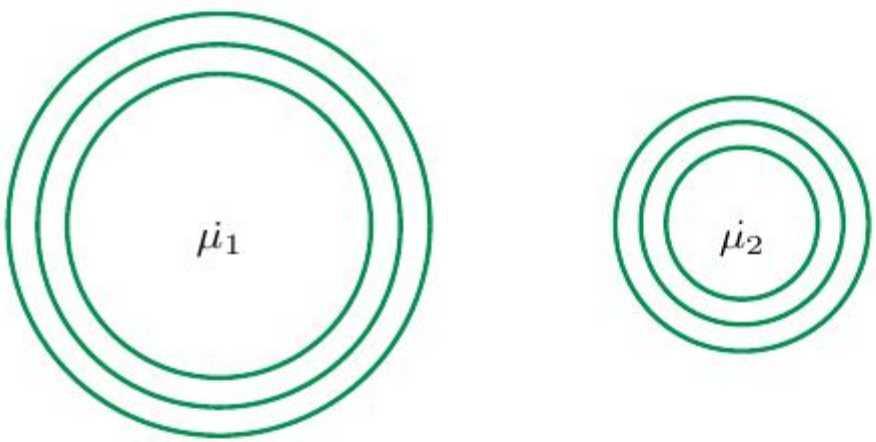
- w, θ dictated by probability model, assuming it is a perfect fit
- Common practice: choose w as above, but fit θ to minimize training/validation error

Different covariances: $\Sigma_1 \neq \Sigma_2$

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

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

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

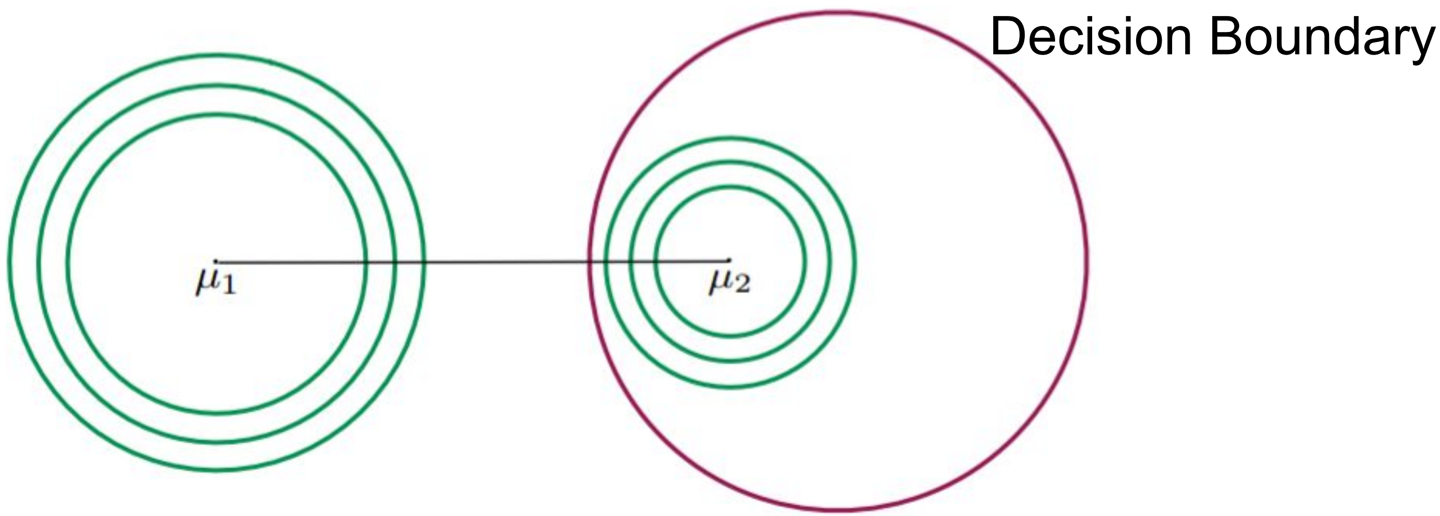


Different covariances: $\Sigma_1 \neq \Sigma_2$

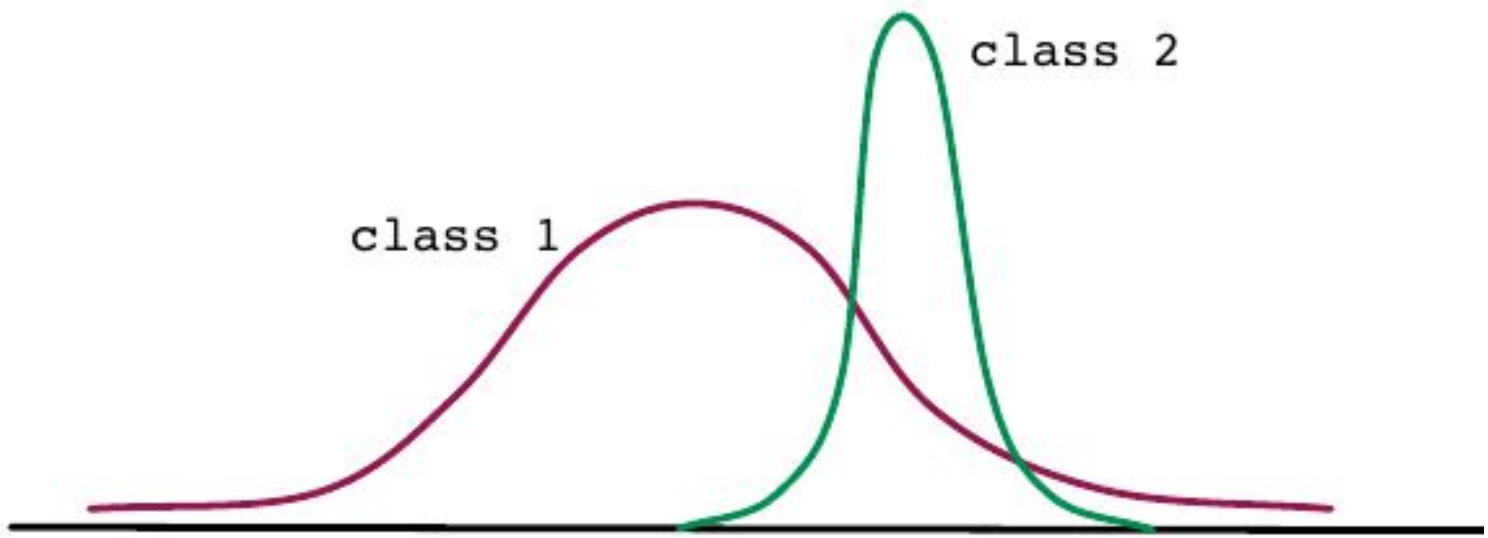
Quadratic boundary: choose class 1 iff $x^T M x + 2w^T x \geq \theta$, where:

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

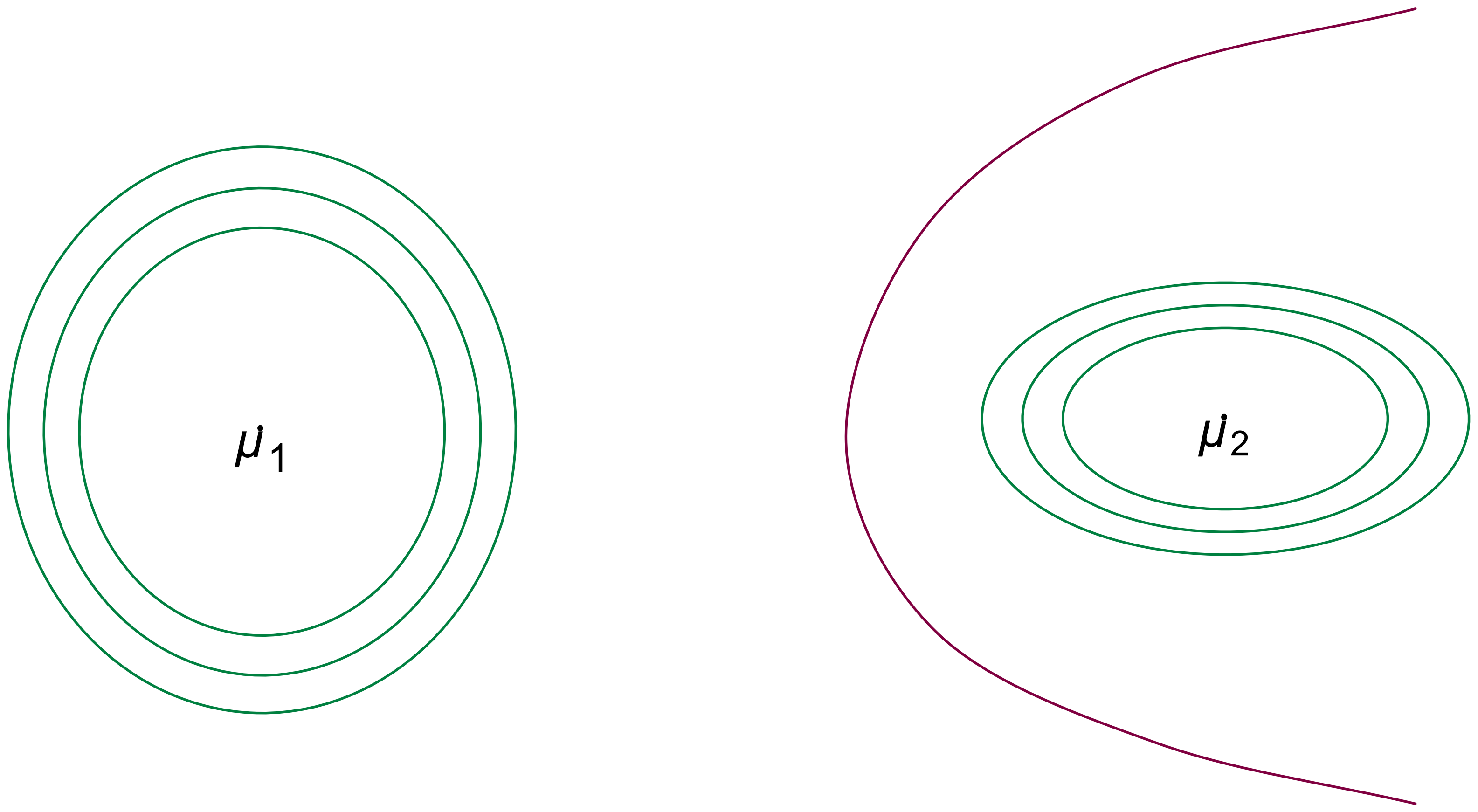
Example 1: $\Sigma_1 = \sigma_1^2 I_p$ and $\Sigma_2 = \sigma_2^2 I_p$ with $\sigma_1 > \sigma_2$



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



Example 3: A parabolic boundary.



Many other possibilities!

Multiclass discriminant analysis

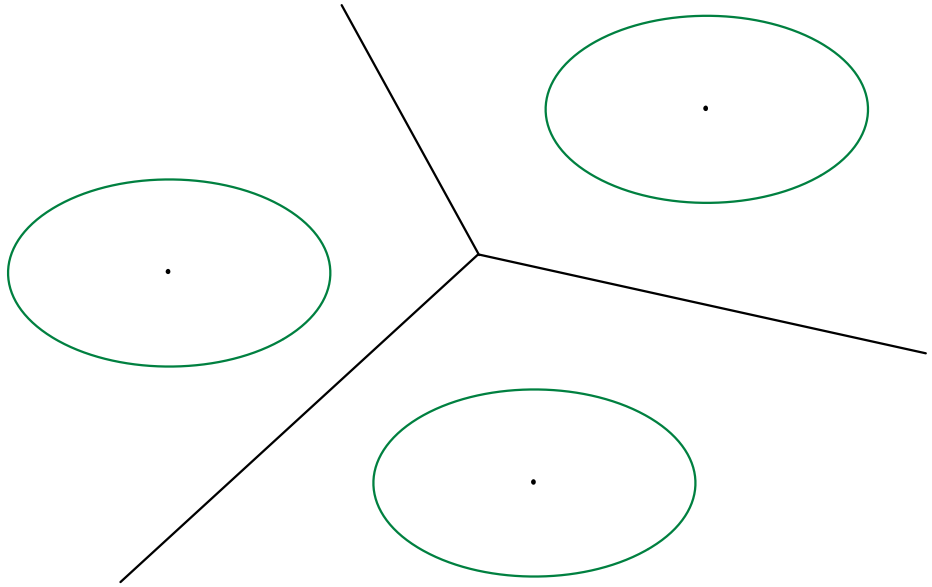
k classes: weights π_j, class-conditional distributions $P_j = N(\mu_j, \Sigma_j)$

Each class has an associated **quadratic** function

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

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

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

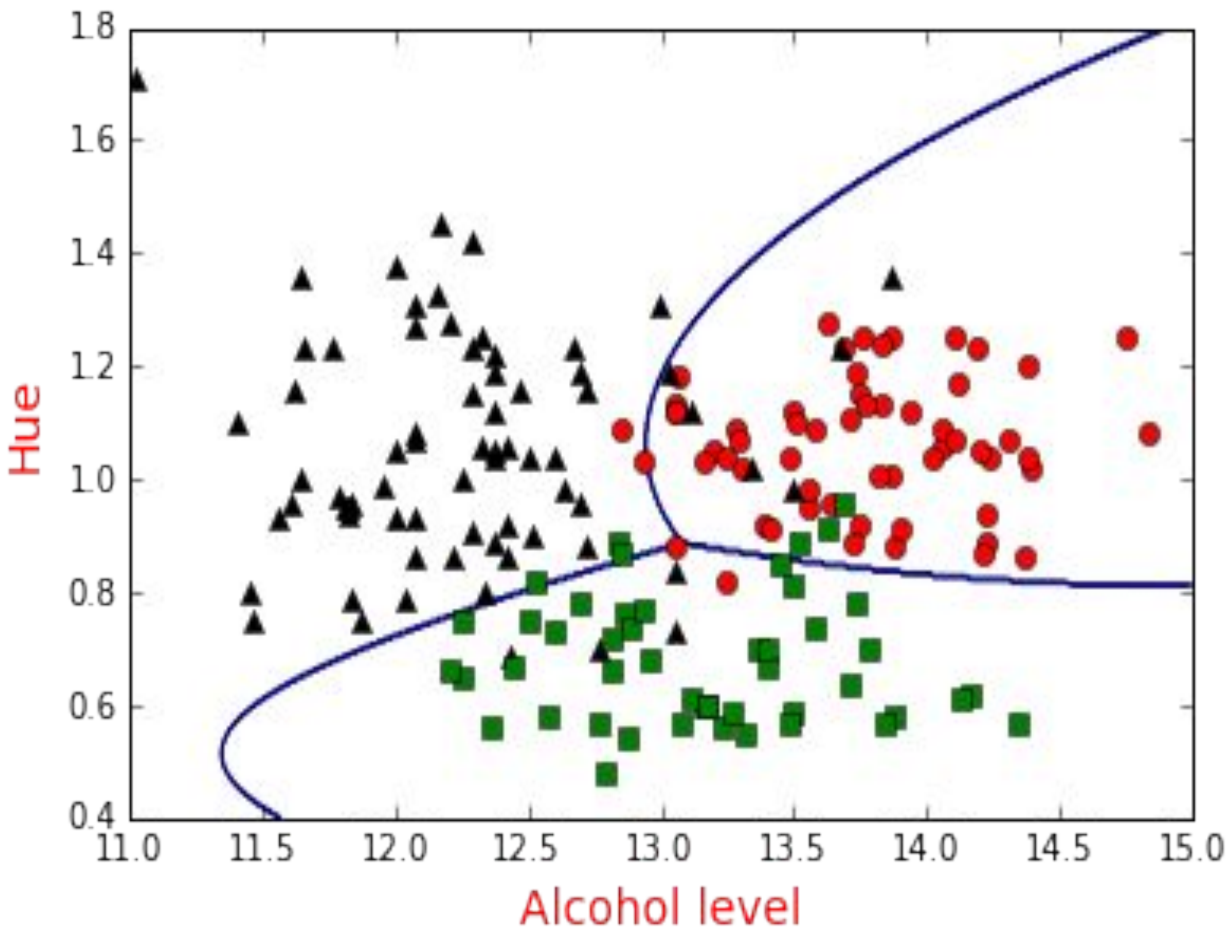


Example: "wine" data set

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 π_j
- mean $\mu_j \in R^{784}$
- covariance matrix $\Sigma_j \in R^{784x784}$

Problem: formula for normal density uses Σ_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

Class Outline

- Parametric Methods
  - Generative Models
    - Naive Bayes
    - Binary Features, Multinomial Features
    - *Hands-On*
    - Gaussian Generative Model
    - Fisher Linear Discriminant Analysis
    - *Hands-On*

Fisher's linear discriminant

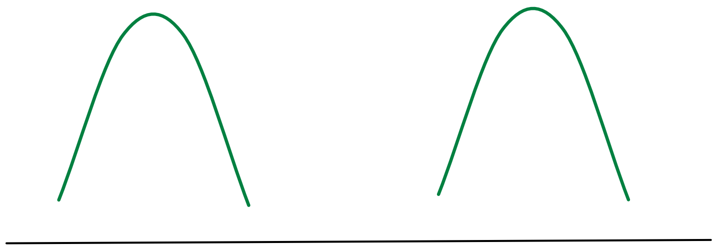
A framework for linear classification without Gaussian assumptions.

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

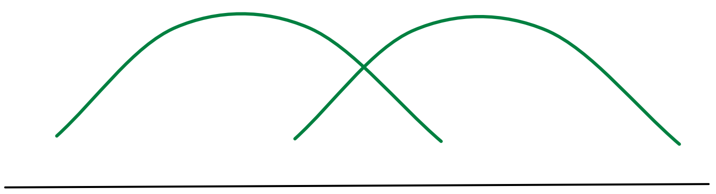
Class 1	Class 2
mean μ_1	mean μ_2
cov Σ_1	cov Σ_2
# pts n_1	# pts n_2

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 μ_1, μ_2; covariances Σ_1, Σ_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 $J(w) = \dfrac{(w \cdot \mu_1 - w \cdot \mu_2)^2}{w^T \Sigma w}$

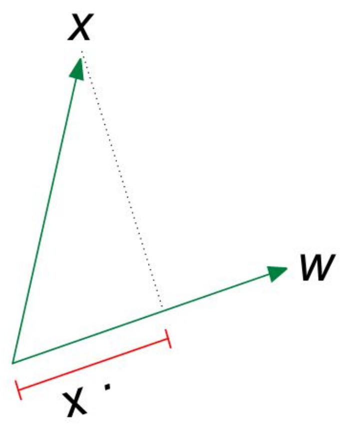
Solution: $w \propto \Sigma^{-1}(\mu_1 - \mu_2)$. Look familiar?

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 Σ_1, Σ_2 are full rank; else project.
2. Since Σ_1 and Σ_2 are p.d., so is their weighted average, Σ.
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: $w = \Sigma^{-1/2} u \propto \Sigma^{-1}(\mu_1 - \mu_2)$.