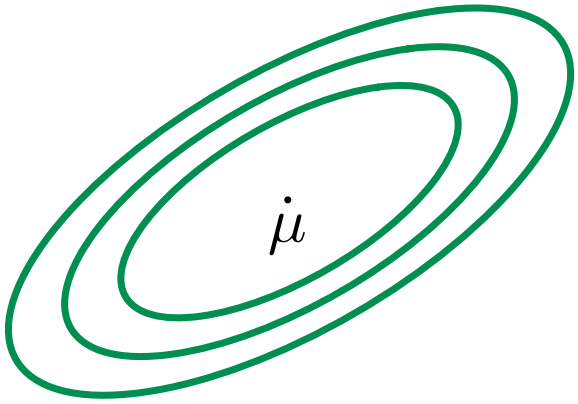
# Generative modeling with the multivariate Gaussian

Instructor: Taylor Berg-Kirkpatrick
Slides: Sanjoy Dasgupta

# The multivariate Gaussian




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

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

Generates points $X = (X_1, X_2, \ldots, X_d)$.

$X^{(i)}$ ← ith training point

$X^{(i)}_j$ ← jth dim of $X^{(i)}$

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

$$\mu_1 = \mathbb{E}X_1, \ \mu_2 = \mathbb{E}X_2, \ldots, \ \mu_d = \mathbb{E}X_d.$$

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

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

$(x-\mu) \in \mathbb{R}^{d\times 1}$

$\Sigma \in \mathbb{R}^{d\times d}$

$\Sigma^{-1} \in \mathbb{R}^{d\times d}$

$(1\times d)(d\times d)(d\times 1)$

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

# Special case: spherical Gaussian

The $X_i$ are independent and all have the same variance $\sigma^2$. What is the covariance matrix $\Sigma$?

$$\Sigma = \sigma^2 I_d$$

$$I_d = \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix} \quad d$$

$$\sigma^2 I_d = \begin{bmatrix} \sigma^2 & & 0 \\ & \sigma^2 & \\ 0 & & \sigma^2 \end{bmatrix}$$

# Spherical Gaussian (cont'd)

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

Simplify the density

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

$$= \frac{1}{C(\Sigma)} \exp\left(-\frac{1}{2}(x_1-\mu_1,\dots,x_d-\mu_d)\begin{bmatrix}\frac{1}{\sigma^2} & & 0\\ & \frac{1}{\sigma^2} & \\ 0 & & \frac{1}{\sigma^2}\end{bmatrix}\begin{pmatrix}x_1-\mu_1\\ \vdots\\ x_d-\mu_d\end{pmatrix}\right)$$

$$= \frac{1}{C(\Sigma)} \exp\left(-\frac{1}{2}(x_1-\mu_1,\dots,x_d-\mu_d)\begin{pmatrix}\frac{1}{\sigma^2}(x_1-\mu_1)\\ \vdots\\ \frac{1}{\sigma^2}(x_d-\mu_d)\end{pmatrix}\right)$$

$$= \frac{1}{C(\Sigma)} \exp\left(-\frac{1}{2}\cdot\sum_{i=1}^{d}\frac{1}{\sigma^2}(x_i-\mu_i)^2\right) = \frac{1}{C(\Sigma)}\prod_i \exp\left(-\frac{\|x_i-\mu_i\|_2^2}{2\sigma^2}\right)$$

shape of univ. $N$

$\frac{1}{\sigma^2}\|x-\mu\|_2^2$

# Spherical Gaussian (cont'd)

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

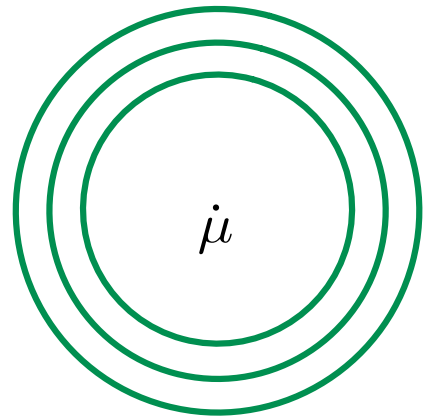
$$\Sigma = \sigma^2 I_d = \text{diag}(\sigma^2, \sigma^2, \ldots, \sigma^2)$$

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

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

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

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

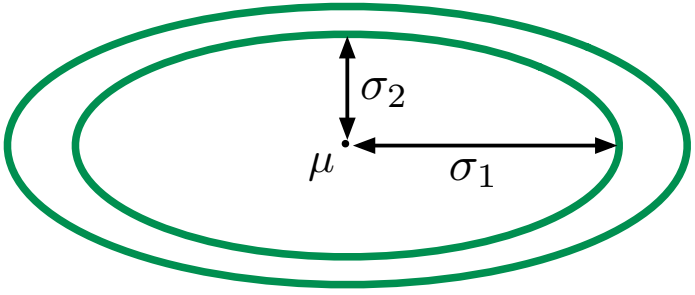
$$\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_d^2) = \begin{bmatrix} \sigma_1^2 & & 0 \\ & \sigma_2^2 & \\ & & \ddots \\ 0 & & \sigma_d^2 \end{bmatrix}$$

(all off-diagonal elements zero).

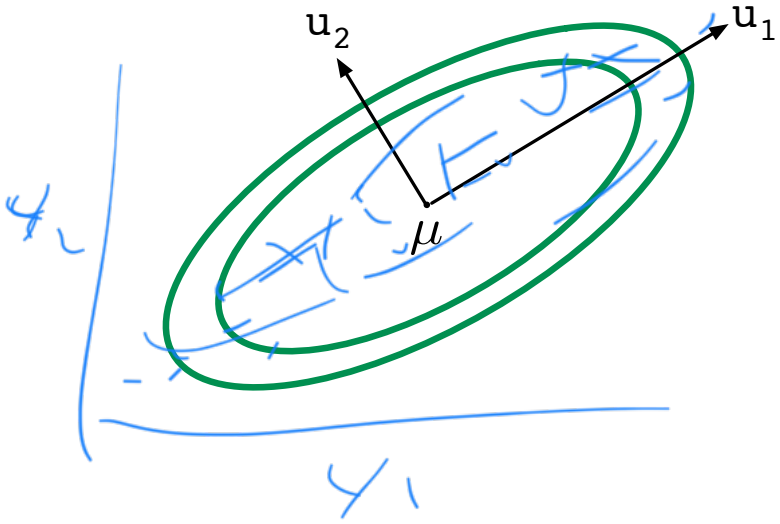
Each $X_i$ is an independent univariate Gaussian $N(\mu_i, \sigma_i^2)$:

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

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



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




- **Eigenvalues** of $\Sigma$ are $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$
- Corresponding **eigenvectors** $u_1, \ldots, u_d$

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

If we write $S = \Sigma^{-1}$ then $S$ is a $d \times d$ 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 $x$.

# How to fit a Gaussian to data

Fit a Gaussian to data points $x^{(1)}, \ldots, x^{(m)} \in \mathbb{R}^d$.

- Mean

$$\mu = \frac{1}{m}\left(x^{(1)} + \cdots + x^{(m)}\right)$$

- Covariance matrix has $i, j$ entry:

$$\Sigma_{ij} = \frac{1}{m}\sum_{k=1}^{m}(x_i^{(k)} - \mu_i)(x_j^{(k)} - \mu_j)$$

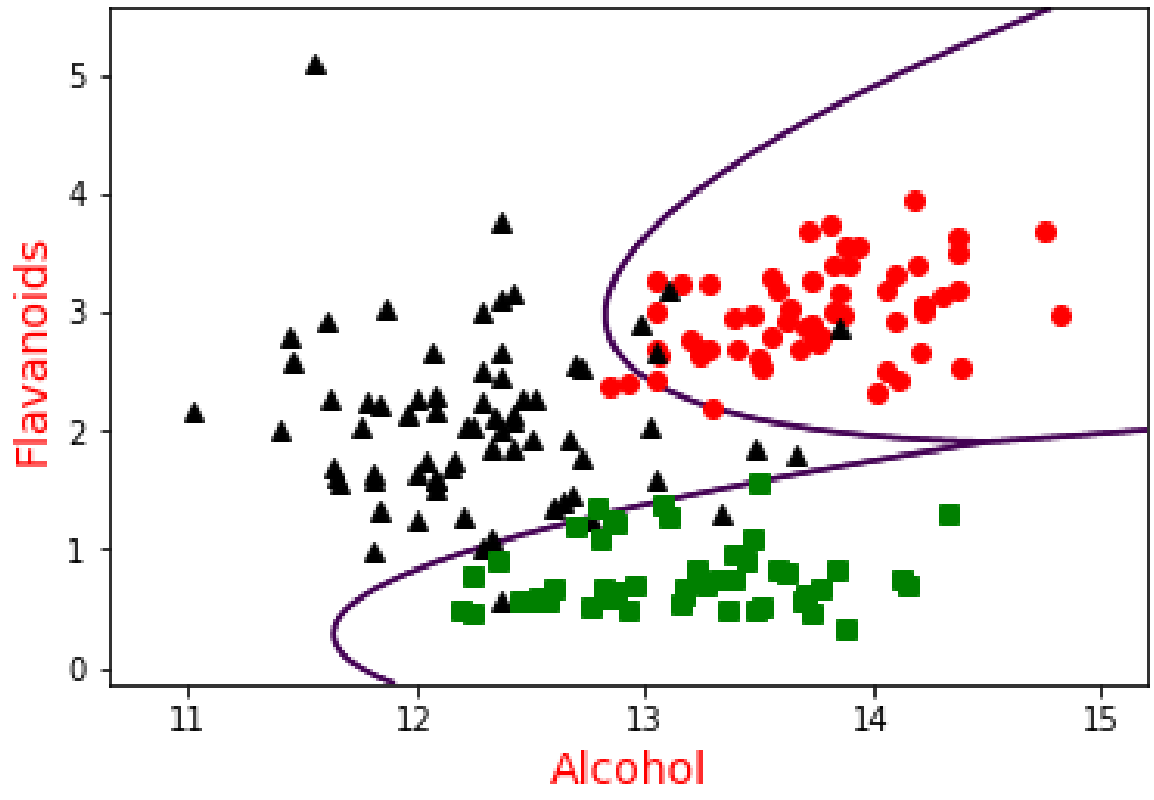
In more compact notation:

$$\Sigma = \frac{1}{m}\sum_{k=1}^{m}(x^{(k)} - \mu)(x^{(k)} - \mu)^T$$

These are the **empirical mean** and **empirical covariance**.

# Back to the winery data

Go from 1 to 2 features: test error goes from 29% to 8%.

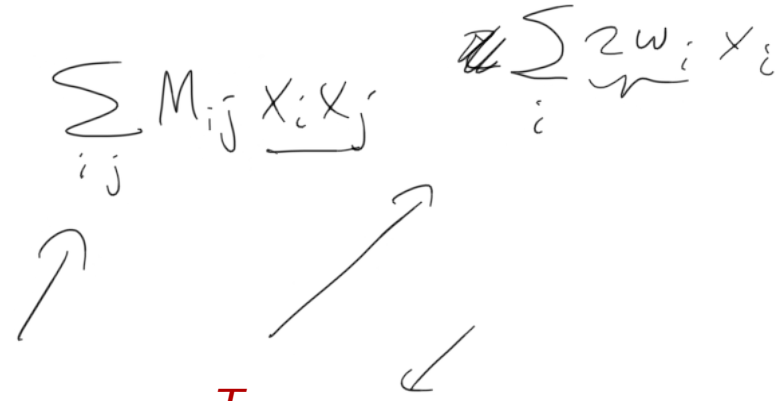



**With all 13 features: error rate goes to zero.**

# Binary classification with Gaussian generative model

We have data from two classes (1 and 2).

- Estimate class probabilities $\pi_1, \pi_2$
- Fit a Gaussian to each class:
  $P_1 = N(\mu_1, \Sigma_1),\ P_2 = N(\mu_2, \Sigma_2)$



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 threshold depending on the various parameters.

**Linear** or **quadratic** decision boundary.

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

Linear decision boundary: choose class 1 iff

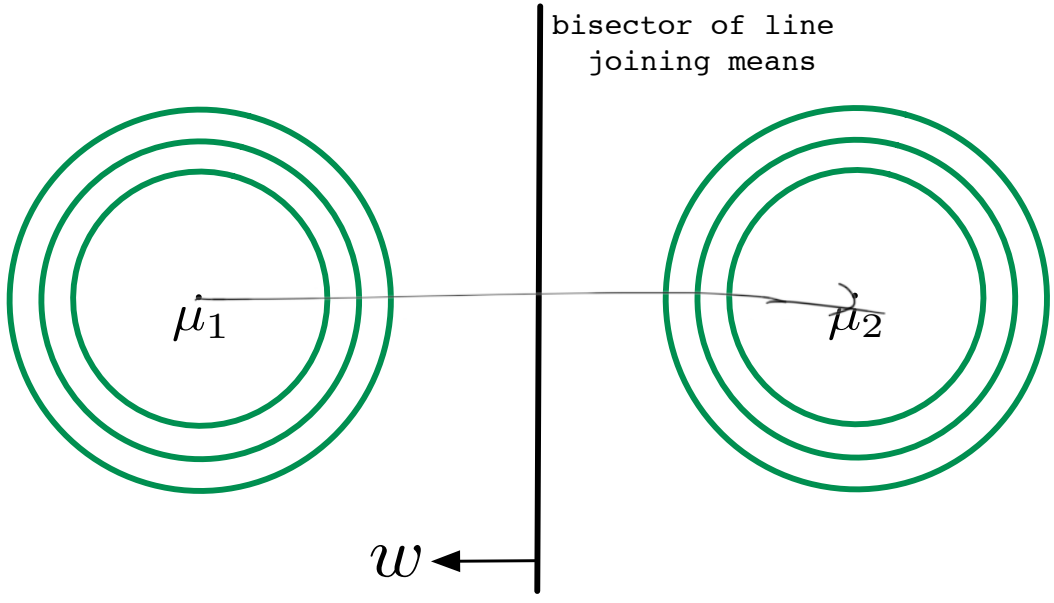
$$x \cdot \underbrace{\Sigma^{-1}(\mu_1 - \mu_2)}_{w} \geq \theta.$$

$w^T x \geq 0$

$\forall x$ s.t. $w^T x = 0$

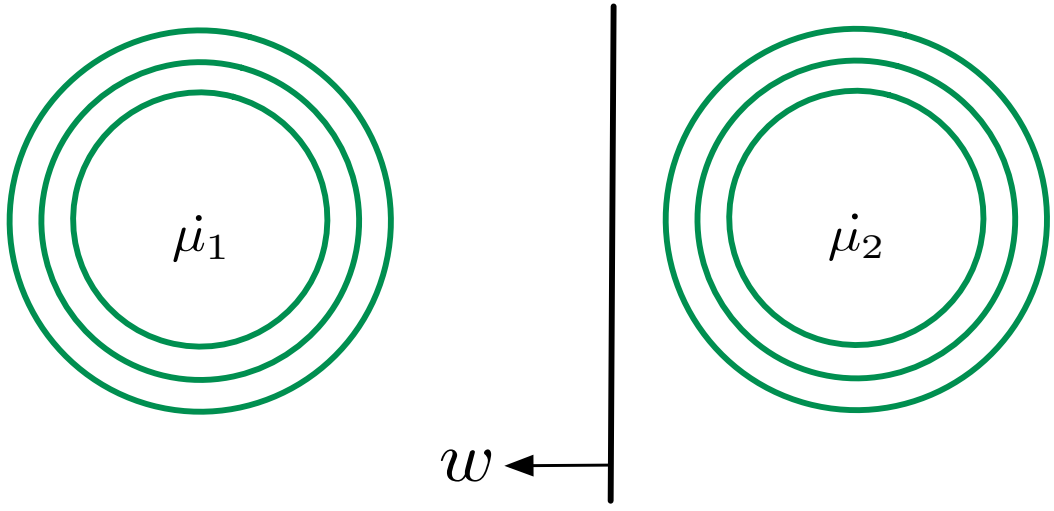


Example 1: Spherical Gaussians with $\Sigma = I_d$ and $\pi_1 = \pi_2$.

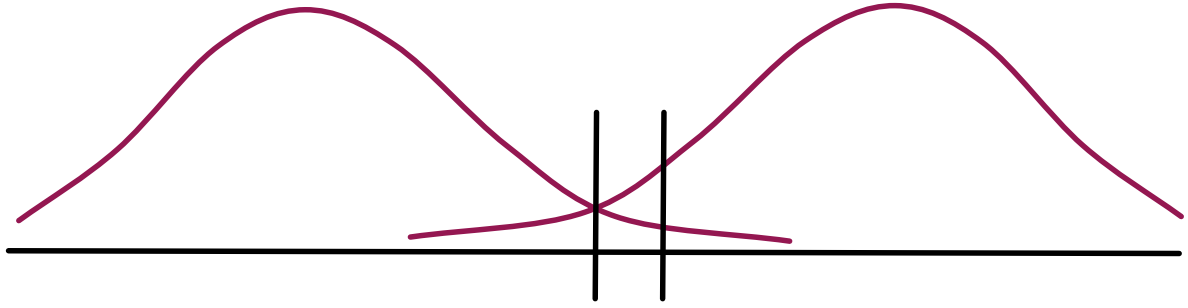


Example 2: Again spherical, but now $\pi_1 > \pi_2$.

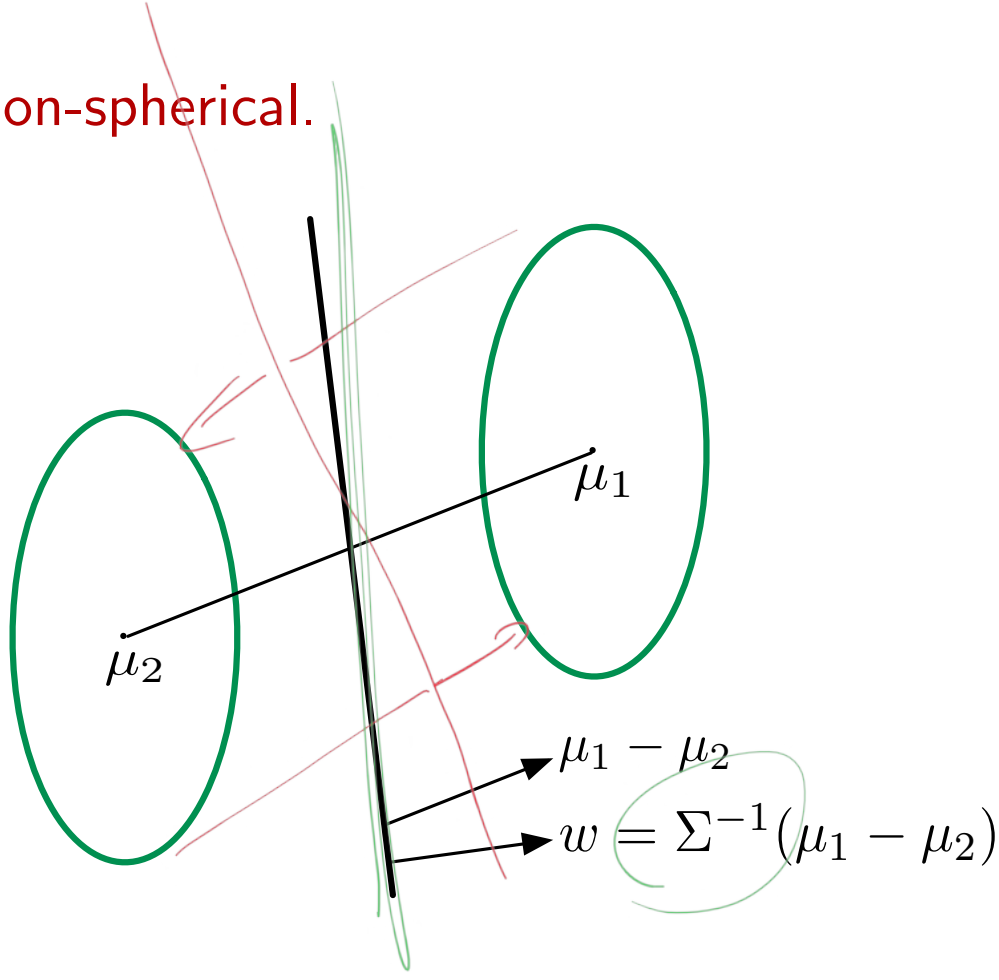
$\Sigma^* = \Sigma + cI_d$




One-d projection onto $w$:



## Example 3: Non-spherical.




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

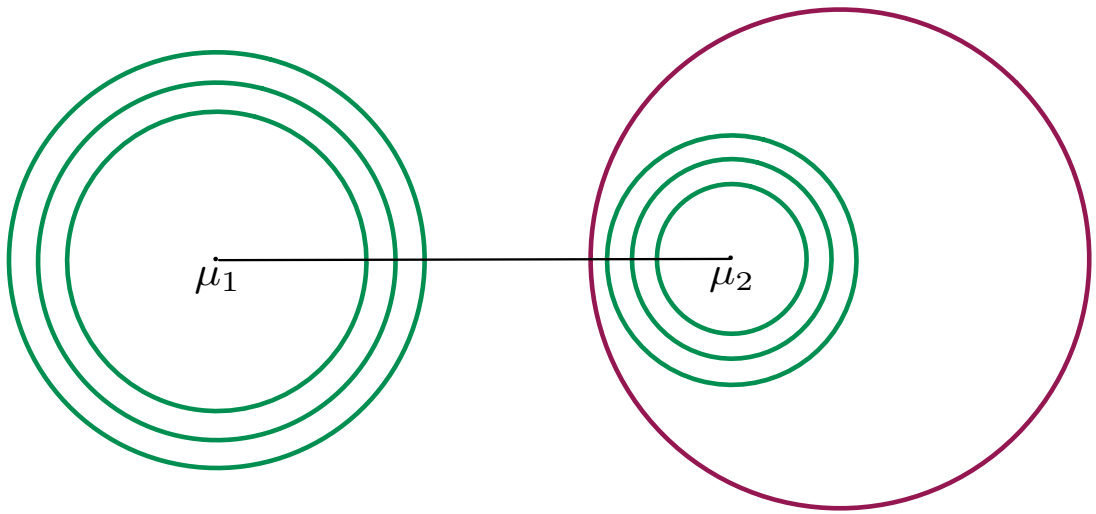
- $w, \theta$ dictated by probability model, assuming it is a perfect fit
- Common practice: choose $w$ as above, but fit $\theta$ to minimize training/validation error

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

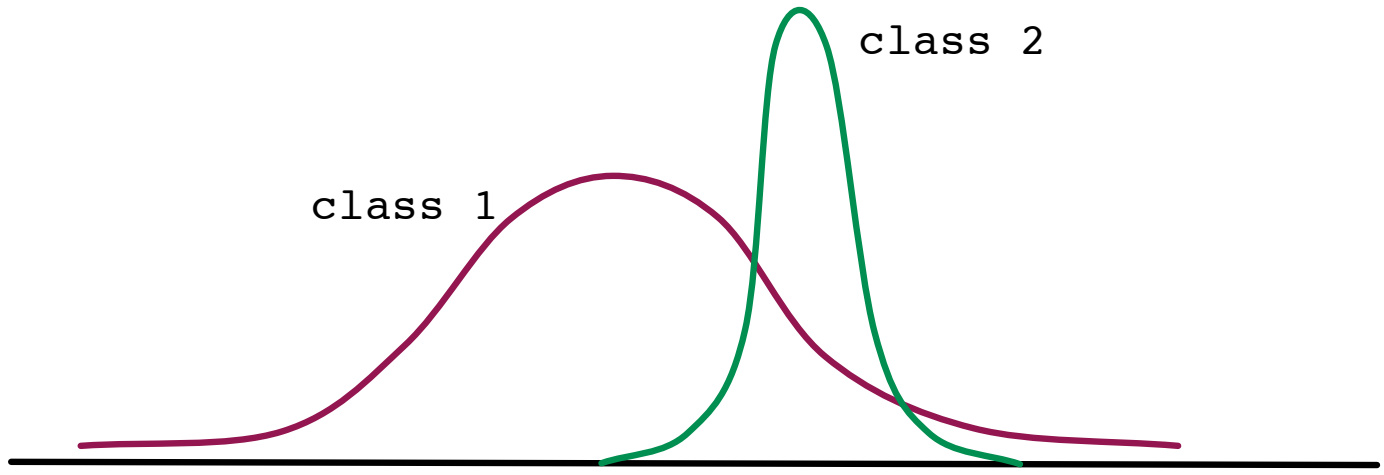
Quadratic boundary: choose class 1 iff $x^T M x + 2 w^T x \geq \theta$, where:

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

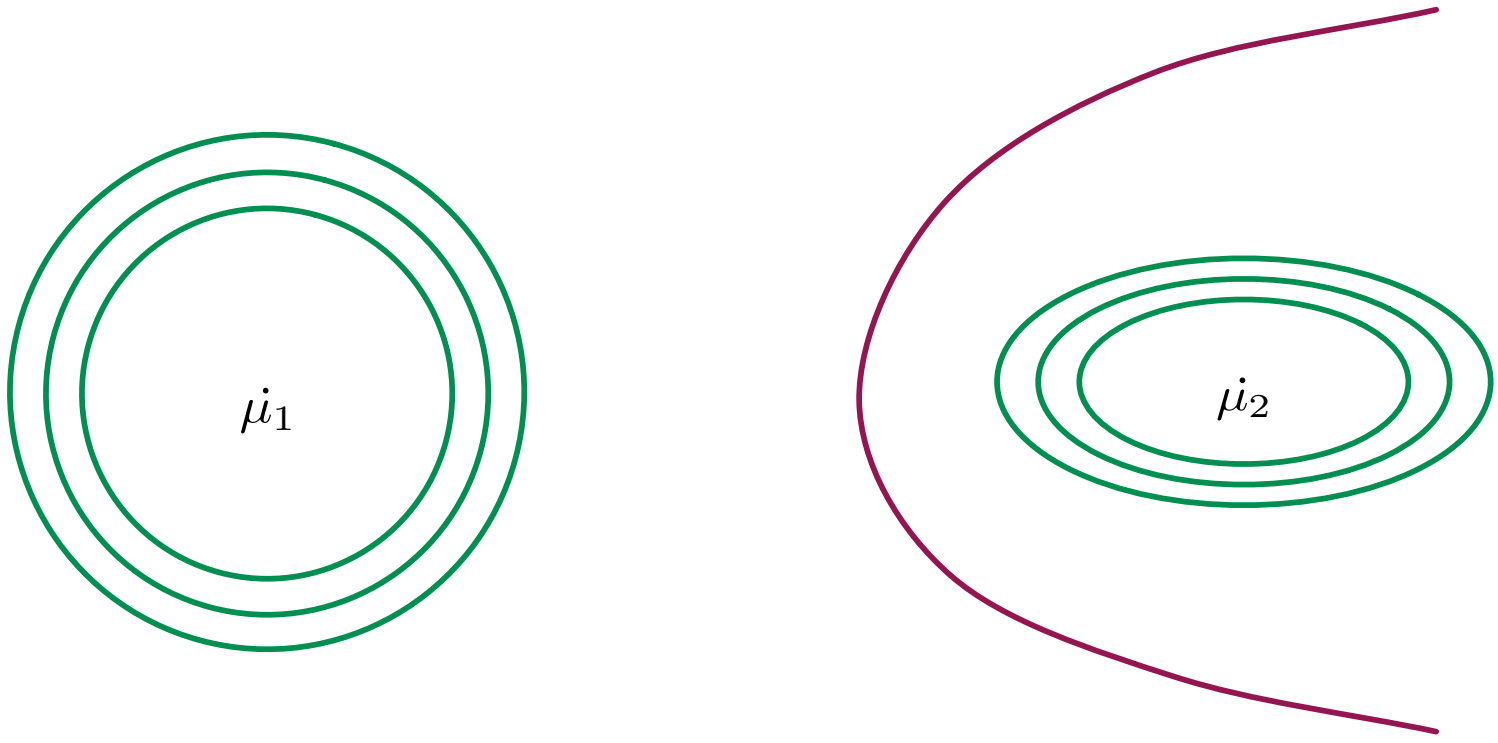
Example 1: $\Sigma_1 = \sigma_1^2 I_d$ and $\Sigma_2 = \sigma_2^2 I_d$ with $\sigma_1 > \sigma_2$



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



Example 3: A parabolic boundary.




Many other possibilities!

# Multiclass discriminant analysis

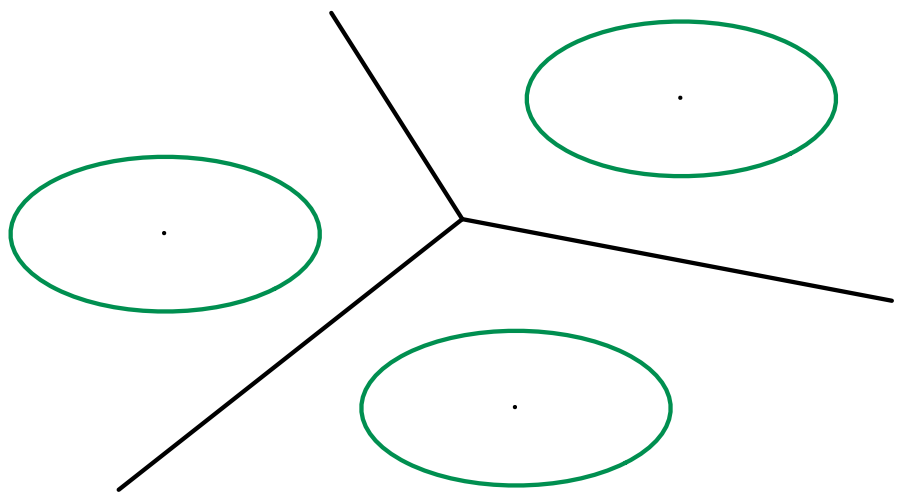
$k$ classes: weights $\pi_j$, class-conditional densities $P_j = N(\mu_j, \Sigma_j)$.

Each class has an associated **quadratic** function

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

To classify point $x$, pick $\arg\max_j f_j(x)$.

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



# Try it on MNIST



To each digit, fit:

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

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

- Need to **regularize/smooth**: $\Sigma_j \to \Sigma_j + cI$
- This is a good idea even without the singularity issue

How to choose $c$? With a **validation set**.

- Divide original training set into training and validation set.
- Fit $\pi_j, \mu_j, \Sigma_j$ to training set
- Choose $c$ with lowest error rate on validation set