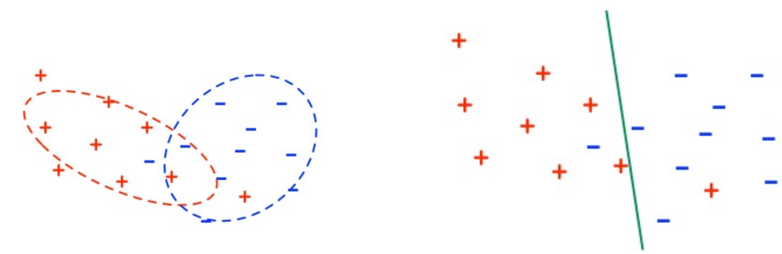
Classification with generative models II

DSE 210

Classification with parametrized models

Classifiers with a fixed number of parameters can represent a limited set of functions. Learning a model is about picking a good approximation.

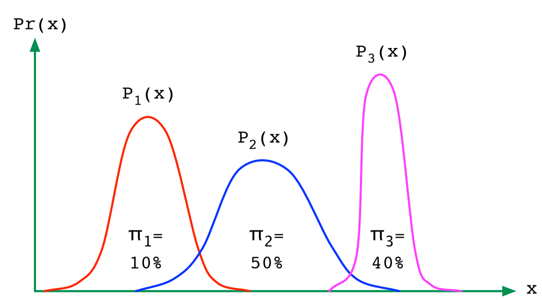
Typically the x's are points in p-dimensional Euclidean space, $\mathbb{R}^p$.



Two ways to classify:

- **Generative**: model the individual classes.
- **Discriminative**: model the decision boundary between the classes.

The Bayes-optimal prediction



Labels $\mathcal{Y} = \{1, 2, \ldots, k\}$, 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 π_j is easy. Estimating the P_j is hard.

Estimating class-conditional distributions

Estimating an arbitrary distribution in $\mathbb{R}^p$:

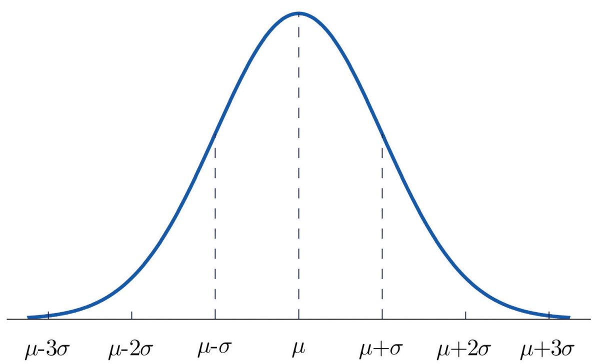
- Can be done, e.g. with kernel density estimation.
- But number of samples needed is exponential in p.

Instead: 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.

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?

Bivariate distributions

Simplest option: treat each variable as independent.

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

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

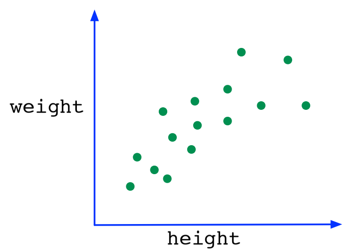
Independence would mean

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

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

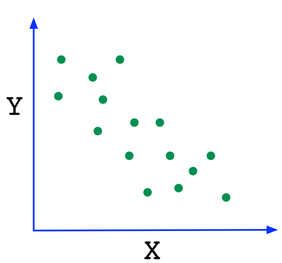
Is this an accurate approximation?
No: we'd expect height and weight to be **positively correlated**.

Types of correlation

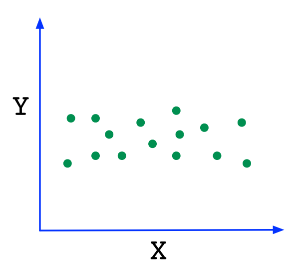


H, W positively correlated.
This also implies

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

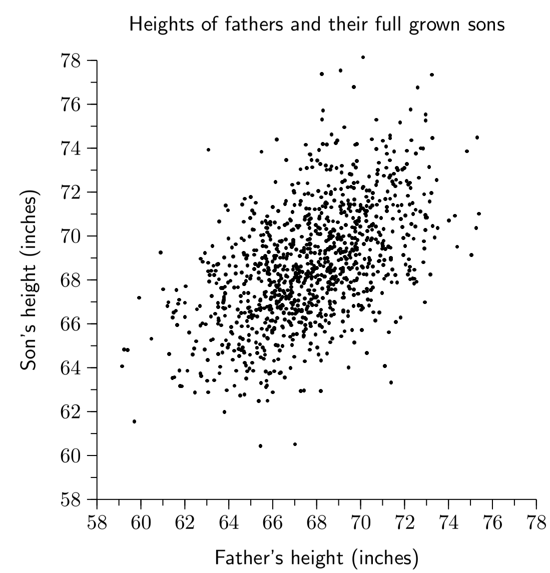


X, Y negatively correlated



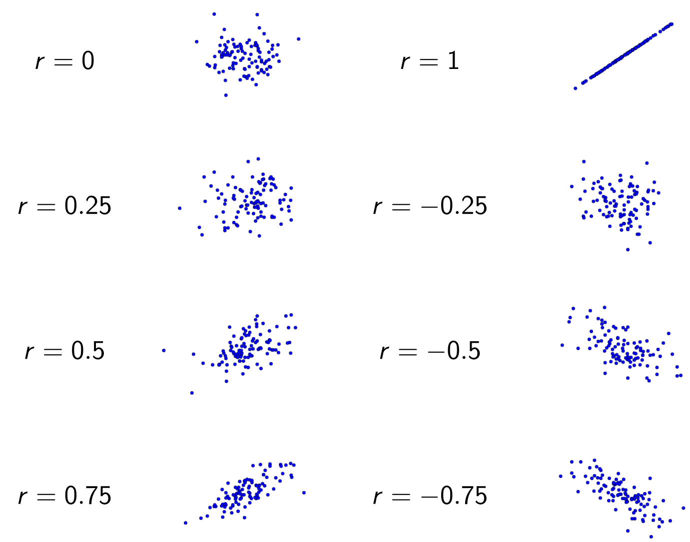
X, Y uncorrelated

Pearson (1903): fathers and sons



How to quantify the degree of correlation?

Correlation pictures



Covariance and correlation

Suppose X has mean μ_X and Y has mean μ_Y.

- Covariance

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

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

- Correlation

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

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

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

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

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

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

In this case, X and Y are negatively correlated.

The bivariate (2-d) Gaussian

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

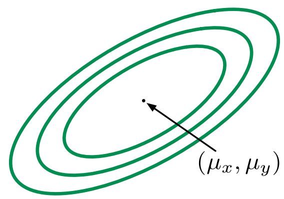
- **Mean** $(\mu_x, \mu_y) \in \mathbb{R}^2$
- **Covariance matrix**

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

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

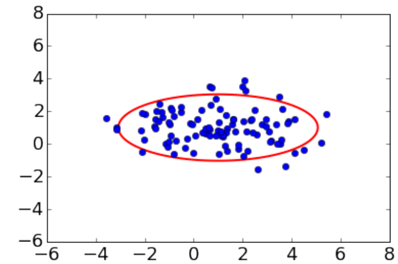
Density $$p(x, y) = \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}\begin{bmatrix} x-\mu_x \\ y-\mu_y \end{bmatrix}^T \Sigma^{-1} \begin{bmatrix} x-\mu_x \\ y-\mu_y \end{bmatrix}\right)$$

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

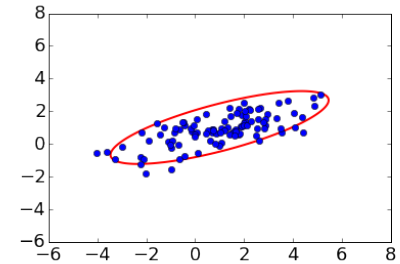


Bivariate Gaussian: examples

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

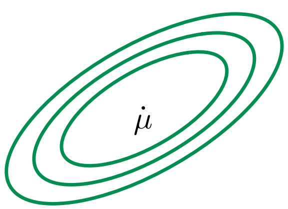


$$\Sigma = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}$$



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

The multivariate Gaussian



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

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

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

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

- 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:
$$\Sigma_{ij} = \Sigma_{ji} = \text{cov}(X_i, X_j) \quad \text{if } i \neq j$$
$$\Sigma_{ii} = \text{var}(X_i)$$
- 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 σ^2. Thus

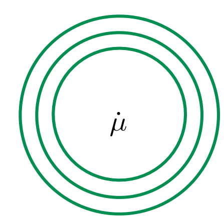
$$\Sigma = \sigma^2 I_p = \text{diag}(\sigma^2, \sigma^2, \ldots, \sigma^2)$$

(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

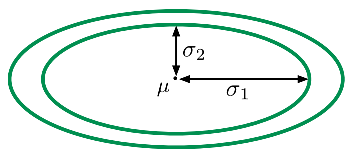
$$\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$$

(all off-diagonal elements zero).

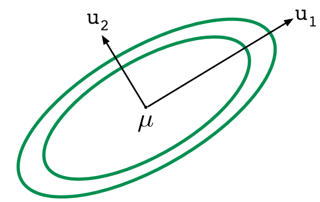
Each X_i is an independent univariate Gaussian $N(\mu_i, \sigma_i^2)$:

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

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:

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

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 θ is a constant depending on the various parameters.

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

Linear decision boundary

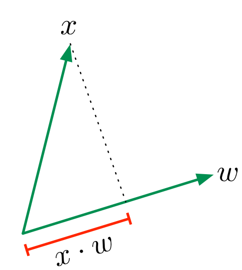
When $\Sigma_1 = \Sigma_2 = \Sigma$: choose class 1 iff

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

What does $x \cdot w$ (or equivalently $x^T w$, or $w^T x$) mean?

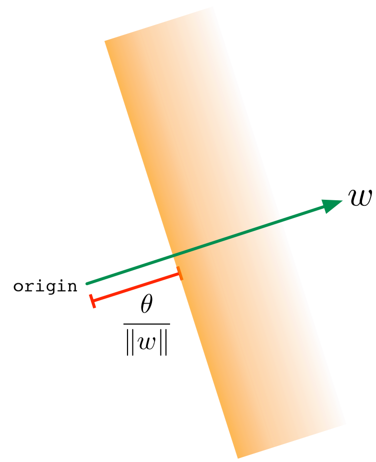
Algebraically: $x \cdot w = w \cdot x = x^T w = w^T x = \sum_{i=1}^{p} x_i w_i$

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

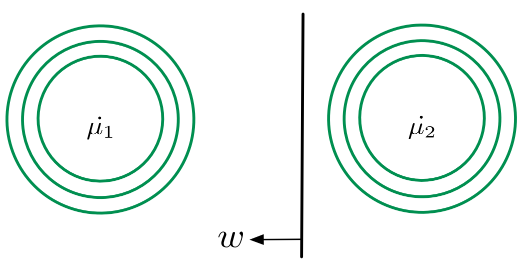


.inear decision boundary

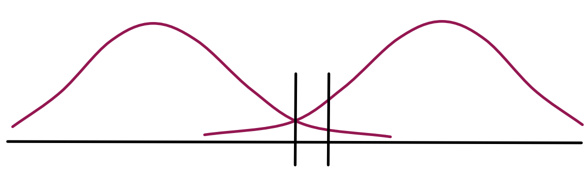
Let w be any vector in $\mathbb{R}^p$. What is meant by decision rule $w \cdot x \geq \theta$?



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



One-d projection onto w:

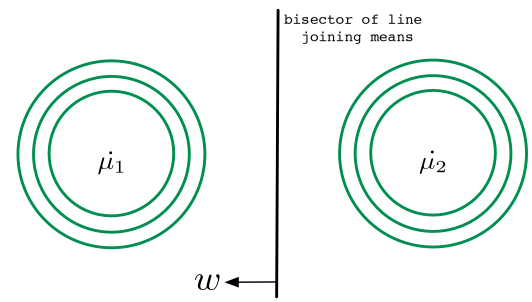


Common covariance: $\Sigma_1 = \Sigma_2 = \Sigma$

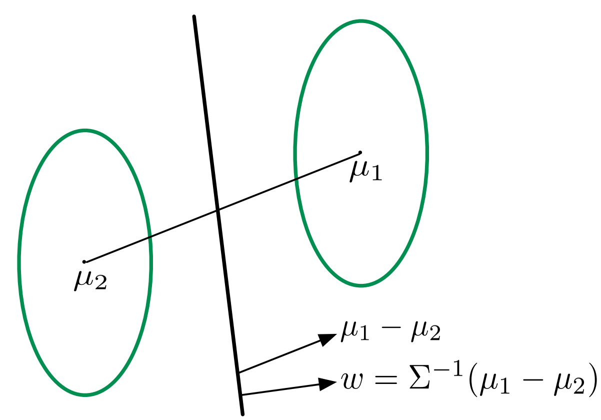
Linear decision boundary: choose class 1 iff

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

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



Example 3: Non-spherical.



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

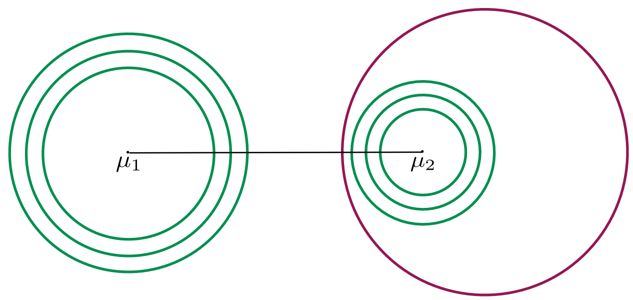
- w, θ dictated by probability model, assuming it is a perfect fit
- Common practice: choose w as above, but fit θ to minimize training/validation error

Different covariances: $\Sigma_1 \neq \Sigma_2$

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

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

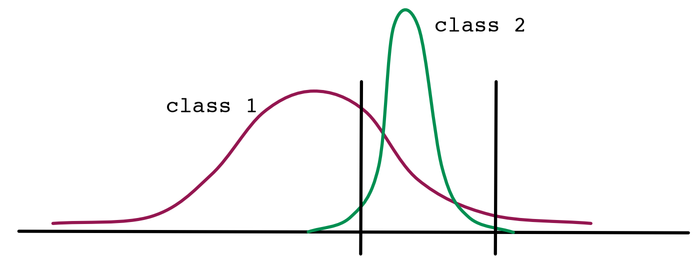
Example 1: $\Sigma_1 = \sigma_1^2 I_p$ and $\Sigma_2 = \sigma_2^2 I_p$ with $\sigma_1 > \sigma_2$



Example 3: A parabolic boundary.

Many other possibilities!

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



Multiclass discriminant analysis

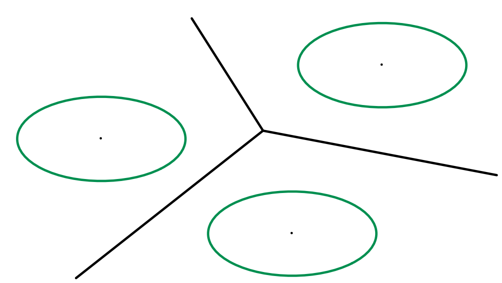
k classes: weights π_j, class-conditional distributions $P_j = N(\mu_j, \Sigma_j)$.

Each class has an associated **quadratic** function

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

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

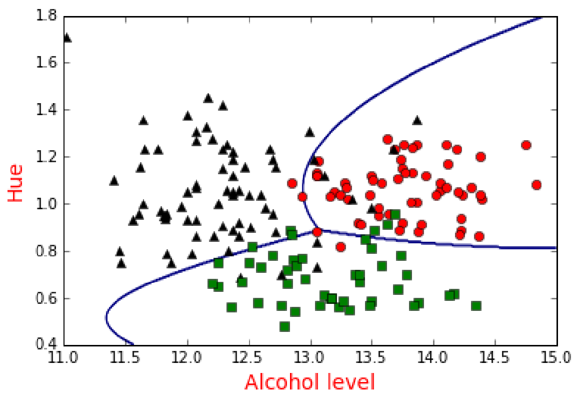
If $\Sigma_1 = \cdots = \Sigma_k$, the boundaries are **linear**.



:xample: "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 $\mu_j \in \mathbb{R}^{784}$
- covariance matrix $\Sigma_j \in \mathbb{R}^{784\times 784}$

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

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

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

- Divide original training set into a training set and a validation set.
- Fit parameters π_j, μ_j, Σ_j to training set
- Choose the constant c that yields lowest error rate on validation set

:isher's linear discriminant

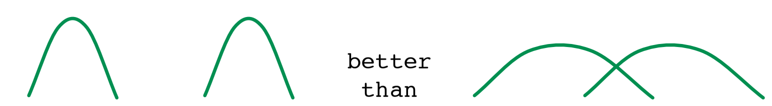
A framework for linear classification without Gaussian assumptions.

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.



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?