

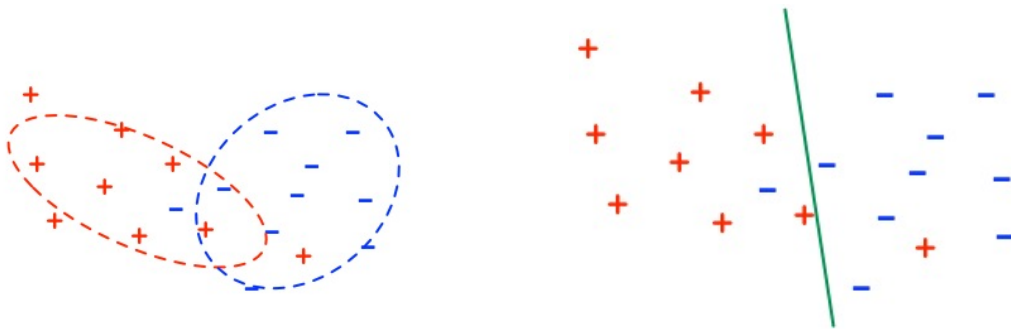
Classification with generative models 2

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.

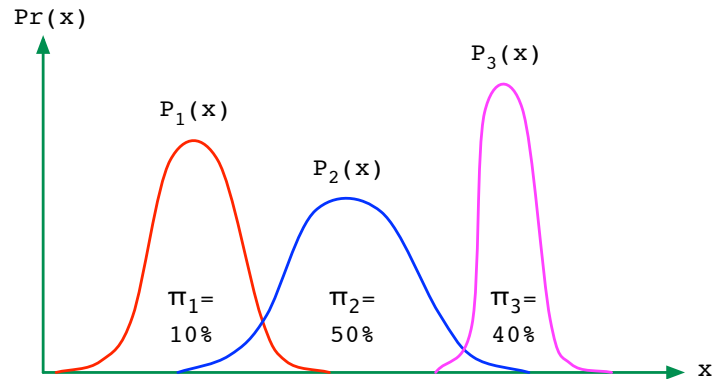
Typically the x 's are points in d -dimensional Euclidean space, \mathbb{R}^d .



Two ways to classify:

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

The Bayes-optimal prediction



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

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: $h^*(x) = \arg \max_j \pi_j P_j(x)$.

The winery prediction problem

Which winery is it from, 1, 2, or 3?



Using one feature ('Alcohol'), error rate is 29%.

What if we use **two** features?

The data set, again

Training set obtained from 130 bottles

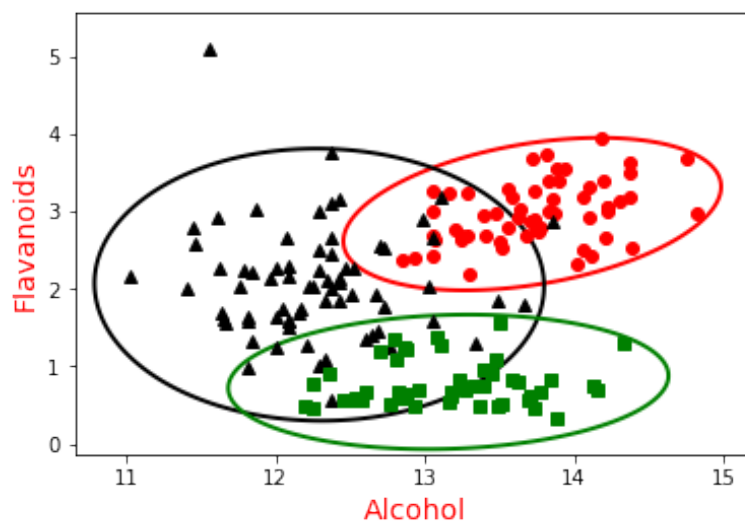
- Winery 1: 43 bottles
- Winery 2: 51 bottles
- Winery 3: 36 bottles
- For each bottle, 13 features:
'Alcohol', 'Malic acid', 'Ash', 'Alcalinity of ash', 'Magnesium',
'Total phenols', 'Flavanoids', 'Nonflavanoid phenols',
'Proanthocyanins',
'Color intensity', 'Hue', 'OD280/OD315 of diluted wines', 'Proline'

Also, a separate test set of 48 labeled points.

This time: 'Alcohol' and 'Flavanoids'.

Why it helps to add features

Better **separation** between the classes!



Error rate drops from 29% to 8%.

Bivariate distributions

Simplest option: treat each variable as independent.

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

H = height

W = weight

Independence would mean

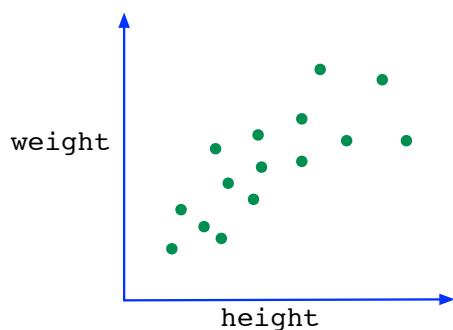
$$\Pr(H = h, W = w) = \Pr(H = h) \Pr(W = w),$$

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

Is this an accurate approximation?

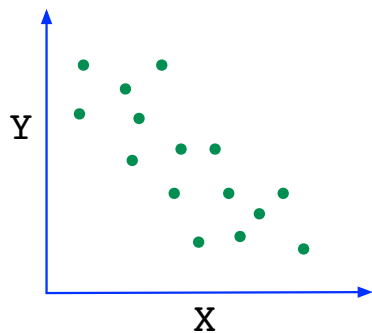
No: we'd expect height and weight to be **positively correlated**.

Types of correlation

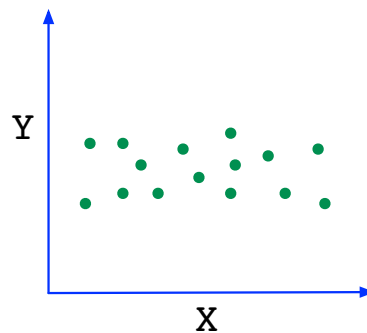


H, W positively correlated.
This also implies

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

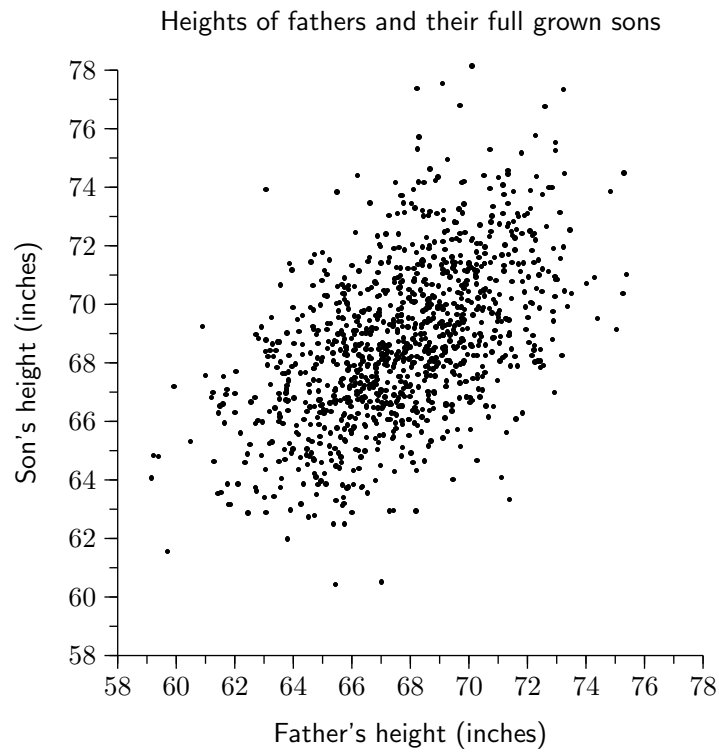


X, Y negatively correlated



X, Y uncorrelated

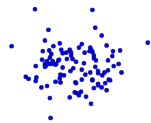
Pearson (1903): fathers and sons



How to quantify the degree of correlation?

Correlation pictures

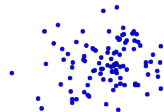
$$r = 0$$



$$r = 1$$



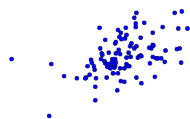
$$r = 0.25$$



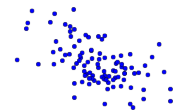
$$r = -0.25$$



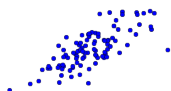
$$r = 0.5$$



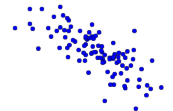
$$r = -0.5$$



$$r = 0.75$$



$$r = -0.75$$



Covariance and correlation

Suppose X has mean μ_X and Y has mean μ_Y .

- Covariance

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

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 $[-1, 1]$.

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

$$\mu_X = 0$$

$$\mu_Y = -1/3$$

$$\text{var}(X) = 1$$

$$\text{var}(Y) = 8/9$$

$$\text{cov}(X, Y) = 0$$

$$\text{corr}(X, Y) = 0$$

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

$$\mu_X = 0$$

$$\mu_Y = 0$$

$$\text{var}(X) = 1$$

$$\text{var}(Y) = 100$$

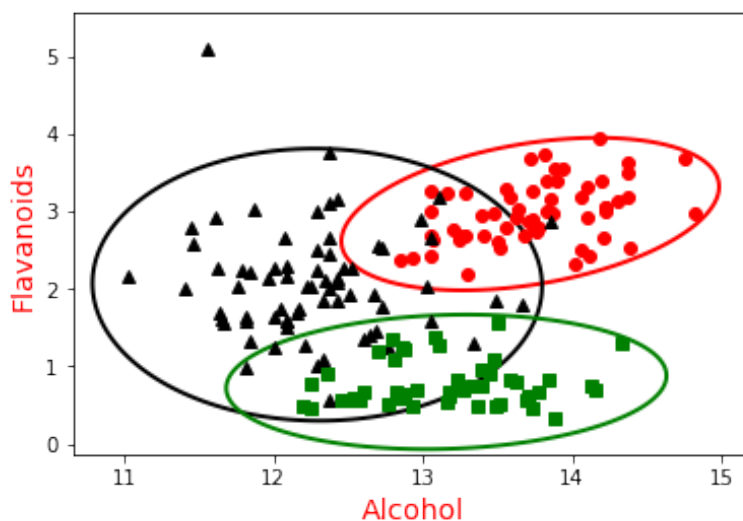
$$\text{cov}(X, Y) = -10/3$$

$$\text{corr}(X, Y) = -1/3$$

In this case, X and Y are negatively correlated.

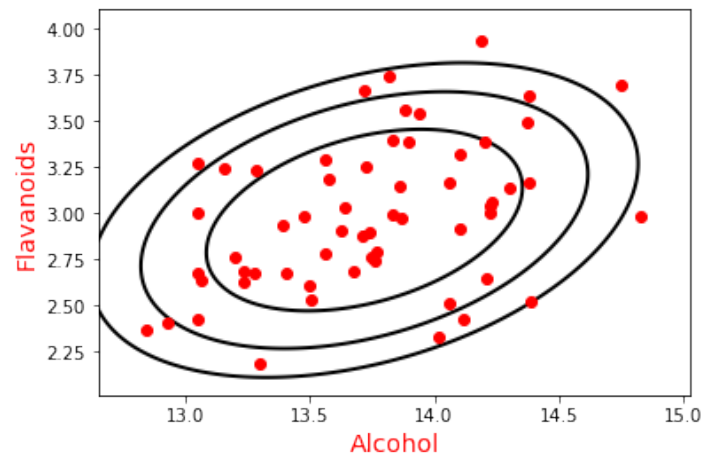
Return to winery example

Better **separation** between the classes!



Error rate drops from 29% to 8%.

The bivariate Gaussian



Model class 1 by a bivariate Gaussian, parametrized by:

$$\text{mean } \mu = \begin{pmatrix} 13.7 \\ 3.0 \end{pmatrix} \text{ and covariance matrix } \Sigma = \begin{pmatrix} 0.20 & 0.06 \\ 0.06 & 0.12 \end{pmatrix}$$

The bivariate (2-d) Gaussian

A distribution over $(x_1, x_2) \in \mathbb{R}^2$, parametrized by:

- **Mean** $(\mu_1, \mu_2) \in \mathbb{R}^2$, where $\mu_1 = \mathbb{E}(X_1)$ and $\mu_2 = \mathbb{E}(X_2)$
- **Covariance matrix** $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$ where
$$\left\{ \begin{array}{l} \Sigma_{11} = \text{var}(X_1) \\ \Sigma_{22} = \text{var}(X_2) \\ \Sigma_{12} = \Sigma_{21} = \text{cov}(X_1, X_2) \end{array} \right\}$$

Density is highest at the mean, falls off in ellipsoidal contours.

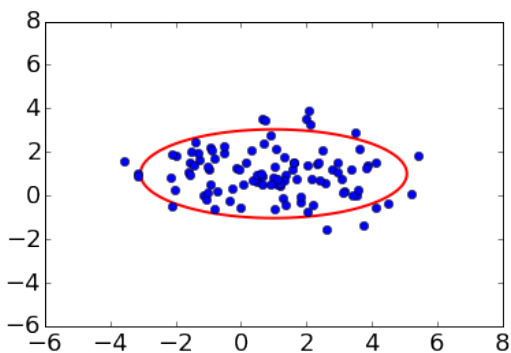
Density of the bivariate Gaussian

- **Mean** $(\mu_1, \mu_2) \in \mathbb{R}^2$, where $\mu_1 = \mathbb{E}(X_1)$ and $\mu_2 = \mathbb{E}(X_2)$
- **Covariance matrix** $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$

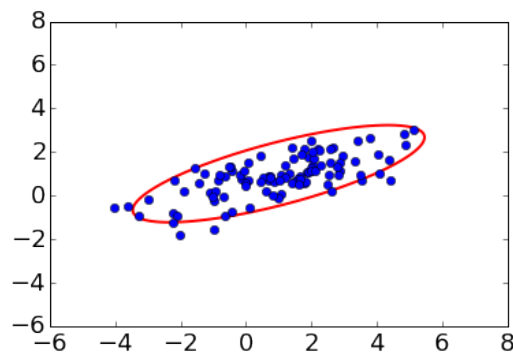
$$\text{Density } p(x_1, x_2) = \frac{1}{2\pi|\Sigma|^{1/2}} \exp \left(-\frac{1}{2} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}^T \Sigma^{-1} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix} \right)$$

Bivariate Gaussian: examples

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



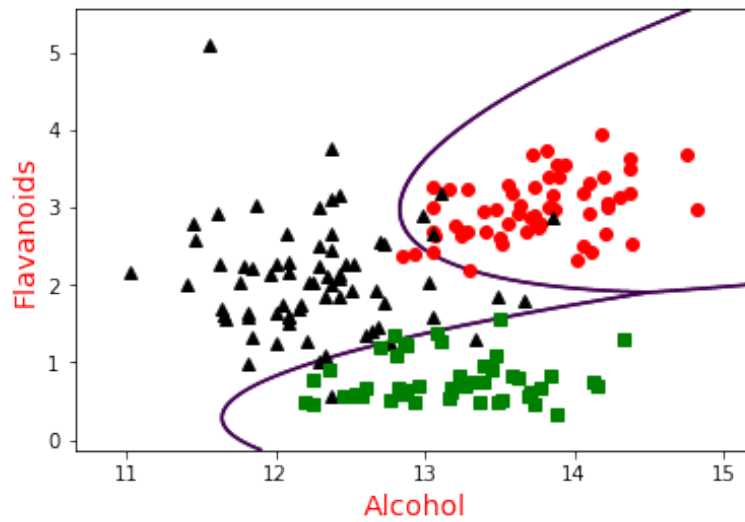
$$\Sigma = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}$$



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

The decision boundary

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



What kind of function is this? And, can we use more features?

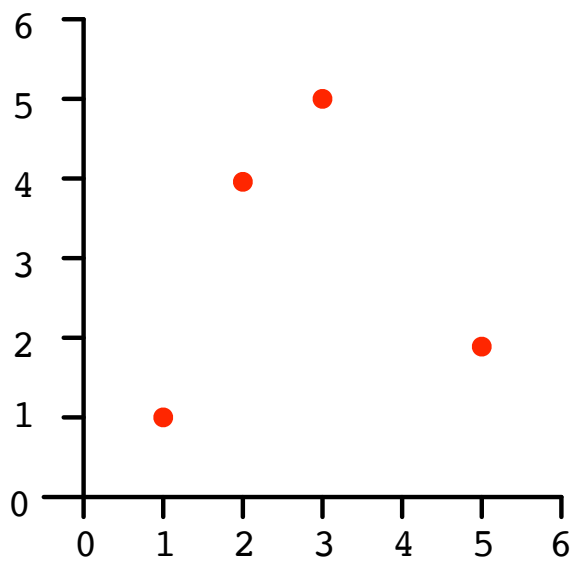
Worksheet 6 — Generative models 2

1. Would you expect the following pairs of random variables to be uncorrelated, positively correlated, or negatively correlated?
 - (a) The weight of a new car and its price.
 - (b) The weight of a car and the number of seats in it.
 - (c) The age in years of a second-hand car and its current market value.
2. Consider a population of married couples in which every wife is exactly 0.9 of her husband's age. What is the correlation between husband's age and wife's age?
3. Each of the following scenarios describes a joint distribution (x, y) . In each case, give the parameters of the (unique) bivariate Gaussian that satisfies these properties.
 - (a) x has mean 2 and standard deviation 1, y has mean 2 and standard deviation 0.5, and the correlation between x and y is -0.5 .
 - (b) x has mean 1 and standard deviation 1, and y is equal to x .
4. Roughly sketch the shapes of the following Gaussians $N(\mu, \Sigma)$. For each, you only need to show a representative contour line which is qualitatively accurate (has approximately the right orientation, for instance).
 - (a) $\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and $\Sigma = \begin{pmatrix} 9 & 0 \\ 0 & 1 \end{pmatrix}$
 - (b) $\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and $\Sigma = \begin{pmatrix} 1 & -0.75 \\ -0.75 & 1 \end{pmatrix}$
5. For each of the two Gaussians in the previous problem, check your answer using Python: draw 100 random samples from that Gaussian and plot it.

Linear algebra primer

DSE 210

Data as vectors and matrices



Matrix-vector notation

Vector $x \in \mathbb{R}^d$:

$$x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_d \end{pmatrix}$$

Matrix $M \in \mathbb{R}^{r \times d}$:

$$M = \begin{pmatrix} M_{11} & M_{12} & \cdots & M_{1d} \\ M_{21} & M_{22} & \cdots & M_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ M_{r1} & M_{r2} & \cdots & M_{rd} \end{pmatrix}$$

M_{ij} = entry at row i , column j

Transpose of vectors and matrices

$$x = \begin{pmatrix} 1 \\ 6 \\ 3 \\ 0 \end{pmatrix} \text{ has transpose } x^T =$$

$$M = \begin{pmatrix} 1 & 2 & 0 & 4 \\ 3 & 9 & 1 & 6 \\ 8 & 7 & 0 & 2 \end{pmatrix} \text{ has transpose } M^T =$$

- $(A^T)_{ij} = A_{ji}$
- $(A^T)^T = A$

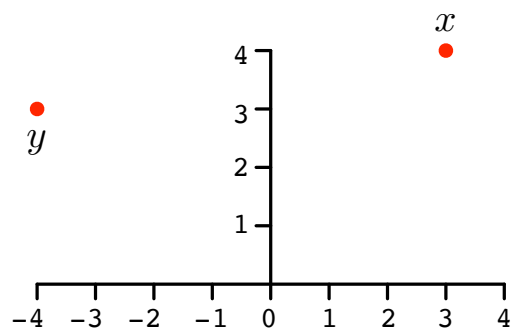
Adding and subtracting vectors and matrices

Dot product of two vectors

Dot product of vectors $x, y \in \mathbb{R}^d$:

$$x \cdot y = x_1 y_1 + x_2 y_2 + \cdots + x_d y_d.$$

What is the dot product between these two vectors?

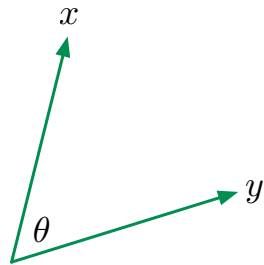


Dot products and angles

Dot product of vectors $x, y \in \mathbb{R}^d$:

$$x \cdot y = x_1 y_1 + x_2 y_2 + \cdots + x_d y_d.$$

Tells us the angle between x and y :



$$\cos \theta = \frac{x \cdot y}{\|x\| \|y\|}.$$

x is **orthogonal** (at right angles) to y if and only if $x \cdot y = 0$ When x, y are **unit vectors** (length 1): $\cos \theta = x \cdot y$ What is $x \cdot x$?

Linear and quadratic functions

In one dimension:

- Linear: $f(x) = 3x + 2$
- Quadratic: $f(x) = 4x^2 - 2x + 6$

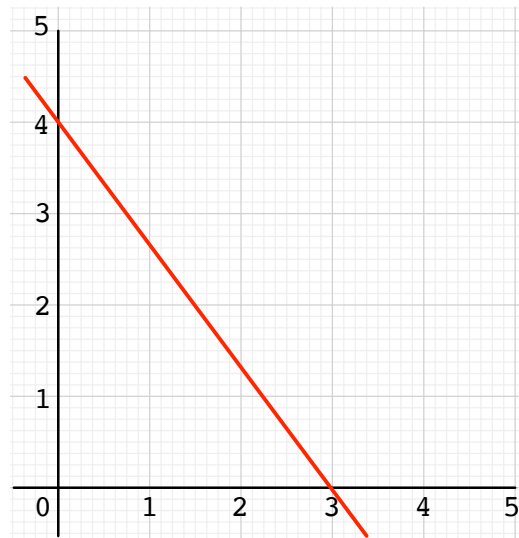
In higher dimension, e.g. $x = (x_1, x_2, x_3)$:

- Linear: $3x_1 - 2x_2 + x_3 + 4$
- Quadratic: $x_1^2 - 2x_1 x_3 + 6x_2^2 + 7x_1 + 9$

Linear functions and dot products

Linear separator

$$4x_1 + 3x_2 = 12:$$



For $x = (x_1, \dots, x_d) \in \mathbb{R}^d$, linear separators are of the form:

$$w_1x_1 + w_2x_2 + \dots + w_dx_d = c.$$

Can write as $w \cdot x = c$, for $w = (w_1, \dots, w_d)$.

More general linear functions

A linear function from \mathbb{R}^4 to \mathbb{R} : $f(x_1, x_2, x_3, x_4) = 3x_1 - 2x_3$

A linear function from \mathbb{R}^4 to \mathbb{R}^3 :

$$f(x_1, x_2, x_3, x_4) = (4x_1 - x_2, x_3, -x_1 + 6x_4)$$

Matrix-vector product

Product of matrix $M \in \mathbb{R}^{r \times d}$ and vector $x \in \mathbb{R}^d$:

The identity matrix

The $d \times d$ **identity matrix** I_d sends each $x \in \mathbb{R}^d$ to itself.

$$I_d = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

Matrix-matrix product

Product of matrix $A \in \mathbb{R}^{r \times k}$ and matrix $B \in \mathbb{R}^{k \times p}$:

Matrix products

If $A \in \mathbb{R}^{r \times k}$ and $B \in \mathbb{R}^{k \times p}$, then AB is an $r \times p$ matrix with (i, j) entry

$(AB)_{ij} = (\text{dot product of } i\text{th row of } A \text{ and } j\text{th column of } B)$

$$= \sum_{\ell=1}^k A_{i\ell} B_{\ell j}$$

- $I_k B = B$ and $A I_k = A$
- Can check: $(AB)^T = B^T A^T$
- For two vectors $u, v \in \mathbb{R}^d$, what is $u^T v$?

Some special cases

For vector $x \in \mathbb{R}^d$, what are $x^T x$ and xx^T ?

Associative but not commutative

- Multiplying matrices is **not commutative**: in general, $AB \neq BA$

$$\begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} =$$

$$\begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} =$$

- But it is **associative**: $ABCD = (AB)(CD) = (A(BC))D$, etc.

Example: if $x \in \mathbb{R}^d$ has length 2, what is $x^T x x^T x x^T x x^T x$?

A special case

Recall: For vector $x \in \mathbb{R}^d$, we have $x^T x = \|x\|^2$.

What about $x^T M x$, for arbitrary $d \times d$ matrix M ?

What is $x^T M x$ for $M = \begin{pmatrix} 1 & 2 \\ 0 & 3 \end{pmatrix}$?

Quadratic functions

Let M be any $d \times d$ (**square**) matrix.

For $x \in \mathbb{R}^d$, the mapping $x \mapsto x^T M x$ is a **quadratic function** from \mathbb{R}^d to \mathbb{R} :

$$x^T M x = \sum_{i,j=1}^d M_{ij} x_i x_j.$$

What is the quadratic function associated with $M = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 3 & 4 & 5 \end{pmatrix}$?

Write the quadratic function $f(x_1, x_2) = x_1^2 + 2x_1x_2 + 3x_2^2$ using matrices and vectors.

Special cases of square matrices

- **Symmetric:** $M = M^T$

$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 3 & 5 & 6 \end{pmatrix}, \quad \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 4 \\ 3 & 4 & 6 \end{pmatrix}$$

- **Diagonal:** $M = \text{diag}(m_1, m_2, \dots, m_d)$

$$\text{diag}(1, 4, 7) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 7 \end{pmatrix}$$

Determinant of a square matrix

Determinant of $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is $|A| = ad - bc$.

Example: $A = \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix}$

Inverse of a square matrix

The **inverse** of a $d \times d$ matrix A is a $d \times d$ matrix B for which $AB = BA = I_d$.

Notation: A^{-1} .

Example: if $A = \begin{pmatrix} 1 & 2 \\ -2 & 0 \end{pmatrix}$ then $A^{-1} = \begin{pmatrix} 0 & -1/2 \\ 1/2 & 1/4 \end{pmatrix}$. Check!

Inverse of a square matrix, cont'd

The **inverse** of a $d \times d$ matrix A is a $d \times d$ matrix B for which $AB = BA = I_d$.

Notation: A^{-1} .

- Not all square matrices have an inverse
- Square matrix A is invertible if and only if $|A| \neq 0$
- What is the inverse of $A = \text{diag}(a_1, \dots, a_d)$?

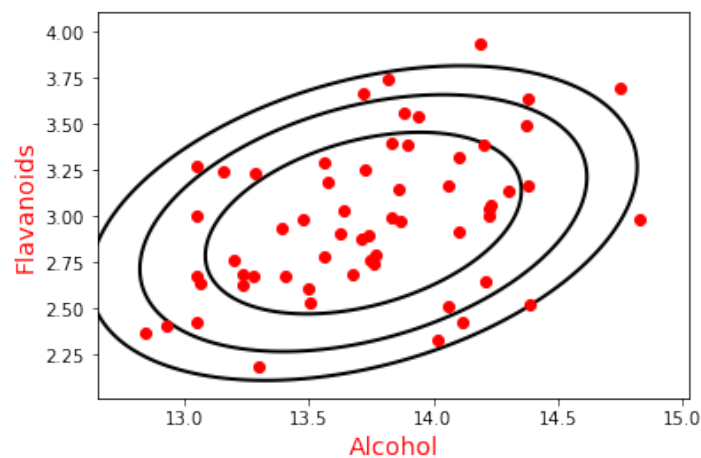
Worksheet 7 — Linear algebra primer

1. Find the unit vector in the same direction as $x = (1, 2, 3)$.
2. Find all unit vectors in \mathbb{R}^2 that are orthogonal to $(1, 1)$.
3. How would you describe the set of all points $x \in \mathbb{R}^d$ with $x \cdot x = 25$?
4. The function $f(x) = 2x_1 - x_2 + 6x_3$ can be written as $w \cdot x$ for $x \in \mathbb{R}^3$. What is w ?
5. For a certain pair of matrices A, B , the product AB has dimension 10×20 . If A has 30 columns, what are the dimensions of A and B ?
6. We have n data points $x^{(1)}, \dots, x^{(n)} \in \mathbb{R}^d$ and we store them in a matrix X , one point per row.
 - (a) What is the dimension of X ?
 - (b) What is the dimension of XX^T ?
 - (c) What is the (i, j) entry of XX^T , simply?
7. Vector x has length 10. What is $x^T x x^T x x^T x$?
8. For $x = (1, 3, 5)$ compute $x^T x$ and $x x^T$.
9. Vectors $x, y \in \mathbb{R}^d$ both have length 2. If $x^T y = 2$, what is the angle between x and y ?
10. The quadratic function $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ given by
$$f(x) = 3x_1^2 + 2x_1x_2 - 4x_1x_3 + 6x_3^2$$
can be written in the form $x^T M x$ for some **symmetric** matrix M . What is M ?
11. Which of the following matrices is necessarily symmetric?
 - (a) AA^T for arbitrary matrix A .
 - (b) $A^T A$ for arbitrary matrix A .
 - (c) $A + A^T$ for arbitrary square matrix A .
 - (d) $A - A^T$ for arbitrary square matrix A .
12. Let $A = \text{diag}(1, 2, 3, 4, 5, 6, 7, 8)$.
 - (a) What is $|A|$?
 - (b) What is A^{-1} ?
13. Vectors $u_1, \dots, u_d \in \mathbb{R}^d$ all have unit length and are orthogonal to each other. Let U be the $d \times d$ matrix whose rows are the u_i .
 - (a) What is UU^T ?
 - (b) What is U^{-1} ?
14. Matrix $A = \begin{pmatrix} 1 & 2 \\ 3 & z \end{pmatrix}$ is singular. What is z ?

Classification with generative models 3

DSE 210

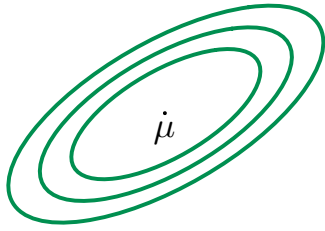
Recall: the bivariate Gaussian



Bivariate Gaussian, parametrized by:

$$\text{mean } \mu = \begin{pmatrix} 13.7 \\ 3.0 \end{pmatrix} \text{ and covariance matrix } \Sigma = \begin{pmatrix} 0.20 & 0.06 \\ 0.06 & 0.12 \end{pmatrix}$$

The multivariate Gaussian



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

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

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

- μ is the vector of coordinate-wise means:

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

- Σ is a matrix containing all pairwise covariances:

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

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

Special case: independent features

Suppose the X_i are independent, and $\text{var}(X_i) = \sigma_i^2$.

What is the covariance matrix Σ , and what is its inverse Σ^{-1} ?

Diagonal Gaussian

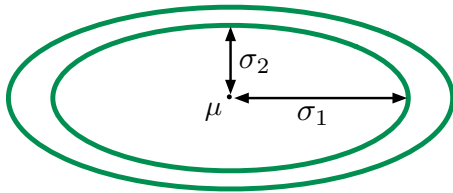
Diagonal Gaussian: the X_i are independent, with variances σ_i^2 . Thus

$$\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_d^2) \text{ (off-diagonal elements zero)}$$

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

$$\Pr(x) = \Pr(x_1)\Pr(x_2)\cdots\Pr(x_d) = \frac{1}{(2\pi)^{d/2}\sigma_1\cdots\sigma_d} \exp\left(-\sum_{i=1}^d \frac{(x_i - \mu_i)^2}{2\sigma_i^2}\right)$$

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



Even more special case: spherical Gaussian

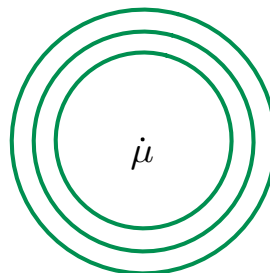
The X_i are independent and all have the same variance σ^2 .

$$\Sigma = \sigma^2 I_d = \text{diag}(\sigma^2, \sigma^2, \dots, \sigma^2) \text{ (diagonal elements } \sigma^2, \text{ rest zero)}$$

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

$$\Pr(x) = \Pr(x_1)\Pr(x_2)\cdots\Pr(x_d) = \frac{1}{(2\pi)^{d/2}\sigma^d} \exp\left(-\frac{\|x - \mu\|^2}{2\sigma^2}\right)$$

Density at a point depends only on its distance from μ :



How to fit a Gaussian to data

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

- Empirical mean

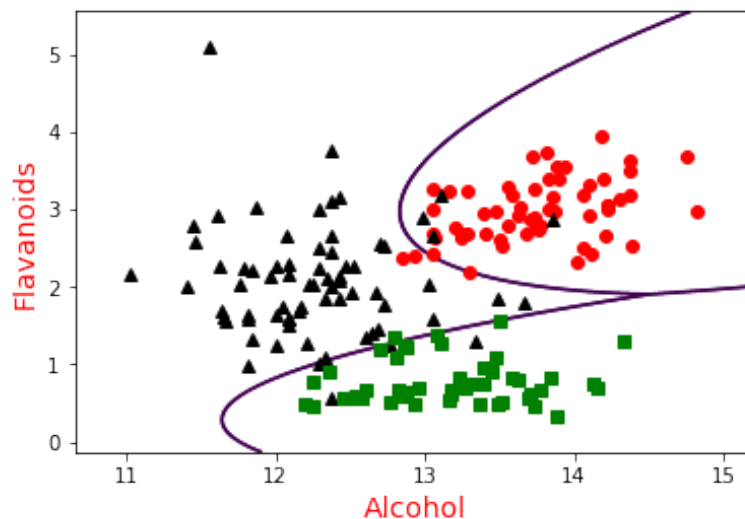
$$\mu = \frac{1}{m} (x^{(1)} + \dots + x^{(m)})$$

- Empirical covariance matrix has i, j entry:

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

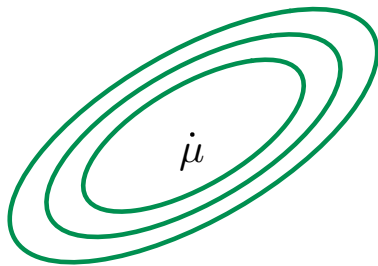
Back to the winery data

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



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

The multivariate Gaussian



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

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

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

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 .

Binary classification with Gaussian generative model

- Estimate class probabilities π_1, π_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 if

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

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 threshold depending on the various parameters.

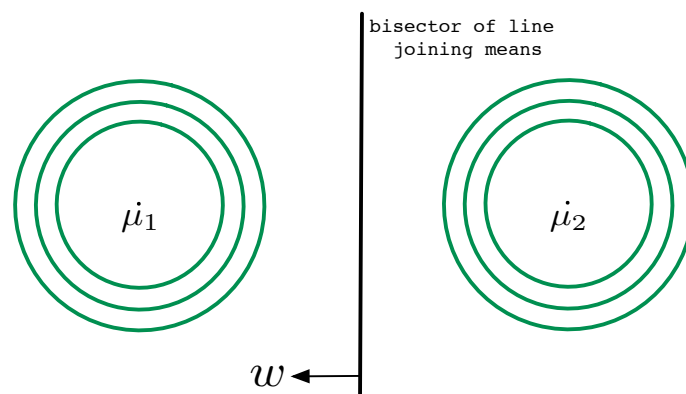
Linear or **quadratic** decision boundary.

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

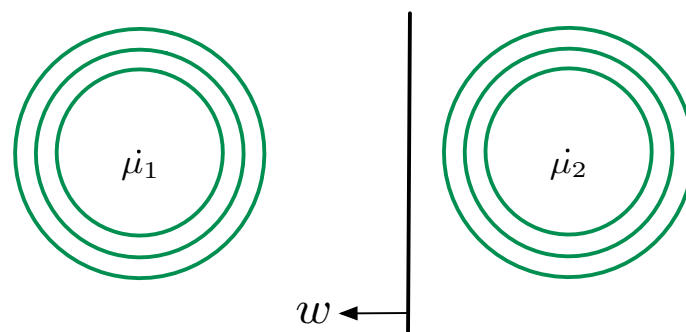
Linear decision boundary: choose class 1 if

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

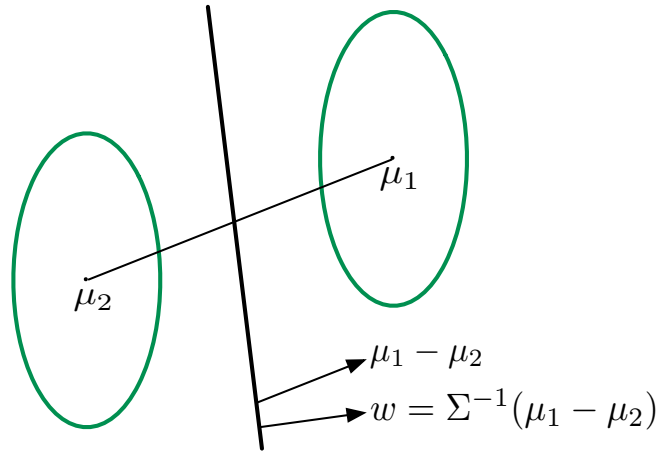
Example 1: Spherical Gaussians with $\Sigma = I_d$ and $\pi_1 = \pi_2$.



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



Example 3: Non-spherical.



Classification rule: $w \cdot x \geq \theta$

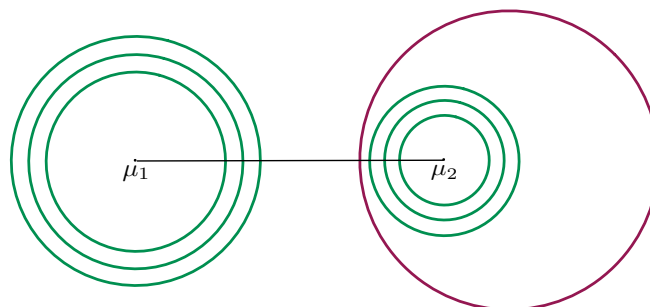
- Choose w as above
- Common practice: fit θ to minimize training or validation error

Different covariances: $\Sigma_1 \neq \Sigma_2$

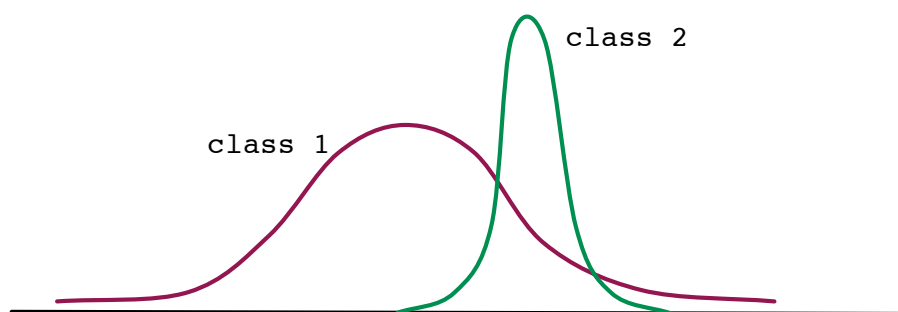
Quadratic boundary: choose class 1 if $x^T M x + 2w^T x \geq \theta$, where:

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

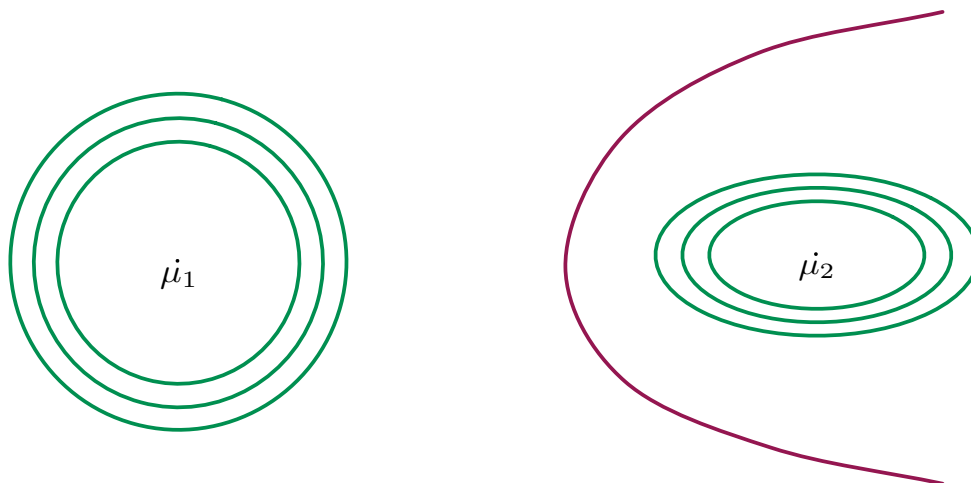
Example 1: $\Sigma_1 = \sigma_1^2 I_d$ and $\Sigma_2 = \sigma_2^2 I_d$ with $\sigma_1 > \sigma_2$



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



Example 3: A parabolic boundary.



Multiclass discriminant analysis

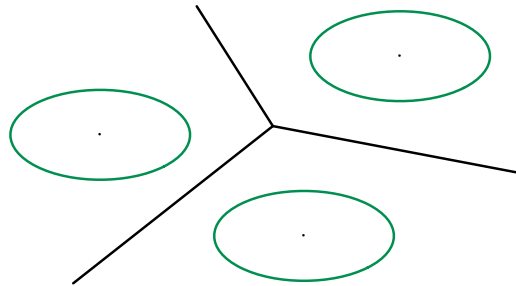
k classes: weights π_j , class-conditional densities $P_j = N(\mu_j, \Sigma_j)$.

Each class has an associated **quadratic** function

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

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

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

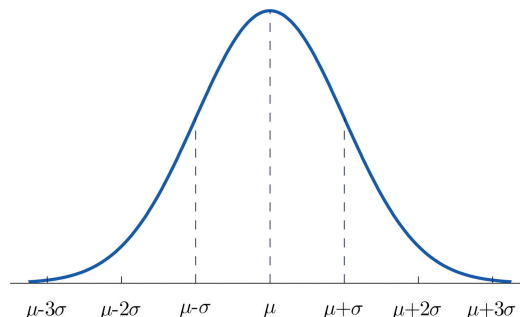


Beyond Gaussians

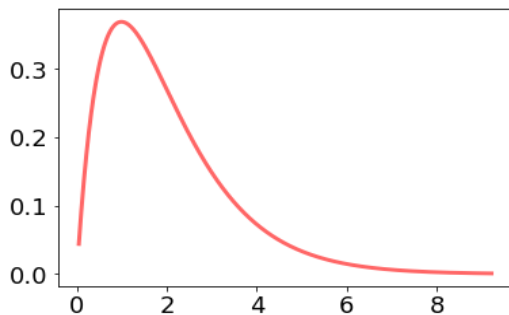
The generative methodology:

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

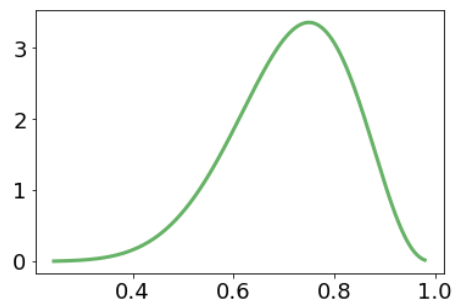
What distribution to use? Are Gaussians enough?



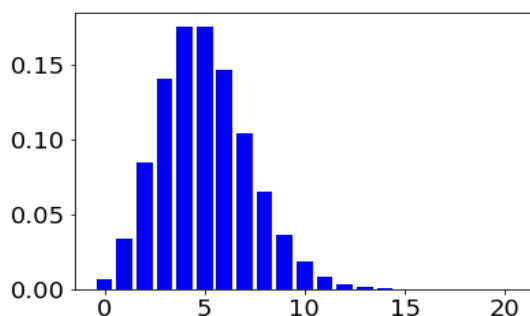
Exponential families of distributions



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

Multivariate distributions

We've described a variety of distributions for **one-dimensional** data.
What about higher dimensions?

① **Naive Bayes:** Treat coordinates as independent.

For $x = (x_1, \dots, x_d)$, fit separate models \Pr_i to each x_i , and assume

$$\Pr(x_1, \dots, x_d) = \Pr_1(x_1)\Pr_2(x_2) \cdots \Pr_d(x_d).$$

This assumption is typically inaccurate.

② **Multivariate Gaussian.**

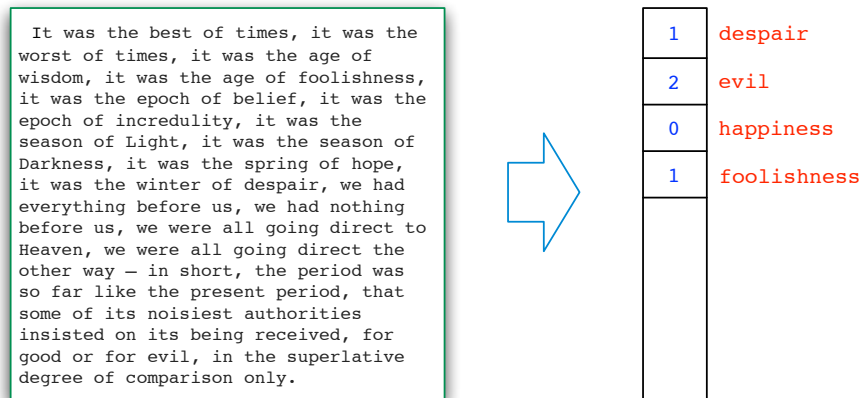
Model correlations between features: we've seen this in detail.

③ **Graphical models.**

Arbitrary dependencies between coordinates.

Handling text data

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



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

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

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 :

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

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

For naive Bayes: one multinomial distribution per class.

- Class probabilities π_1, \dots, π_k
- Multinomials $p^{(1)} = (p_{11}, \dots, p_{1|V|}), \dots, p^{(k)} = (p_{k1}, \dots, p_{k|V|})$

Classify document x as

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

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

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 on classification.

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

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

Worksheet 8 — Generative models 3

1. Consider the linear classifier $w \cdot x \geq \theta$, where

$$w = \begin{pmatrix} -3 \\ 4 \end{pmatrix} \quad \text{and} \quad \theta = 12.$$

Sketch the decision boundary in \mathbb{R}^2 . Make sure to label precisely where the boundary intersects the coordinate axes, and also indicate which side of the boundary is the positive side.

2. How many parameters are needed to specify a diagonal Gaussian in \mathbb{R}^d ?

3. *Text classification using multinomial Naive Bayes.*

- (a) For this problem, you'll be using the *20 Newsgroups* data set. There are several versions of it on the web. You should download "20news-bydate.tar.gz" from

<http://qwone.com/~jason/20Newsgroups/>

Unpack it and look through the directories at some of the files. Overall, there are roughly 19,000 documents, each from one of 20 newsgroups. The label of a document is the identity of its newsgroup. The documents are divided into a training set and a test set.

- (b) The same website has a processed version of the data, "20news-bydate-matlab.tgz", that is particularly convenient to use. Download this and also the file "vocabulary.txt". Look at the first training document in the processed set and the corresponding original text document to understand the relation between the two.
- (c) The words in the documents constitute an overall vocabulary V of size 61188. Build a multinomial Naive Bayes model using the training data. For each of the 20 classes $j = 1, 2, \dots, 20$, you must have the following:
- π_j , the fraction of documents that belong to that class; and
 - P_j , a probability distribution over V that models the documents of that class.

In order to fit P_j , imagine that all the documents of class j are strung together. For each word $w \in V$, let P_{jw} be the fraction of this concatenated document occupied by w . Well, almost: you will need to do smoothing (just add one to the count of how often w occurs).

- (d) Write a routine that uses this naive Bayes model to classify a new document. To avoid underflow, work with logs rather than multiplying together probabilities.
- (e) Evaluate the performance of your model on the test data. What error rate do you achieve?
- (f) If you have the time and inclination: see if you can get a better-performing model.
- Split the training data into a smaller training set and a validation set. The split could be 80-20, for instance. You'll use this training set to estimate parameters and the validation set to decide between different options.

- Think of 2-3 ways in which you might improve your earlier model. Examples include: (i) replacing the frequency f of a word in a document by $\log(1 + f)$, (ii) removing stopwords; (iii) reducing the size of the vocabulary; etc. Estimate a revised model for each of these, and use the validation set to choose between them.
 - Evaluate your final model on the test data. What error rate do you achieve?
4. *Handwritten digit recognition using a Gaussian generative model.* In class, we mentioned the MNIST data set of handwritten digits. You can obtain it from:

<http://yann.lecun.com/exdb/mnist/index.html>

In this problem, you will build a classifier for this data, by modeling each class as a multivariate (784-dimensional) Gaussian.

- (a) Upon downloading the data, you should have two training files (one with images, one with labels) and two test files. Unzip them.

In order to load the data into Python you will find the following code helpful:

<http://cseweb.ucsd.edu/~dasgupta/dse210/loader.py>

For instance, to load in the training data, you can use:

```
x,y = loadmnist('train-images-idx3-ubyte', 'train-labels-idx1-ubyte')
```

This will set x to a 60000×784 array where each row corresponds to an image, and y to a length-60000 array where each entry is a label (0-9). There is also a routine to display images: use `displaychar(x[0])` to show the first data point, for instance.

- (b) Split the training set into two pieces – a training set of size 50000, and a separate *validation set* of size 10000. Also load in the test data.
- (c) Now fit a Gaussian generative model to the training data of 50000 points:
- Determine the class probabilities: what fraction π_0 of the training points are digit 0, for instance? Call these values π_0, \dots, π_9 .
 - Fit a Gaussian to each digit, by finding the mean and the covariance of the corresponding data points. Let the Gaussian for the j th digit be $P_j = N(\mu_j, \Sigma_j)$.

Using these two pieces of information, you can classify new images x using Bayes' rule: simply pick the digit j for which $\pi_j P_j(x)$ is largest.

- (d) One last step is needed: it is important to smooth the covariance matrices, and the usual way to do this is to add in cI , where c is some constant and I is the identity matrix. What value of c is right? Use the validation set to help you choose. That is, choose the value of c for which the resulting classifier makes the fewest mistakes on the validation set. What value of c did you get?
- (e) Turn in an iPython notebook that includes:
- All your code.
 - Error rate on the MNIST test set.
 - Out of the misclassified test digits, pick five at random and display them. For each instance, list the posterior probabilities $\Pr(y|x)$ of each of the ten classes.