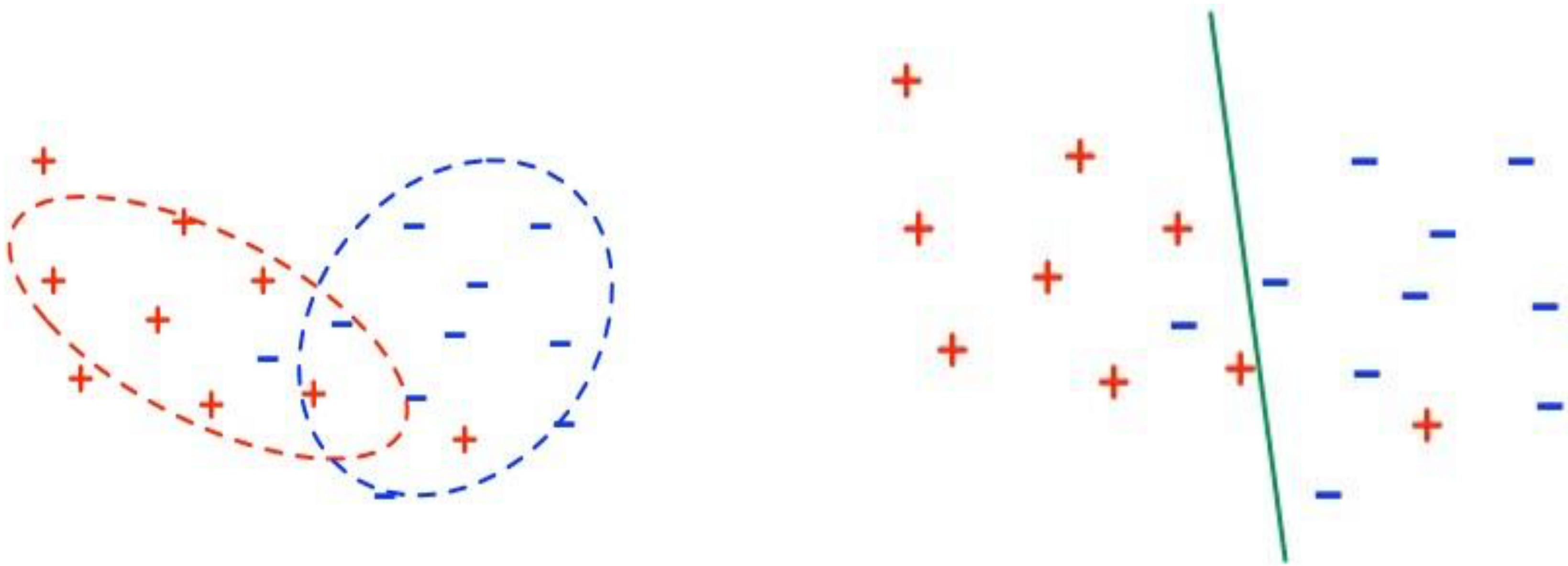
Classification with Generative Models

DSE 220

Classification with parametrized models

Classifiers with a fixed no. 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, R^p



Two ways to classify:

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

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

Example: Suppose cancer is related to age

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

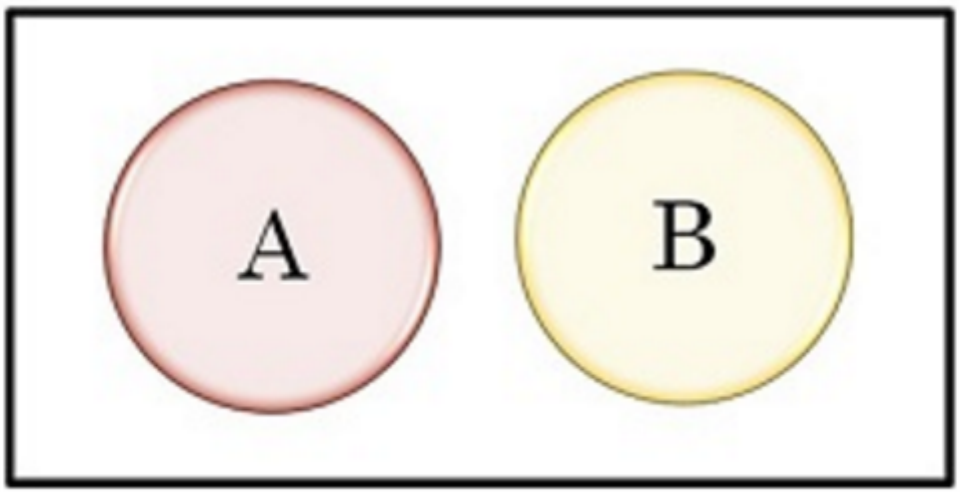
- Disjoint events cannot happen at the same time.
- e.g.: when tossing a coin, the result can either be heads or tails but cannot be both.

Independent

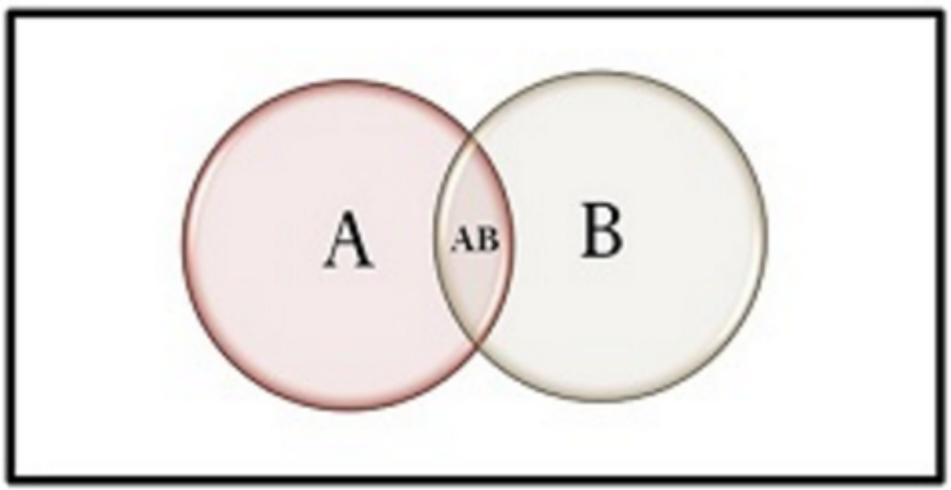
- Occurrence of one event does not influence the other(s).
- e.g.: when tossing two coins, the result of one flip does not affect the result of the other.

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,

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

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

Example: Sex bias in graduate admissions In 1969, there were 12673 applicants for graduate study at Berkeley. 44% of the male applicants were accepted, and 35% of the female applicants.

Over the sample space of applicants, define:

M = male

F = female

A = admitted

So: $\Pr(A|M) = 0.44$ and $\Pr(A|F) = 0.35$.

In every department, the accept rate for female applicants was at least as high as the accept rate for male applicants. How could this be?

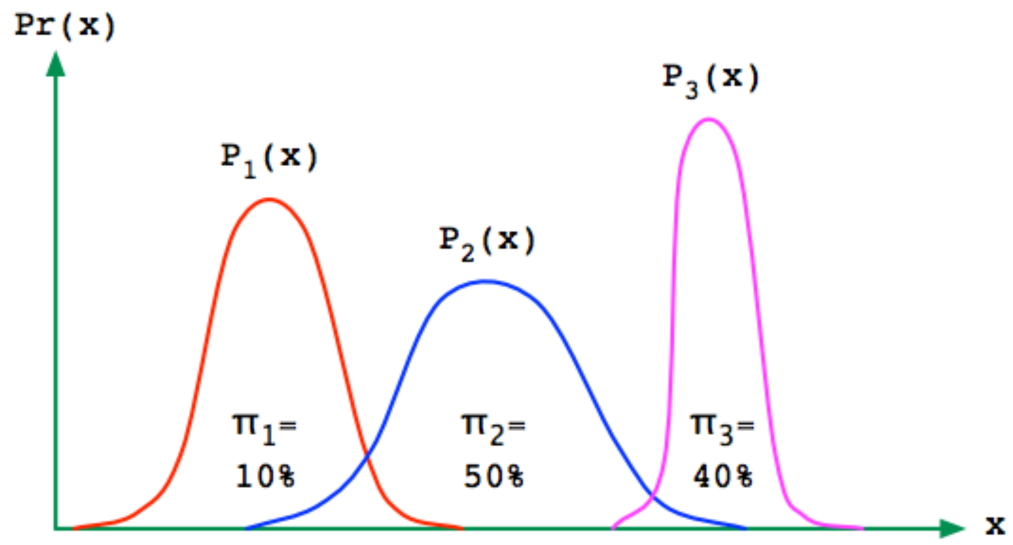
Generative models

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.

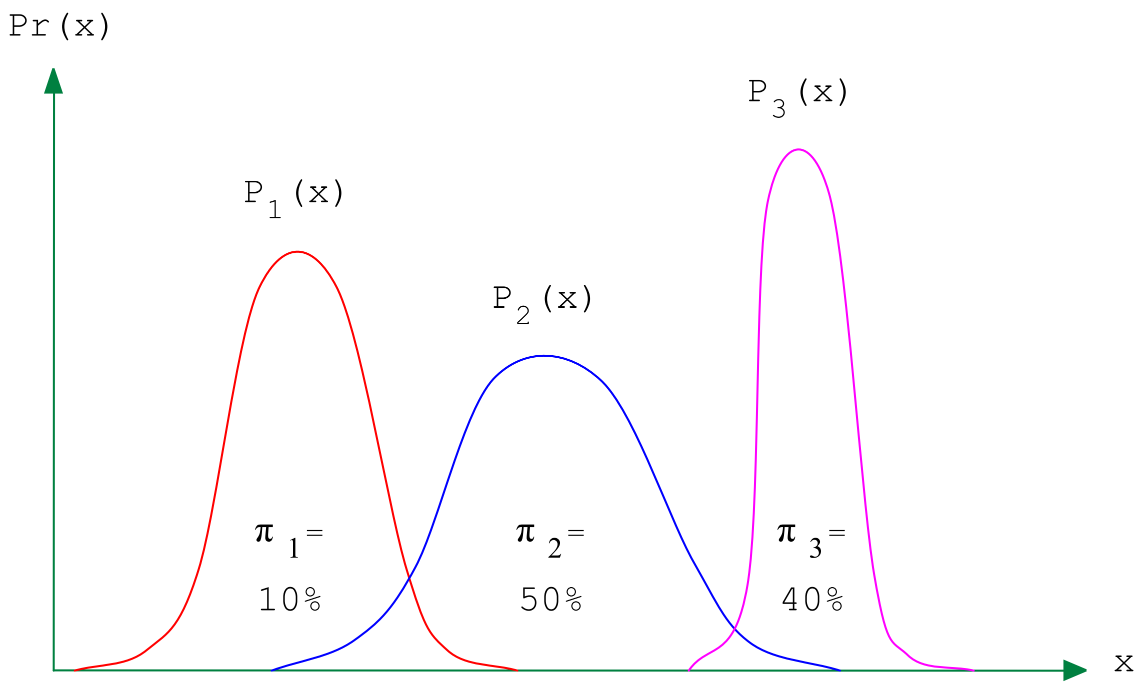
Example:
X = R
$Y = \{1, 2, 3\}$



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 $Y = \{1, 2, \dots, k\}$, 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 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.

Naive Bayes

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

$$P(feature_i\,,\ feature_j \mid label) = P(feature_i \mid label)\ .\ P(feature_i \mid label)$$

$$\overset{\text{posterior}}{p(label \mid features)} = \frac{\overset{\text{prior}}{p(label)}\,\overset{\text{likelihood}}{p(features \mid label)}}{\underset{\text{evidence}}{p(features)}}$$

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

Example: MNIST

Binarized MNIST:

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

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

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

Example: MNIST

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,

$$n_j = \text{\# of instances of class } j$$

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

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(θ) = probability of observing the given data as a function of θ.

Maximum Likelihood estimate of θ = value of θ that maximises Likelihood(θ).

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

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

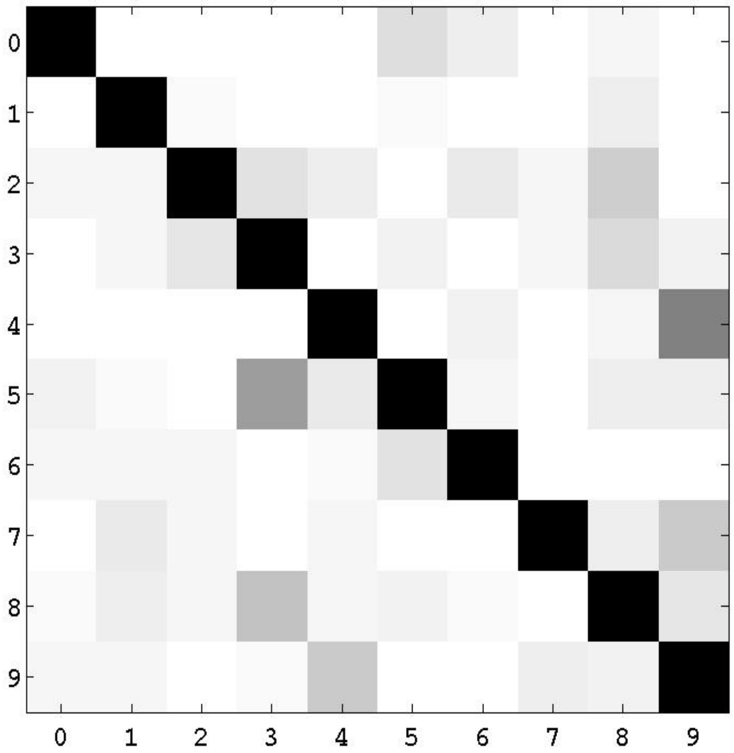
To avoid underflow: take the log:

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

A linear classifier!

Example: MNIST

Result of training: mean vectors for each class.



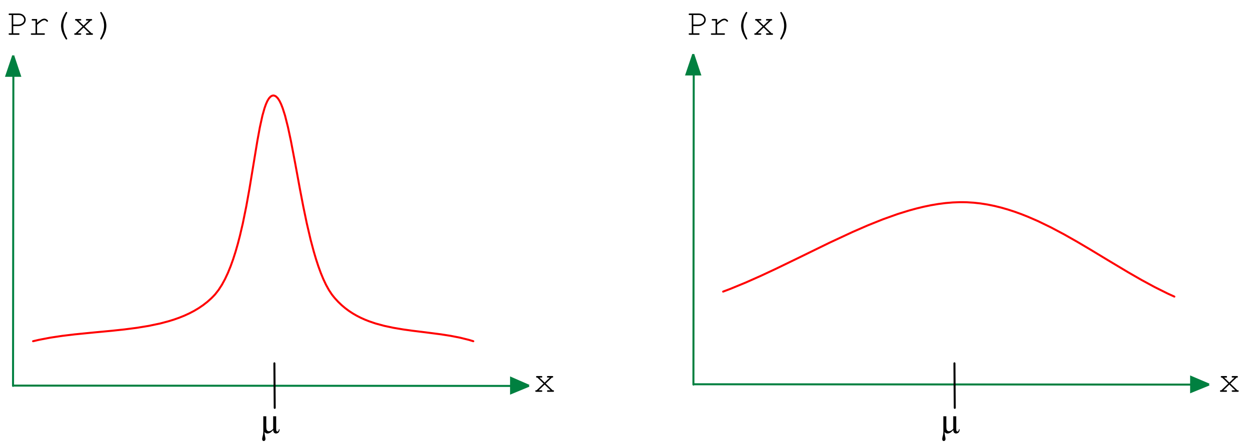
Test error rate: 15.54%.

Visualization of the “confusion matrix” →

Variance

If you had to summarize the entire distribution of a r.v. X by a single number, you would use the mean (or median). Call it μ.

But these don't capture the *spread* of X :



What would be a good measure of spread? How about the average distance away from the mean: $E(|X - \mu|)$?

For convenience, take the square instead of the absolute value.

$$\textbf{Variance:} \qquad \mathrm{var}(X) = \mathbb{E}(X-\mu)^2 = \mathbb{E}(X^2) - \mu^2,$$

where $\mu = \mathbb{E}(X)$. The variance is always ≥ 0.

Variance: example

Recall: $\mathrm{var}(X) = \mathbf{E}(X - \mu)^2 = \mathbf{E}(X^2) - \mu^2$, where $\mu = \mathbf{E}(X)$.

Toss a coin of bias p. Let $X \in \{0, 1\}$ be the outcome.

$$\mathbf{E}(X) = p$$

$$\mathbf{E}(X^2) = p$$

$$\mathbf{E}(X - \mu)^2 = p^2 \cdot (1 - p) + (1 - p)^2 \cdot p = p(1 - p)$$

$$\mathbf{E}(X^2) - \mu^2 = p - p^2 = p(1 - p)$$

This variance is highest when $p = 1/2$ (fair coin).

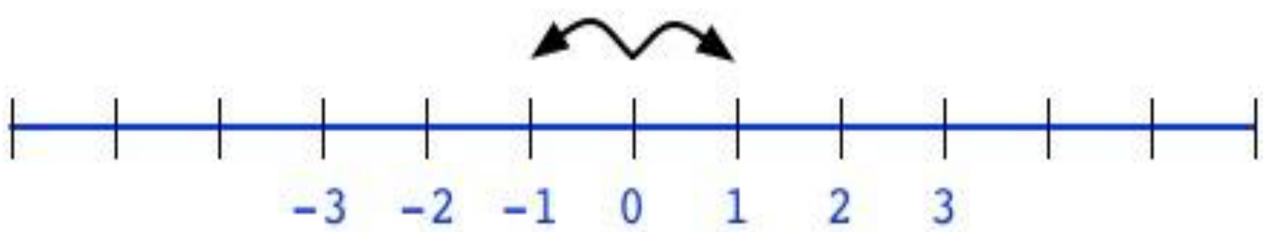
The standard deviation of X is $\mathrm{std}(X) = \sqrt{\mathrm{var}(X)}$.
It is the average amount by which X differs from its mean.

It is the average amount by which X differs from its mean.

Variance of a sum

var($X_1 + \cdots + X_k$) = var(X_1) + $\cdots$ + var(X_k) if the X_i are independent.

<u>Symmetric random walk</u>. A drunken man sets out from a bar. At each time step, he either moves one step to the right or one step to the left, with equal probabilities. Roughly where is he after n steps?

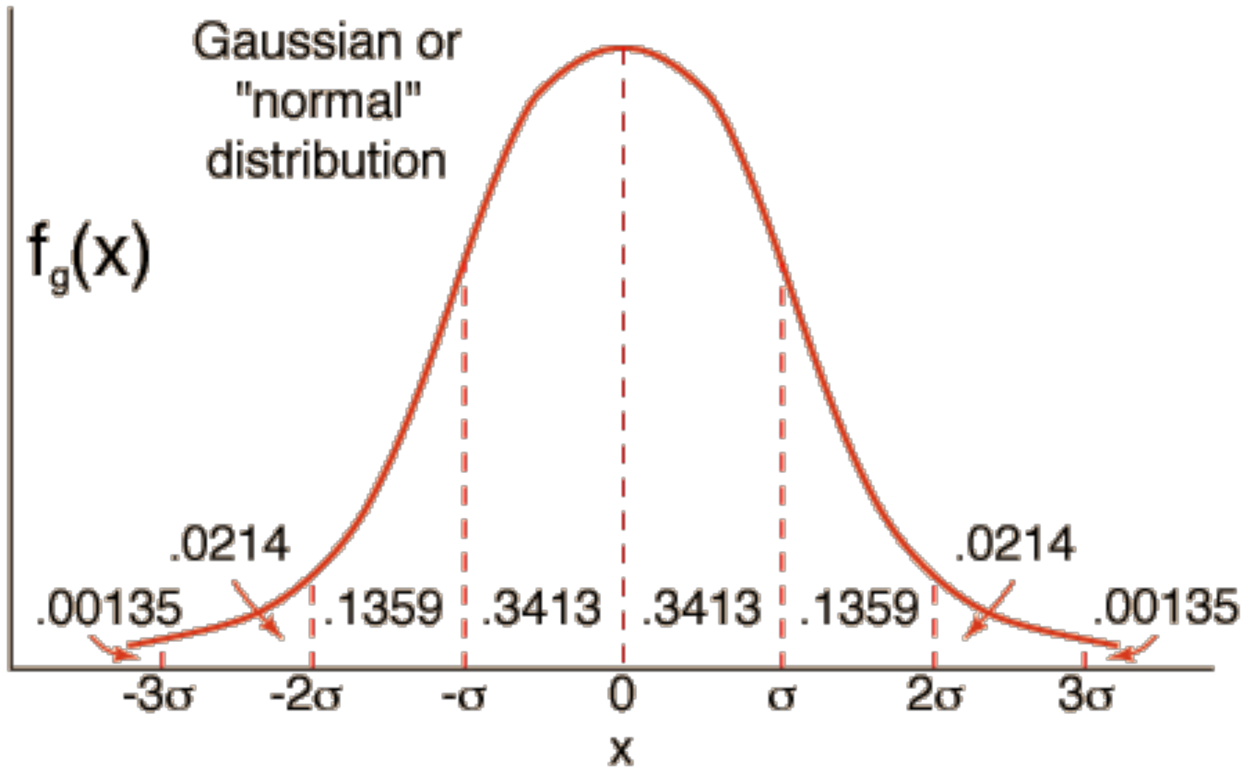


Let $X_i \in \{-1, 1\}$ be his i^{th} step. Then E(X_i) = ?0 and var(X_i) = ?1.

His position after n steps is $X = X_1 + \ldots + X_n$.

$$\mathbb{E}(X) = 0$$
$$\text{var}(X) = n$$
$$\text{stddev}(X) = \sqrt{n}$$

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

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

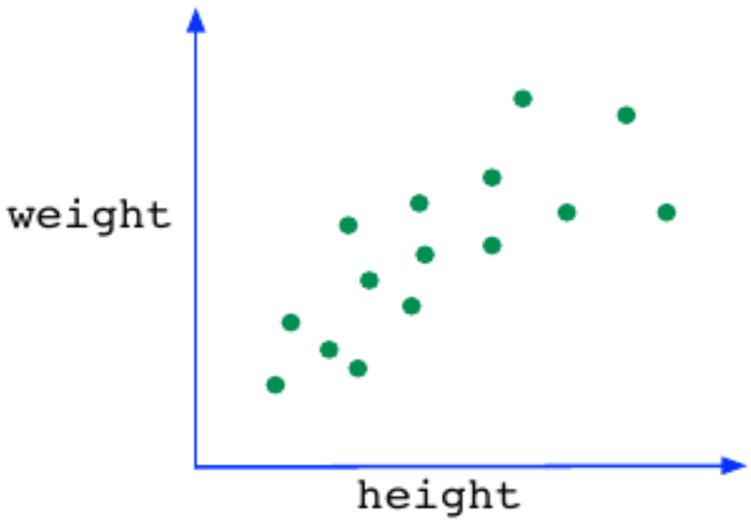
Independence would mean

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

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

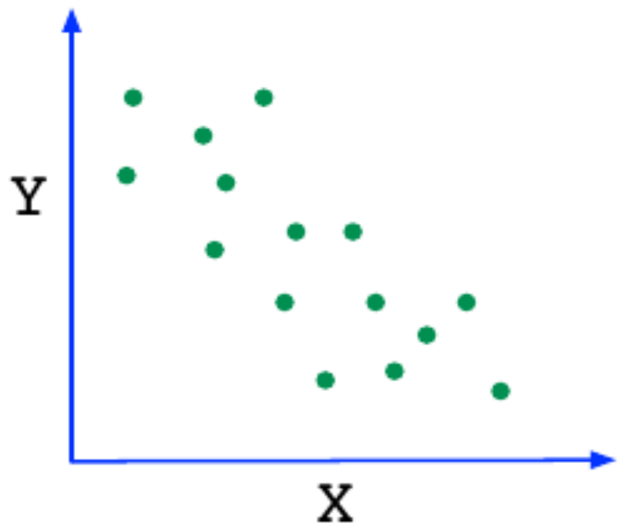
Is this an accurate approximation?
No: we'd expect height and weight to be **positively correlated**.

Types of correlation

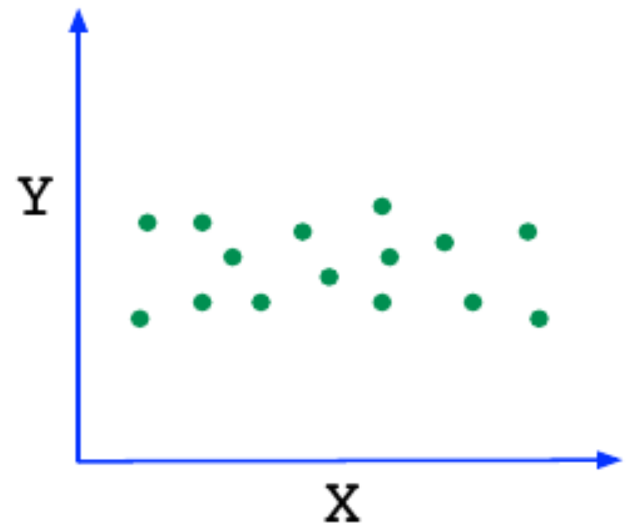


H, W positively correlated.
This also implies

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

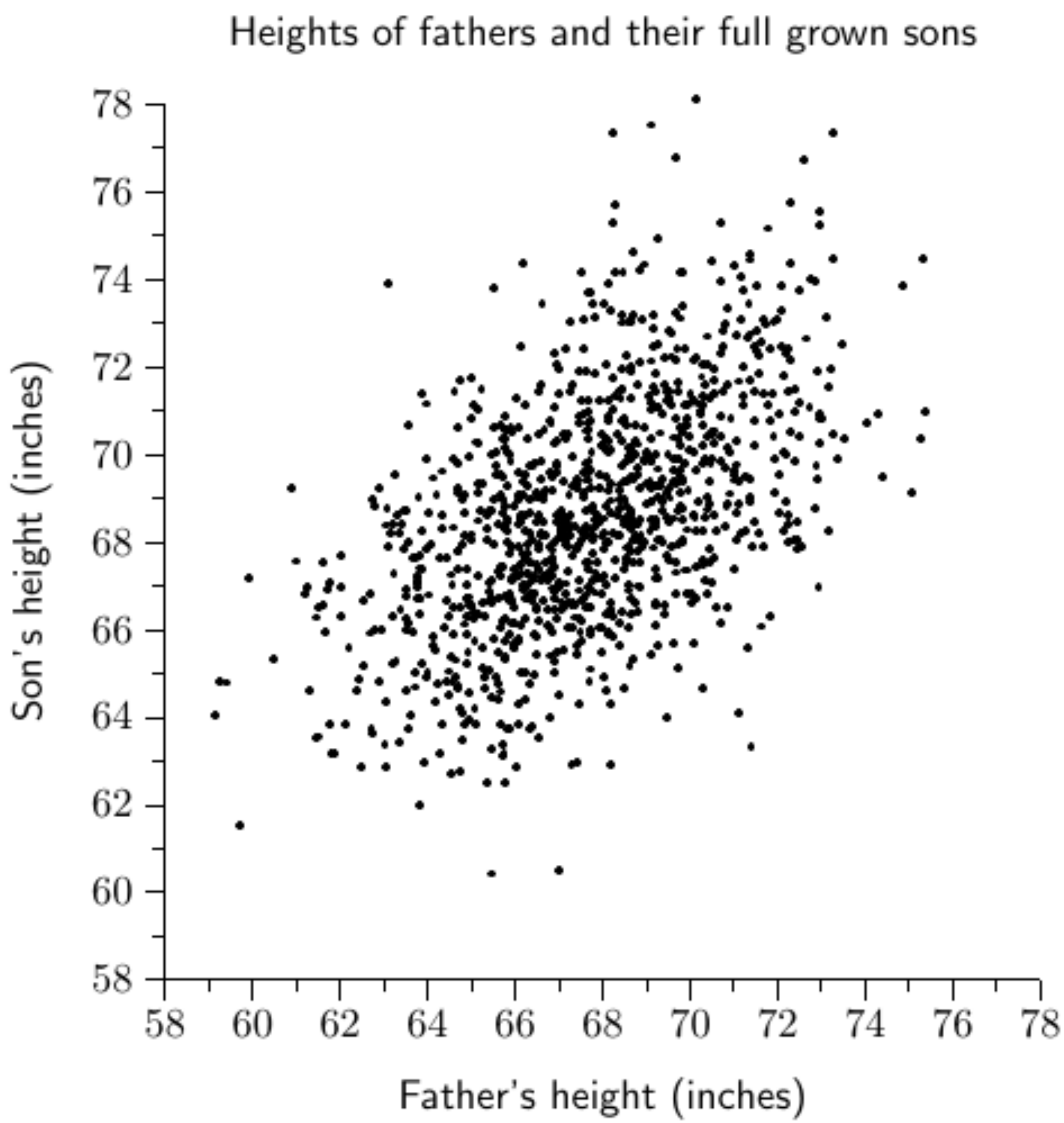


X, Y negatively correlated



X, Y uncorrelated

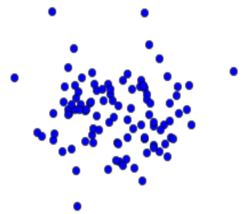
Pearson (1903): fathers and sons



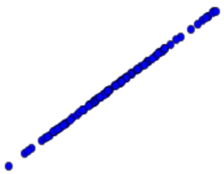
How to quantify the degree of correlation?

Correlation pictures

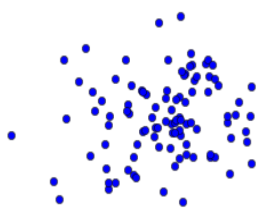
$r = 0$



$r = 1$

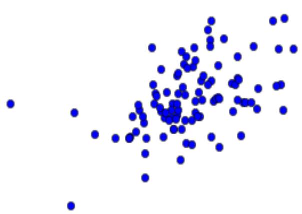


$r = 0.25$

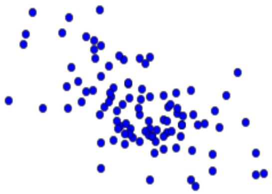


$r = -0.25$

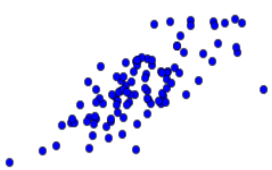
$r = 0.5$



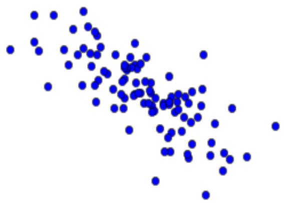
$r = -0.5$



$r = 0.75$



$r = -0.75$



Covariance and correlation

Suppose X has mean μ_x and Y has mean μ_Y

- Covariance

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

Maximized when $X = Y$, in which case it is 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

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

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

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

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

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	$1/6$

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

In this case, X and Y are negatively correlated.

The bivariate (2-d) Gaussian

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

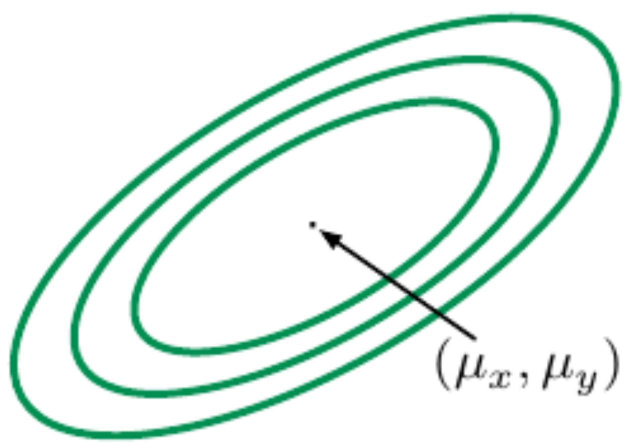
- **Mean** $(\mu_x, \mu_y) \in \mathbb{R}^2$
- **Covariance matrix**

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

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

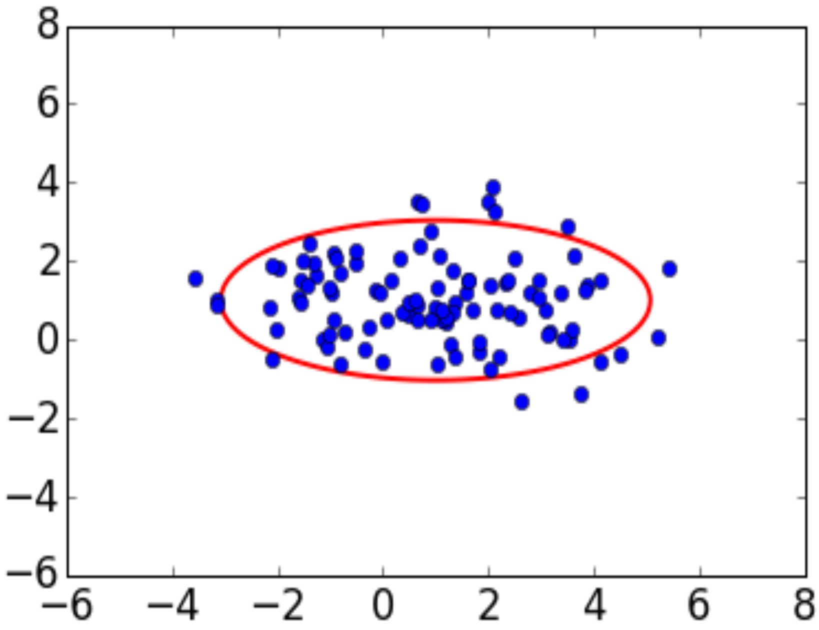
Density $$p(x, y) = \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}\begin{bmatrix} x - \mu_x \\ y - \mu_y \end{bmatrix}^T \Sigma^{-1} \begin{bmatrix} x - \mu_x \\ y - \mu_y \end{bmatrix}\right)$$

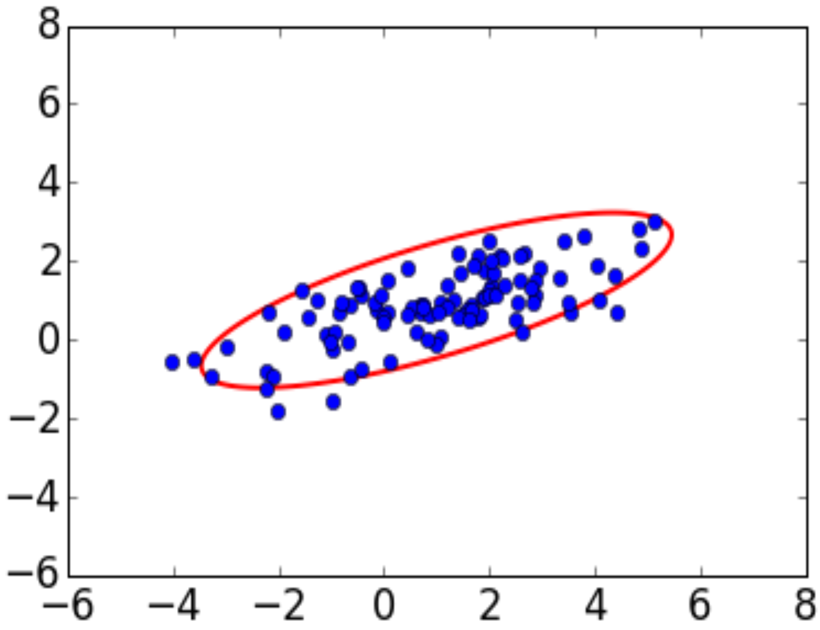
The density is highest at the mean, and falls off in ellipsoidal contours.



Bivariate Gaussian: examples

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

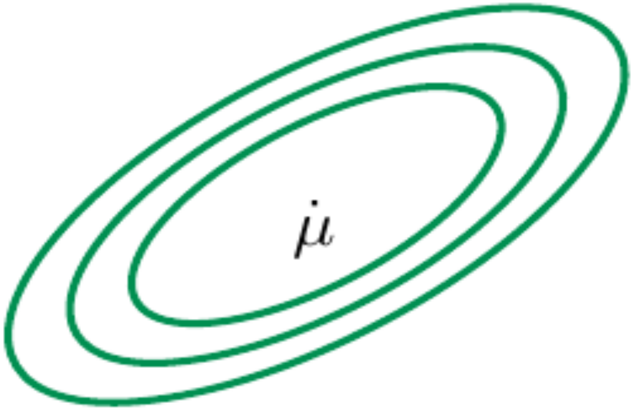




$$\Sigma = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}$$

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

The multivariate Gaussian



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

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

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

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:

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

- 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

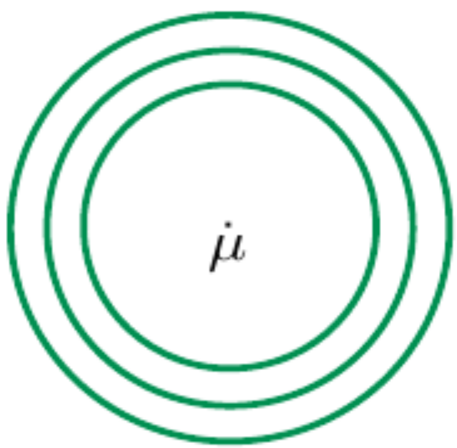
$$\Sigma = \sigma^2 I_p = \text{diag}(\sigma^2, \sigma^2, \ldots, \sigma^2)$$

(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

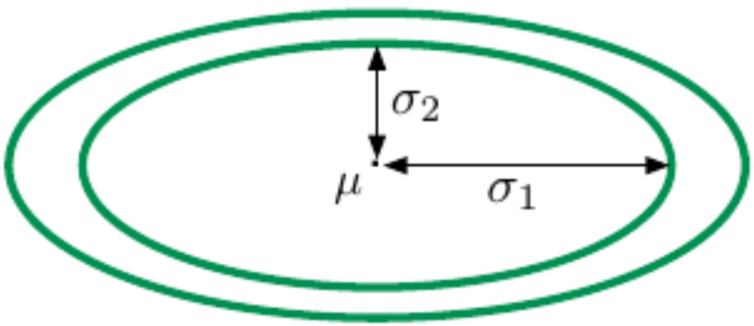
$$\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$$

(all off-diagonal elements zero).

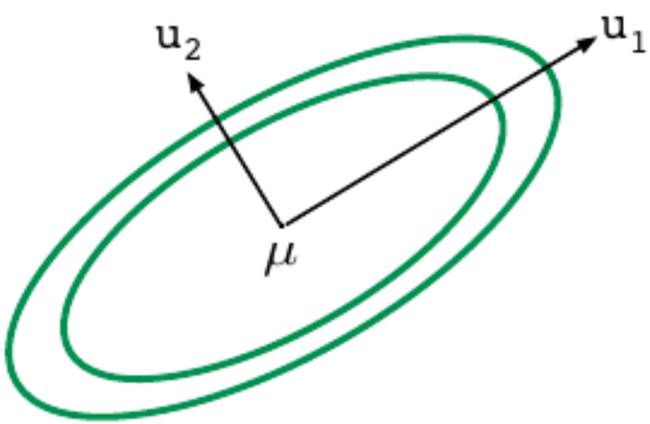
Each X_i is an independent univariate Gaussian $N(\mu_i, \sigma_i^2)$:

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

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:

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

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.

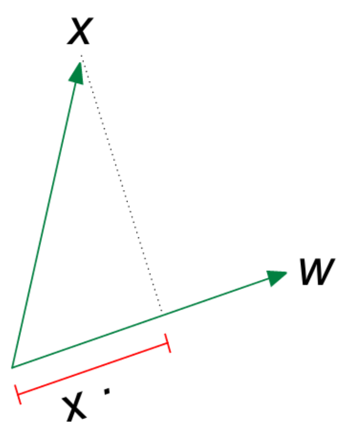
$\Sigma_1 = \Sigma_2$: linear decision boundary. Otherwise, quadratic boundary.

Linear decision boundary

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

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

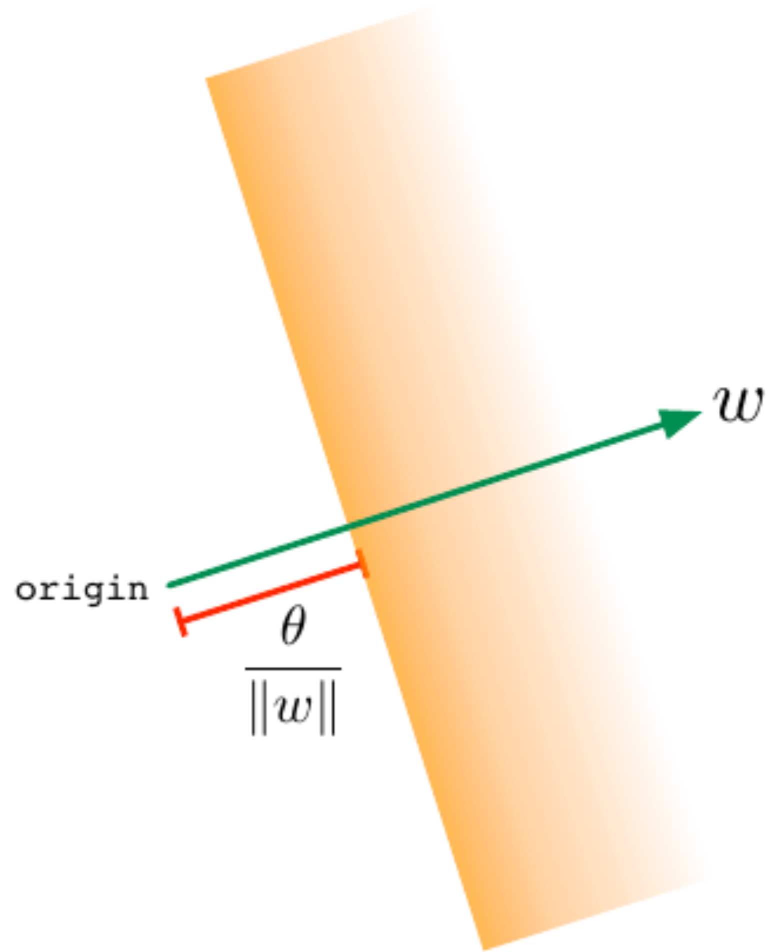
Geometric picture: Suppose *w* is a unit vector (that is, $||w|| = 1$). Then $x \cdot w$ is the **projection** of vector *x* onto direction *w* .



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 $w \cdot x \geq \theta$?

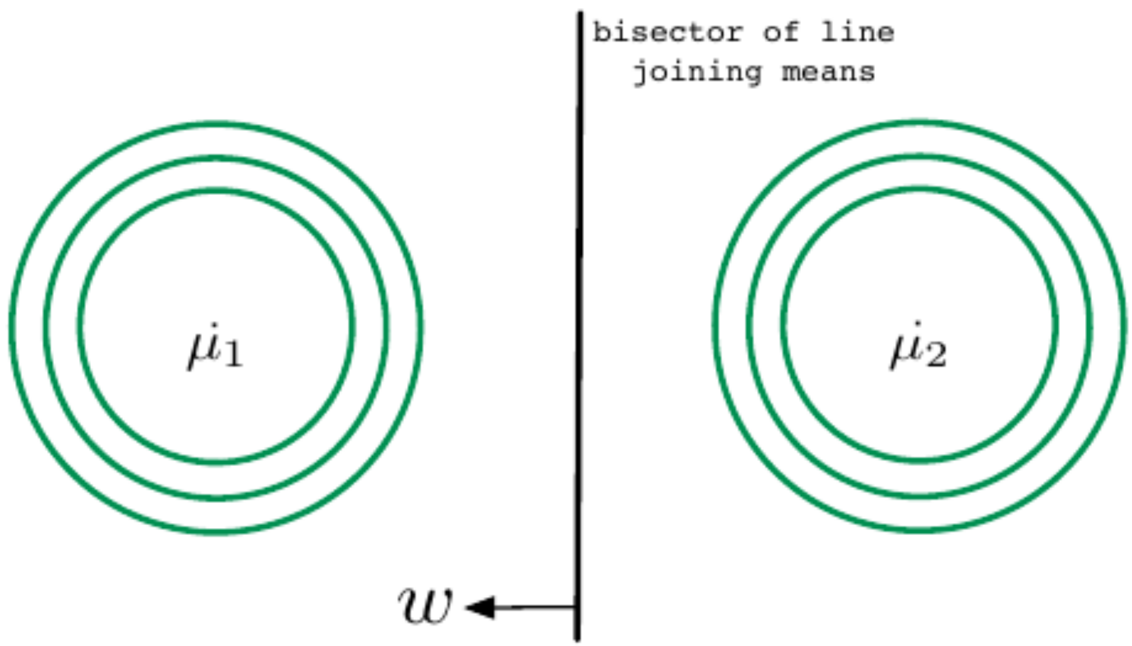


Common covariance: $\Sigma_1 = \Sigma_2 = \Sigma$

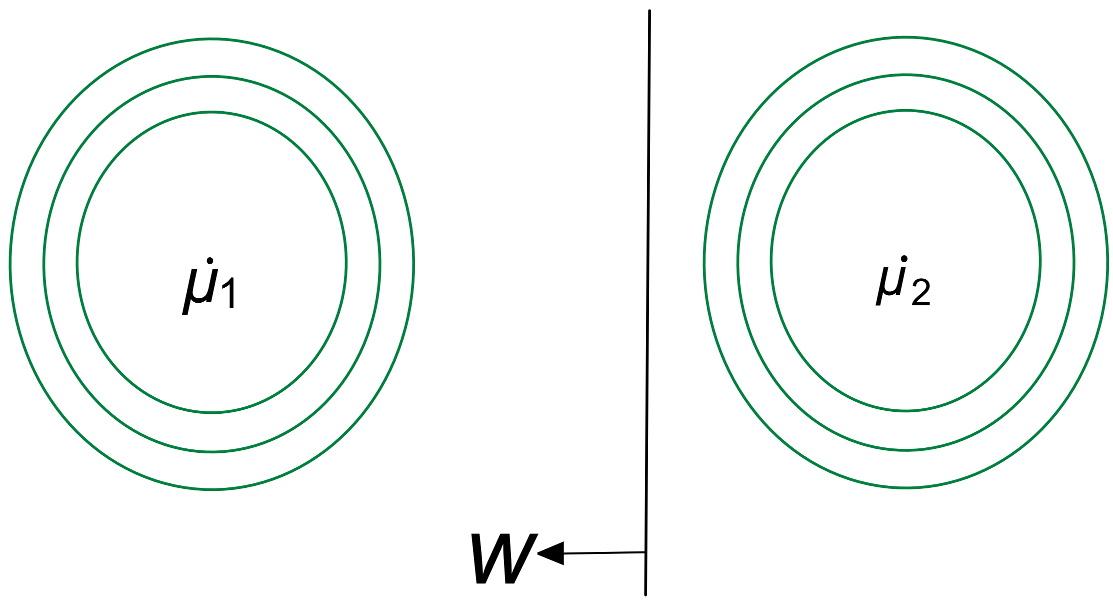
Linear decision boundary: choose class 1 iff

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

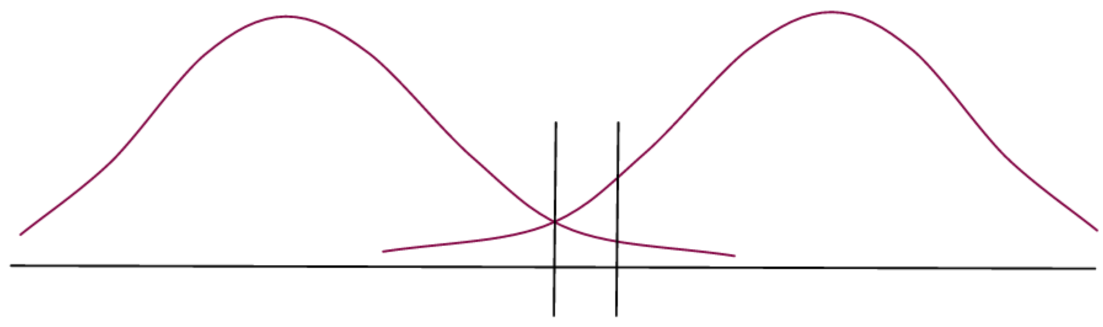
Example 1: Spherical Gaussians with $\Sigma = I_p$ and $\pi_1 = \pi_2$.



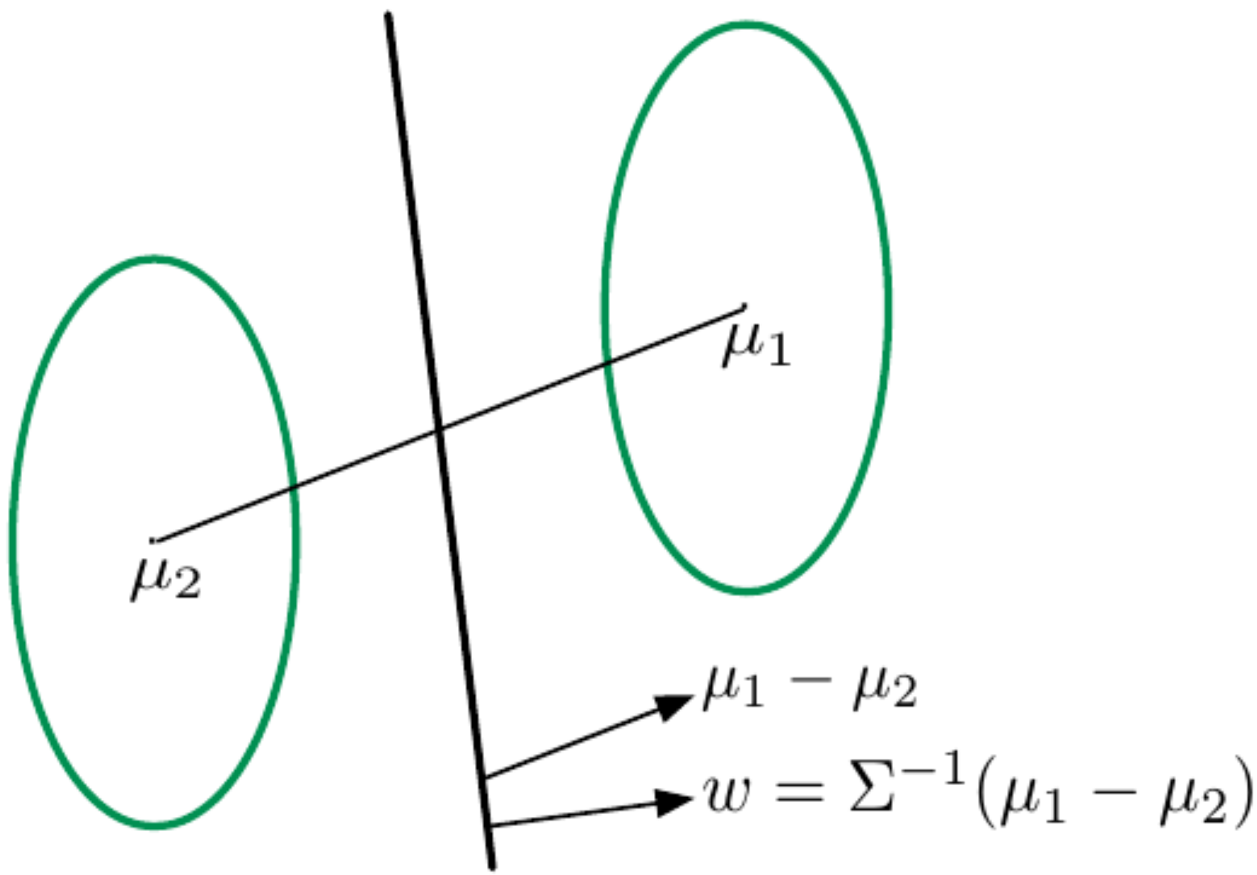
Example 2: Again spherical, but now $\pi_1 > \pi_2$.



One-d projection onto w :



Example 3: Non-spherical.



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

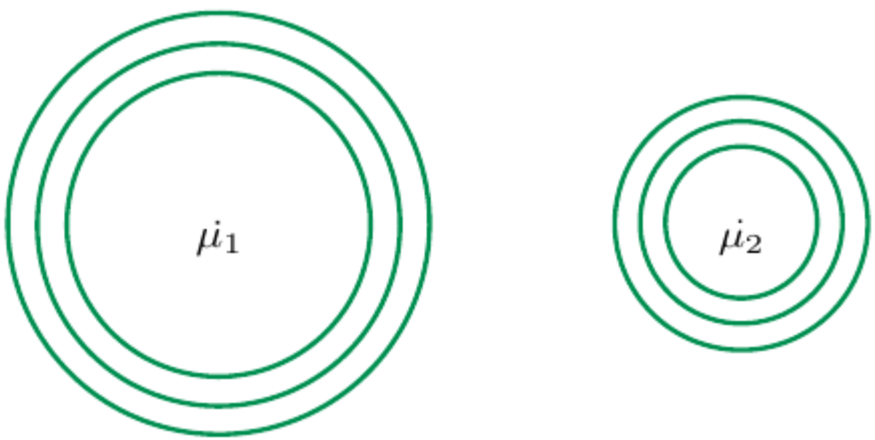
- w, θ dictated by probability model, assuming it is a perfect fit
- Common practice: choose w as above, but fit θ to minimize training/validation error

Different covariances: $\Sigma_1 \neq \Sigma_2$

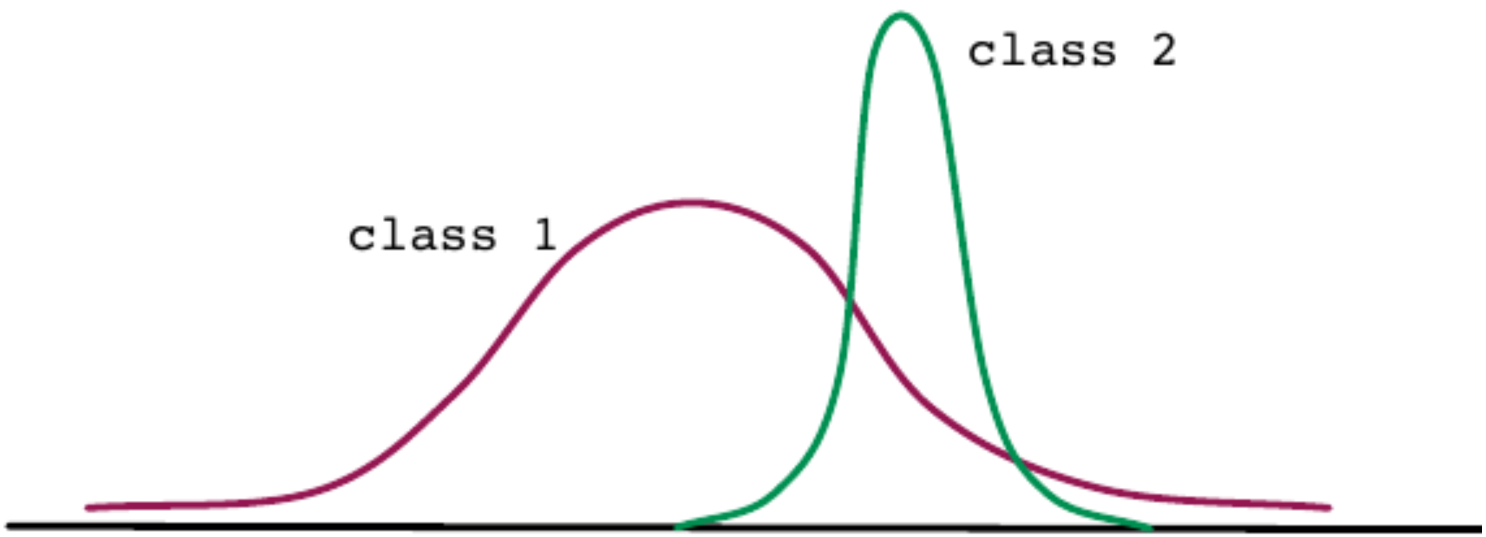
Quadratic boundary: choose class 1 iff $x^T M x + 2w^T x \geq \theta$, where:

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

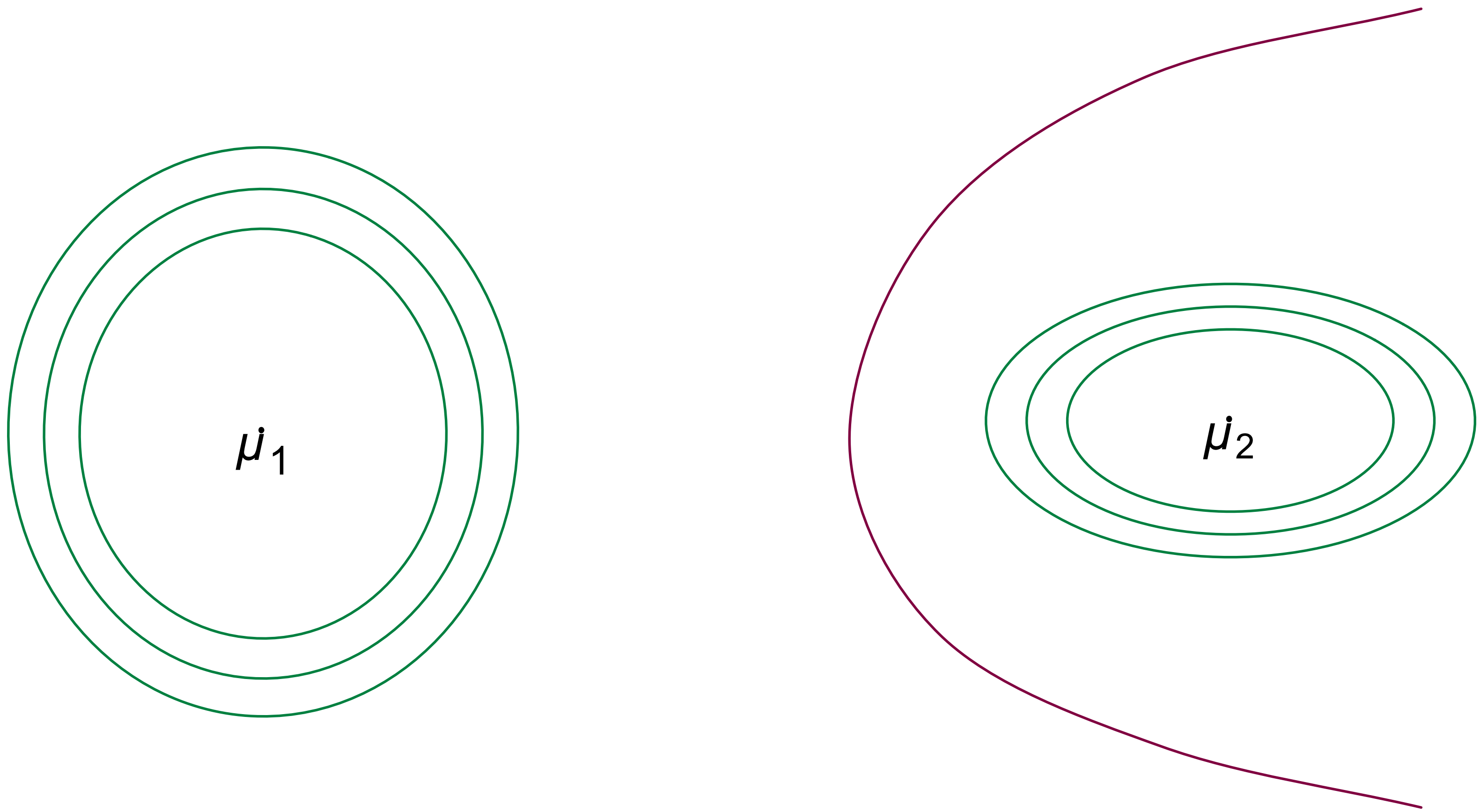
Example 1: $\Sigma_1 = \sigma_1^2 I_p$ and $\Sigma_2 = \sigma_2^2 I_p$ with $\sigma_1 > \sigma_2$



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



Example 3: A parabolic boundary.



Many other possibilities!

Multiclass discriminant analysis

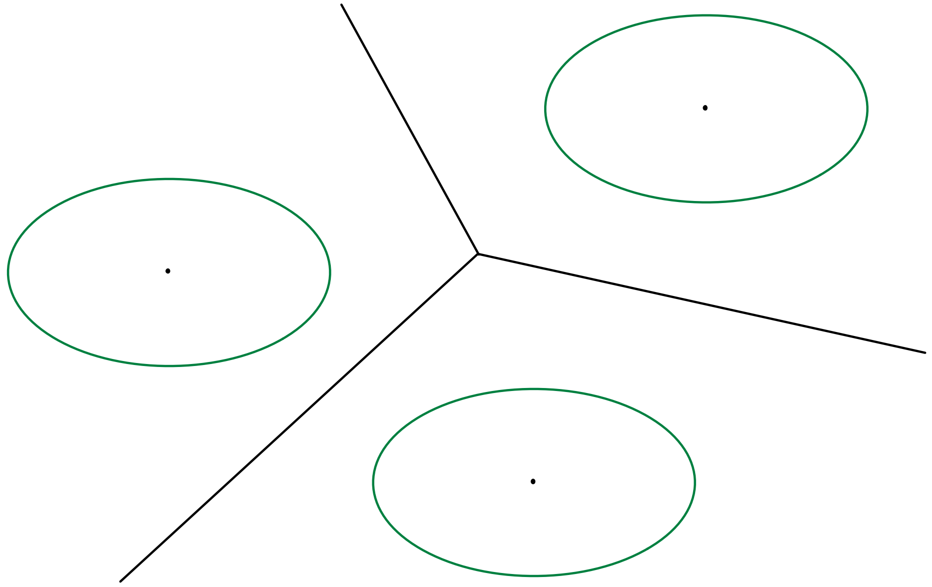
k classes: weights π_j, class-conditional distributions $P_j = N(\mu_j, \Sigma_j)$

Each class has an associated **quadratic** function

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

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

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

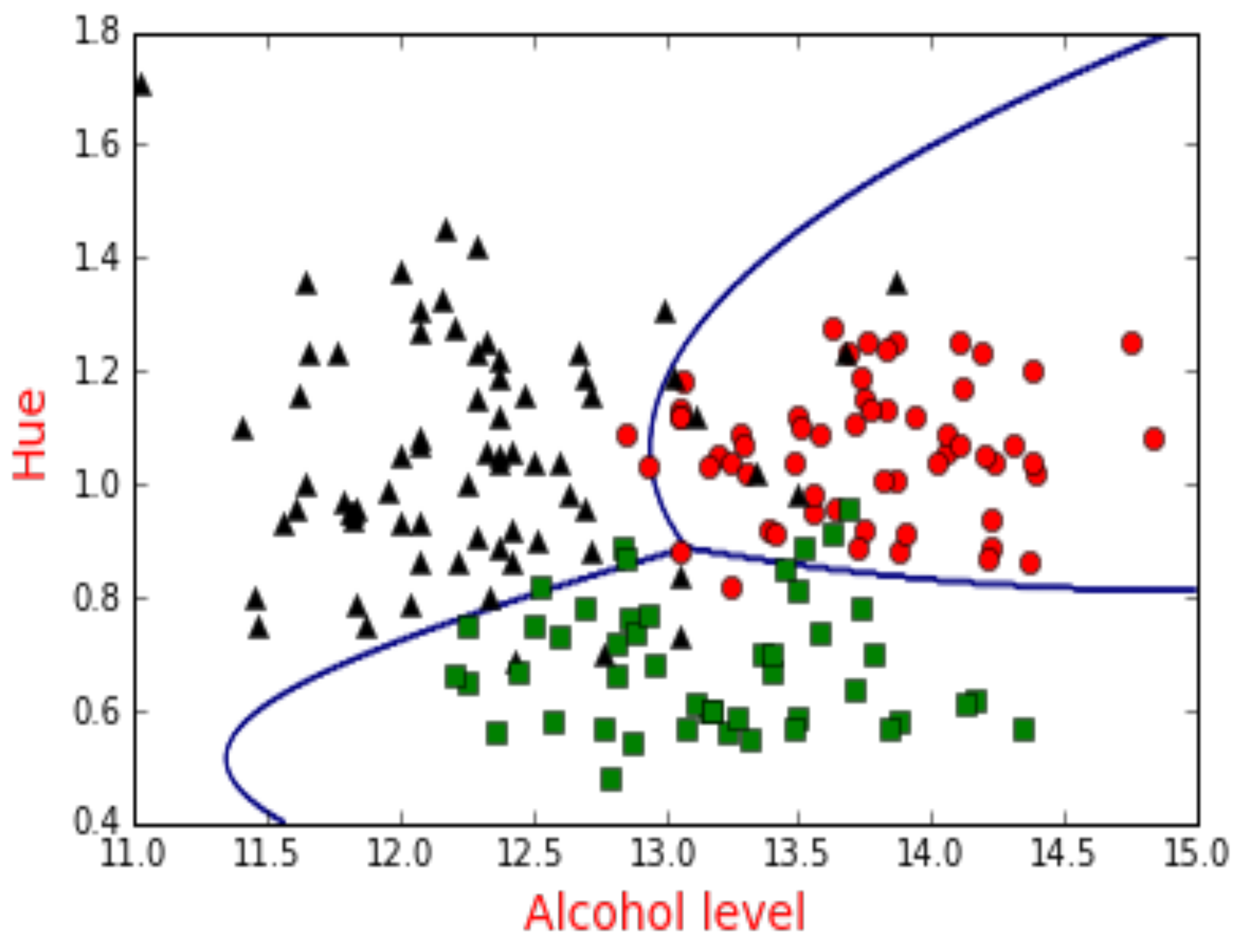


Example: “wine” data set

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

Error rate with regularization: **???**

Fisher's linear discriminant

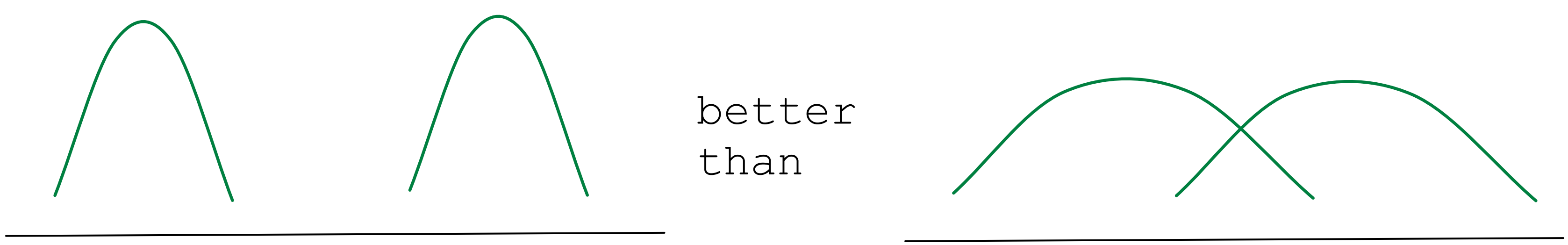
A framework for linear classification without Gaussian assumptions.

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.



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?

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: $w = \Sigma^{-1/2} u \propto \Sigma^{-1}(\mu_1 - \mu_2)$.