The Poisson distribution

DSE 210

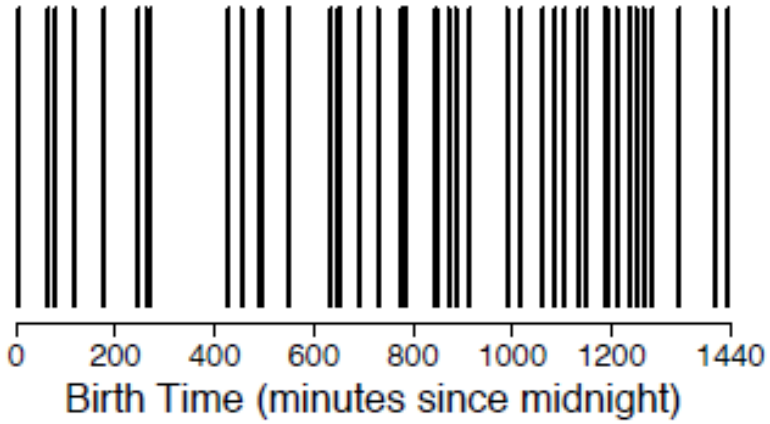
Examples

Many experimental situation occur in which we observe the counts of events within a set unit of time, area, volume, length etc. For example,

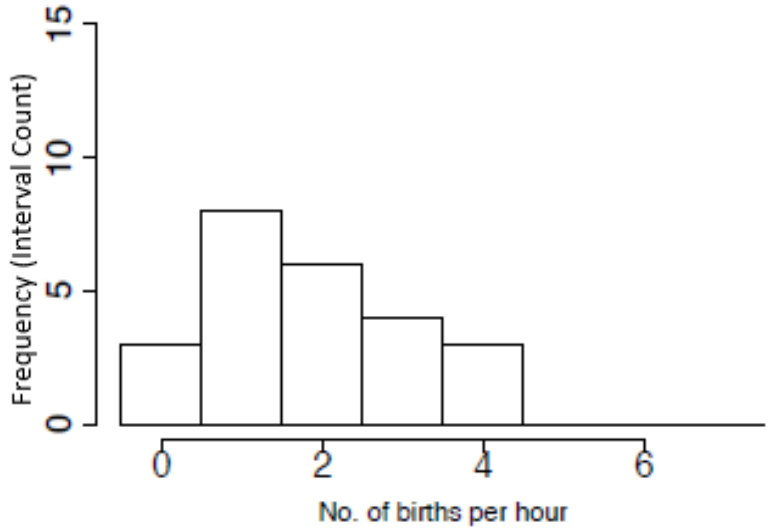
- The number of cases of a disease in different towns;
- Flying bomb hits on London in WWII
- The number of particles emitted by a radioactive source in a given time;
- The number of births per hour during a given day.

Example: The number of births in a day

The births in a given day happened like this:



and the histogram of these birth times per hour.



The Poisson Distribution

The Poisson distribution is a discrete probability distribution for the counts of events that occur randomly in a given interval of time (or space).

If we let $X =$ The number of events in a given interval.

Then, if the mean number of events per interval is λ

The probability of observing x events in a given interval is given by

$$\mathrm{P(X = x)} = \mathrm{e}^{-\lambda}\frac{\lambda^x}{x!} \qquad x = 0, 1, 2, 3, 4, \ldots$$

The Poisson Distribution

Note: e is a mathematical constant. $\mathrm{e} \approx 2.718282$. There should be a button on your calculator $\boxed{\mathrm{e}^x}$ that calculates powers of e.

If the probabilities of X are distributed in this way, we write

$$X \sim \text{Poisson}(\lambda).$$

λ is the **parameter** of the distribution. We *say* X follows a Poisson distribution with parameter λ

Note: A Poisson random variable can take on any positive integer value. In contrast, the Binomial distribution always has a finite upper limit.

Example: Hospital Births

Births in a hospital occur randomly with Poisson distribution at an average rate of 1.8 births per hour.

What is the probability of observing 4 births in a given hour at the hospital?

Let X = No. of births in a given hour

$\text{Mean rate } \lambda = 1.8$ Then, $X \sim$ Poisson(1.8)

We can now use the formula to calculate the probability of observing exactly 4 births in a given hour

$$P(X = 4) = \mathrm{e}^{-1.8}\frac{1.8^4}{4!} = 0.0723$$

Example: Hospital Births

What about the probability of observing more than or equal to 2 births in a given hour at the hospital?

We want $P(X \geq 2) = P(X = 2) + P(X = 3) + \ldots$

i.e. an infinite number of probabilities to calculate

but

$$\begin{aligned}
P(X \geq 2) &= P(X = 2) + P(X = 3) + \ldots \\
&= 1 - P(X < 2) \\
&= 1 - (P(X = 0) + P(X = 1)) \\
&= 1 - \left(\mathrm{e}^{-1.8} \frac{1.8^0}{0!} + \mathrm{e}^{-1.8} \frac{1.8^1}{1!} \right) \\
&= 1 - (0.16529 + 0.29753) \\
&= 0.537
\end{aligned}$$

Example: Disease Incidence

Suppose there is a disease, whose average incidence is 2 per million people. What is the probability that a city of 1 million people has at least twice the average incidence?

Twice the average incidence would be 4 cases.

We can reasonably suppose the random variable

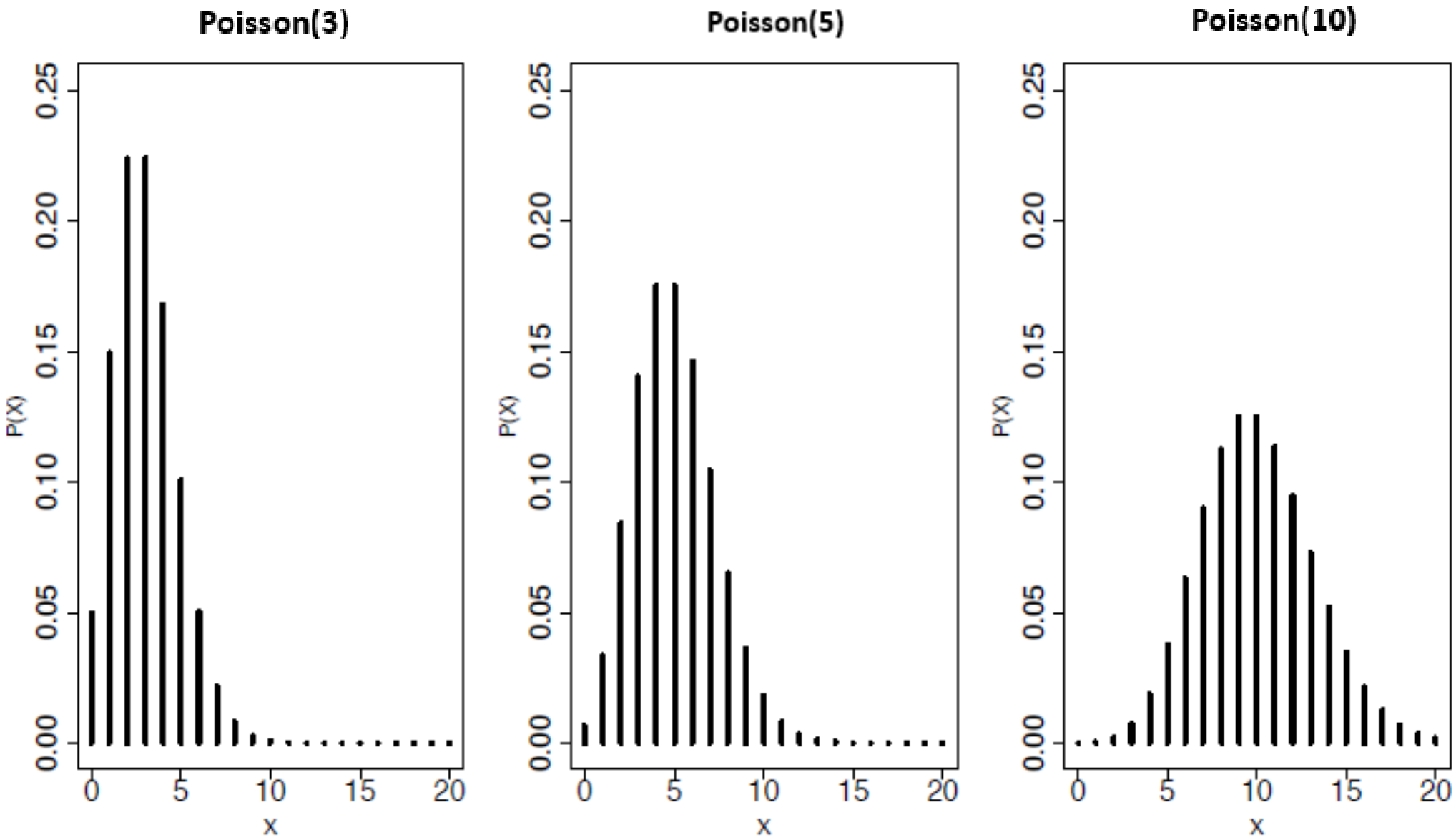
X=number of cases in 1 million people

has Poisson distribution with parameter 2.

Then

$$P(X \geq 4) = 1 - P(X \leq 3) = 1 - \left(e^{-2}\frac{2^0}{0!} + e^{-2}\frac{2^1}{1!} + e^{-2}\frac{2^2}{2!} + e^{-3}\frac{2^3}{3!} \right)$$
$$= 0.143.$$

The Shape of the Poisson Distribution



The Shape of the Poisson Distribution

We observe that the Poisson distributions

1 - are unimodal (have one clear peak),

2 - exhibit positive skew (that decreases as λ increases),

3 - are centered roughly on λ,

4 - have variance (spread) that increases as λ increases.

Mean and Variance

- The mean and the variance of a Poisson random variable are as follows:

If $X \sim$ Poisson(λ)

- Mean: $\mathbb{E}X = \lambda$
- Variance: $\mathbb{E}(X - \lambda)^2 = \lambda$

Changing the Size of The Interval

Suppose we know that births in a hospital occur randomly at an average rate of 1.8 births per hour.

What is the probability that we observe 5 births in a given 2 hour interval?

Well, if births occur randomly at a rate of 1.8 births per 1 hour interval
Then births occur randomly at a rate of 3.6 births per 2 hour interval

Let Y = No. of births in a 2 hour period

Then $Y \sim$ Poisson(3.6)

$$P(Y = 5) = \mathrm{e}^{-3.6}\frac{3.6^5}{5!} = 0.13768$$

Changing the Size of The Interval

The previous example illustrates the following rule:

If X ~ Poisson(λ) on 1-unit interval,

then Y ~ Poisson($k\lambda$) on k-unit intervals.

Sum of Two Poisson Variables

Now suppose we know that

- in hospital A births occur randomly at an average rate of 2.3 births per hour
- in hospital B births occur randomly at an average rate of 3.1 births per hour

What is the probability that we observe 7 births in total from the two hospitals in a given 1 hour period?

To answer this question we can use the following rule:

If $X \sim \text{Poisson}(\lambda_1)$ on 1-unit interval,

and $Y \sim \text{Poisson}(\lambda_2)$ on 1-unit interval,

then $X + Y \sim \text{Poisson}(\lambda_1 + \lambda_2)$ on 1-unit interval.

Sum of two Poisson Variables

So if we let

$$X = \text{No. of births in a given hour at hospital A}$$

and

$$Y = \text{No. of births in a given hour at hospital B}$$

Then X ~ Poisson(2.3), Y ~ Poisson(3.1), X + Y ~ Poisson(5.4)

$$\Rightarrow P(X+Y=7) = \mathrm{e}^{-5.4}\frac{5.4^7}{7!} = 0.11999$$

Example: Disease Incidence

Suppose

- disease A occurs with incidence 1.7 per million,
- disease B occurs with incidence 2.9 per million.

Statistics are compiled, in which these diseases are not distinguished, but simply are all called cases of disease "AB".

What is the probability that a city of 1 million people has at least 6 cases of AB?

If Z=Number of cases of AB, then $P \sim$ Poisson(4.6)Thus,

$$\begin{aligned} P(Z \geq 6) &= 1 - P(Z \leq 5) \\ &= 1 - e^{-4.6}\left(\frac{4.6^0}{0!} + \frac{4.6^1}{1!} + \frac{4.6^2}{2!} + \frac{4.6^3}{3!} + \frac{4.6^4}{4!} + \frac{4.6^5}{5!}\right) \\ &= 0.314. \end{aligned}$$

Fitting a Poisson Distribution

- Consider the birth times example we saw at the beginning. The example had a total of 44 births in 24-hour interval.

- Therefore the mean birth rate is 44/24 = 1.8333

- What would be the expected counts if birth times were Poisson distributed i.e. what is the expected histogram for a Poisson random variable with mean rate $\lambda = 1.8333$?

- Using the Poisson formula we can calculate the probabilities of obtaining each possible value.

Fitting a Poisson Distribution

In practice we group values with low probability into one category.

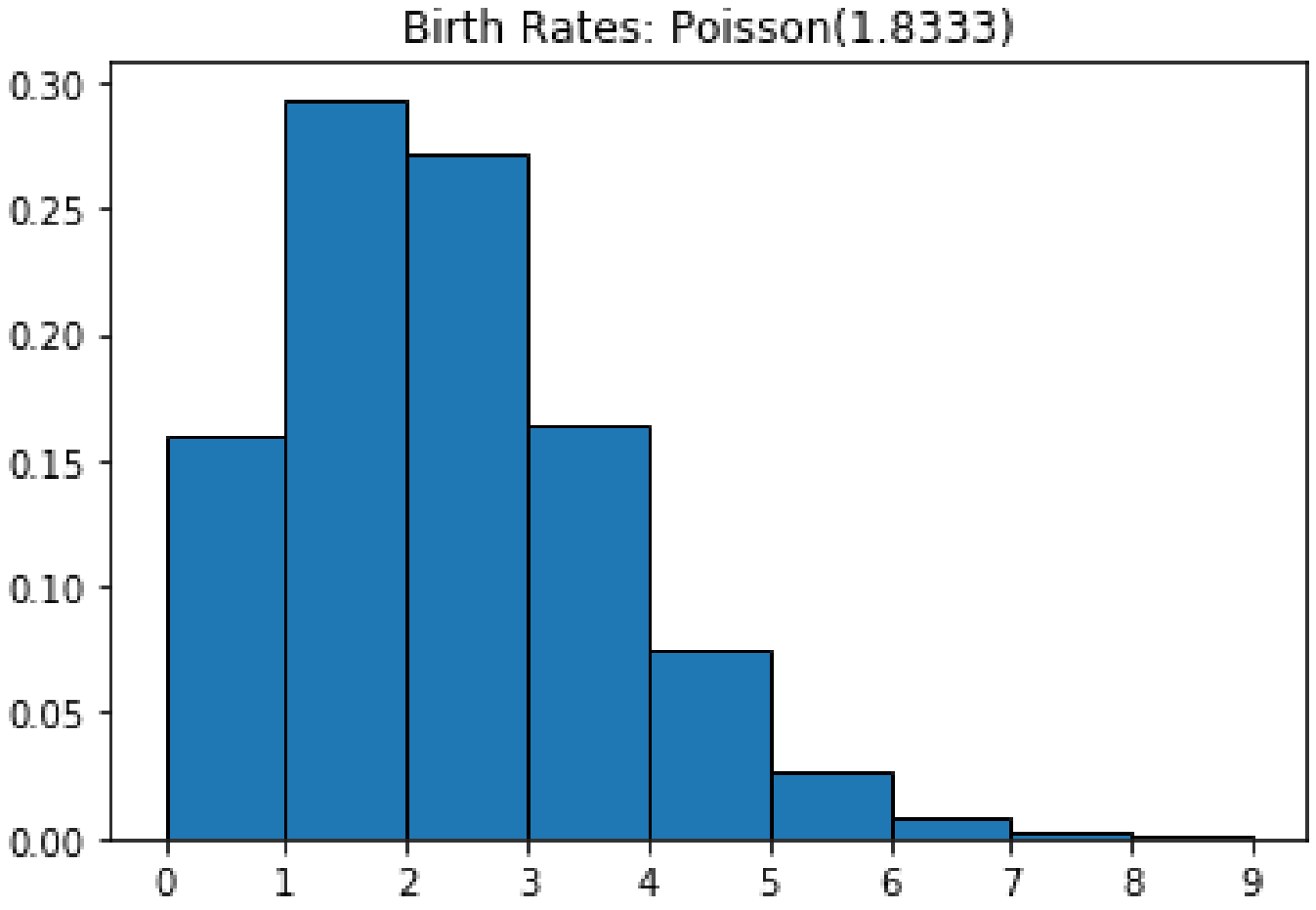
x	0	1	2	3	4	5	≥ 6
$P(X = x)$	0.159	0.293	0.268	0.164	0.075	0.027	0.011

Then if we observe 24 hour intervals we can calculate the expected frequencies as $24 \times P(X = x)$ for each value of x.

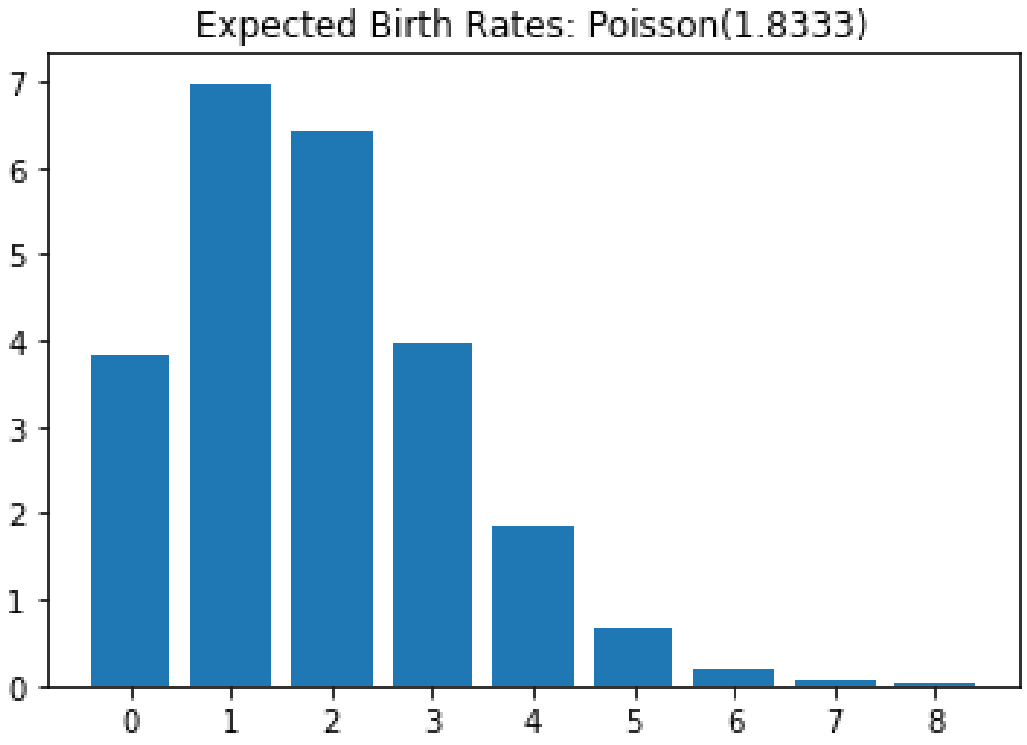
x	0	1	2	3	4	5	≥ 6
Expected freq.	3.837	7.035	6.448	3.941	1.806	0.662	0.271

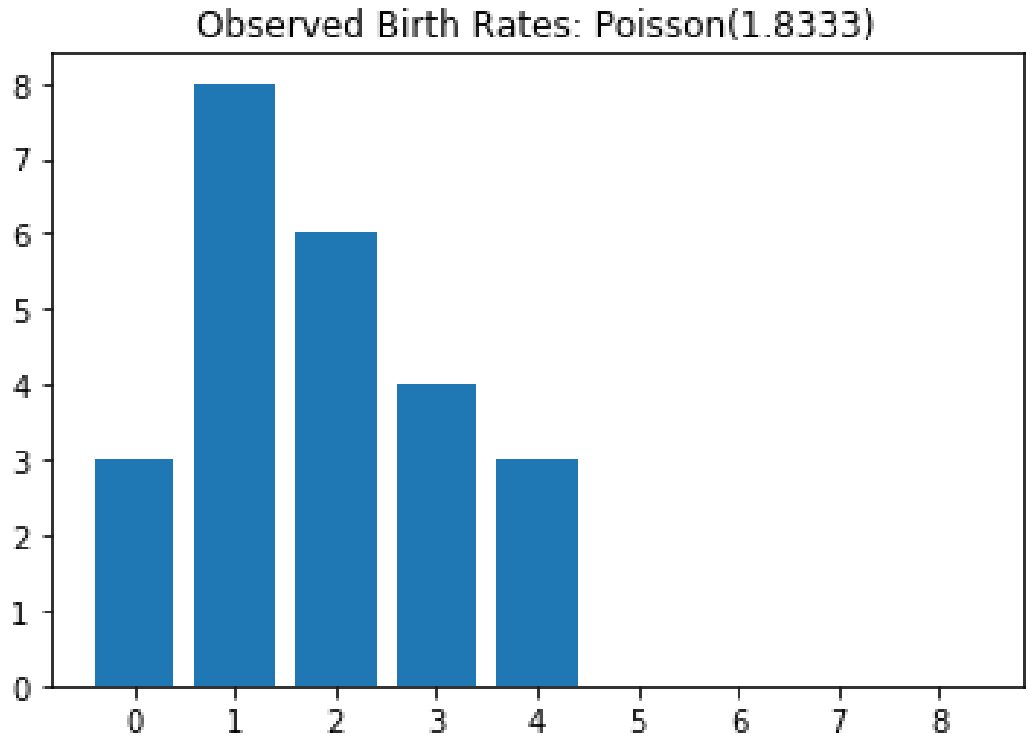
We say we have fitted a Poisson distribution to the data.

Fitting a Poisson Distribution: Probability Density



Fitting a Poisson Distribution: Histogram of the Fit





Fitting a Poisson Distribution

This consists of 3 steps

1. Estimating the parameters of the distribution from the data
2. Calculating the probability distribution
3. Multiplying the probability distribution by the number of observations

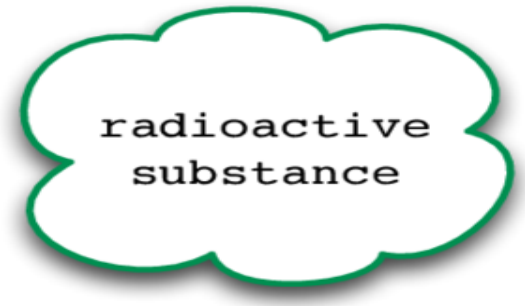
Fitting a Poisson Distribution

- Once we have fitted a distribution to the data, we can compare the expected frequencies to those we observed from the real birth dataset. We see that the agreement is quite good.

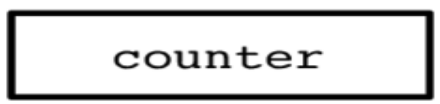
x	0	1	2	3	4	5	≥ 6
Expected	3.837	7.035	6.448	3.941	1.806	0.662	0.271
Observed	3	8	6	4	3	0	0

Poisson: examples

Rutherford's experiments with radioactive disintegration (1920)

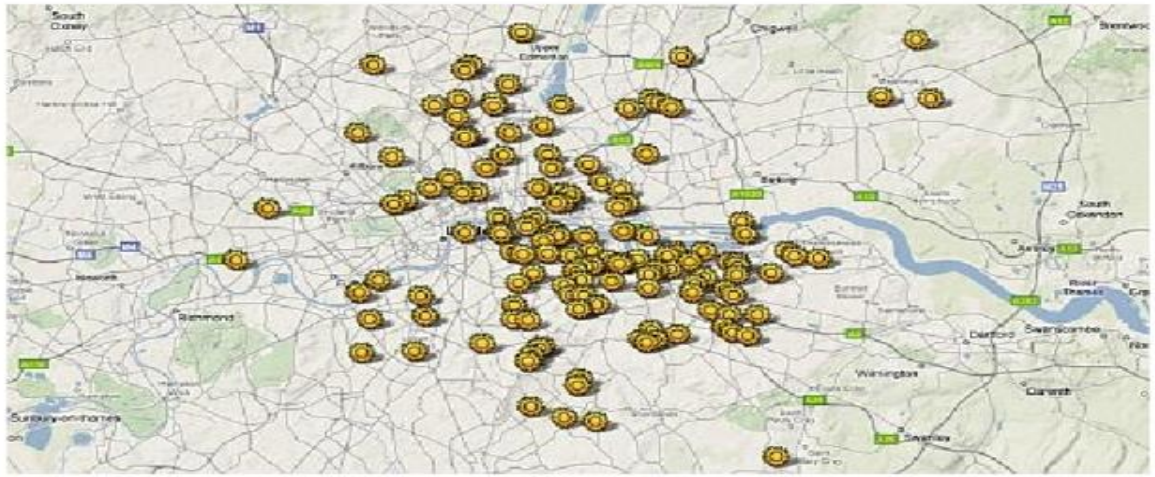


- $N = 2608$ intervals of 7.5 seconds
- $N_k = \#$ intervals with k particles
- Mean: 3.87 particles per interval



k	0	1	2	3	4	5	6	7	8	≥ 9
N_k	57	203	383	525	532	408	273	139	45	43
P(3.87)	54.4	211	407	526	508	394	254	140	67.9	46.3

Flying bomb hits on London in WWII



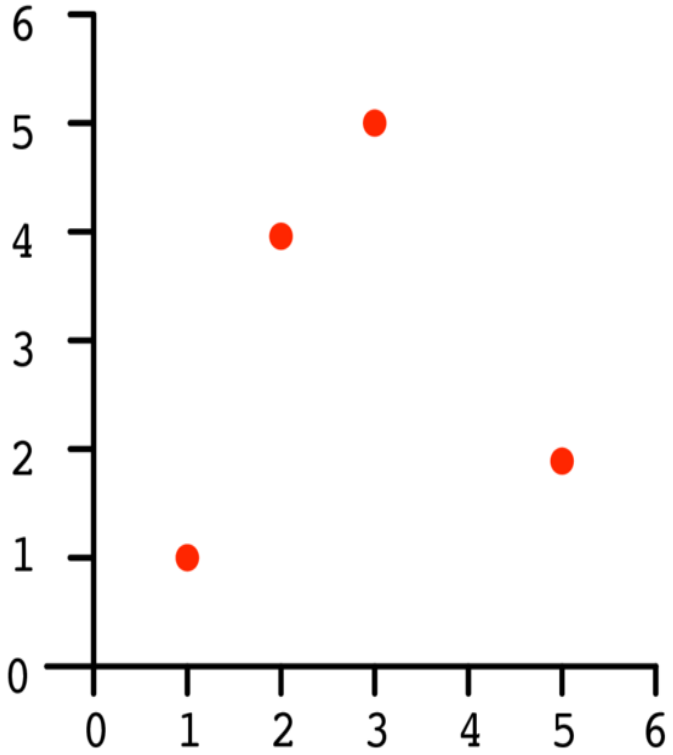
- Area divided into 576 regions, each 0.25 km^2
- N_k = # regions with k hits
- Mean: 0.93 hits per region

k	0	1	2	3	4	≥ 5
N_k	229	211	93	35	7	1
P(0.93)	226.8	211.4	98.54	30.62	7.14	1.57

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

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

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

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

Adding and subtracting vectors and matrices

Example 1:

Add the matrices.

$$\begin{bmatrix} 1 & 5 \\ -4 & 3 \end{bmatrix} + \begin{bmatrix} 2 & -1 \\ 4 & -1 \end{bmatrix}$$

First note that both addends are 2×2 matrices, so we can add them.

$$\begin{bmatrix} 1 & 5 \\ -4 & 3 \end{bmatrix} + \begin{bmatrix} 2 & -1 \\ 4 & -1 \end{bmatrix} = \begin{bmatrix} 1+2 & 5+(-1) \\ -4+4 & 3+(-1) \end{bmatrix}$$

$$= \begin{bmatrix} 3 & 4 \\ 0 & 2 \end{bmatrix}$$

Adding and subtracting vectors and matrices

Example 2:

Subtract.

$$\begin{bmatrix} 4 & 5 & 6 \\ 2 & 3 & 4 \end{bmatrix} - \begin{bmatrix} 2 & 4 & 6 \\ 1 & 2 & 3 \end{bmatrix}$$

Subtract corresponding entries.

$$\begin{bmatrix} 4 & 5 & 6 \\ 2 & 3 & 4 \end{bmatrix} - \begin{bmatrix} 2 & 4 & 6 \\ 1 & 2 & 3 \end{bmatrix} = \begin{bmatrix} 4-2 & 5-4 & 6-6 \\ 2-1 & 3-2 & 4-3 \end{bmatrix}$$

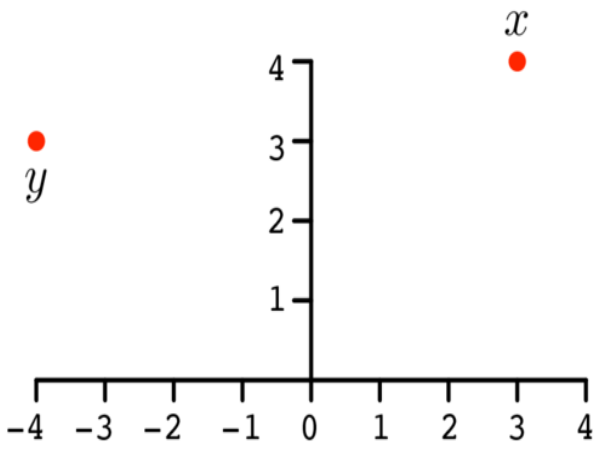
$$= \begin{bmatrix} 2 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

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?



$$x = \begin{pmatrix} 3 \\ 4 \end{pmatrix}$$

$$y = \begin{pmatrix} -4 \\ 3 \end{pmatrix}$$

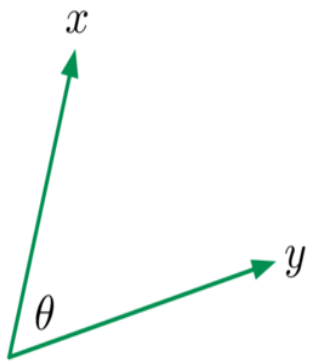
$$\mathbf{x} \cdot \mathbf{y} = 3(-4) + (4)3 = 0$$

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

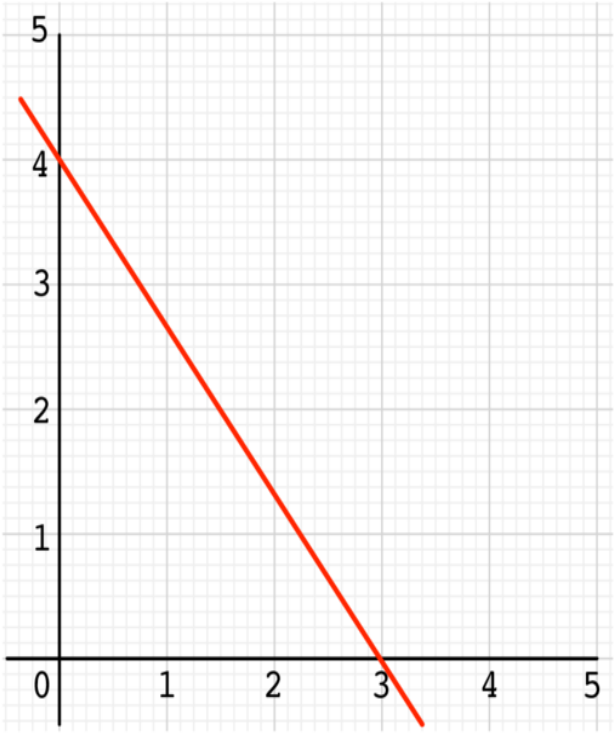
In higher dimension, e.g. $x = (x_1, x_2, x_3)$:

- Linear: $3x_1 - 2x_2 + x_3 + 4$
- Quadratic: $x_1^2 - 2x_1x_3 + 6x_2^2 + 7x_1 + 9$

Linear functions and dot products

Linear separator

$4x_1 + 3x_2 = 12$:



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

$$w_1 x_1 + w_2 x_2 + \cdots + w_d x_d = c.$$

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

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

$$Ax = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1d} \\ a_{21} & a_{22} & \dots & a_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ a_{r1} & a_{r2} & \dots & a_{rd} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1d}x_d \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2d}x_d \\ \vdots \\ a_{r1}x_1 + a_{r2}x_2 + \cdots + a_{rd}x_d \end{bmatrix}$$

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

If **A** is an $m \times n$ matrix and **B** is an $n \times p$ matrix,

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1p} \\ b_{21} & b_{22} & \cdots & b_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{np} \end{pmatrix}$$

the *matrix product* **C** = **AB** is defined to be the $m \times p$ matrix:

$$\mathbf{C} = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1p} \\ c_{21} & c_{22} & \cdots & c_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ c_{m1} & c_{m2} & \cdots & c_{mp} \end{pmatrix}$$

such that

$$c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{in}b_{nj} = \sum_{k=1}^{n} a_{ik}b_{kj},$$

for $i = 1, \ldots, m$ and $j = 1, \ldots, p$.

Example:

Let A be a 2x3 matrix, and B be a 3x2 matrix:

$$A = \begin{bmatrix} 0 & 4 & -2 \\ -4 & -3 & 0 \end{bmatrix} \qquad B = \begin{bmatrix} 0 & 1 \\ 1 & -1 \\ 2 & 3 \end{bmatrix}$$

Then AB is:

$$\begin{aligned} AB &= \begin{bmatrix} 0 & 4 & -2 \\ -4 & -3 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & -1 \\ 2 & 3 \end{bmatrix} \\ &= \begin{bmatrix} 0\cdot 0 + 4\cdot 1 - 2\cdot 2 & 0\cdot 1 + 4\cdot(-1) - 2\cdot 3 \\ -4\cdot 0 - 3\cdot 1 + 0\cdot 2 & -4\cdot 1 - 3\cdot(-1) + 0\cdot 3 \end{bmatrix} \\ &= \begin{bmatrix} 0+4-4 & 0-4-6 \\ 0-3+0 & -4+3+0 \end{bmatrix} \\ &= \begin{bmatrix} 0 & -10 \\ -3 & -1 \end{bmatrix}. \end{aligned}$$

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?

$$x^T x = \begin{bmatrix} x_1 & x_2 & \cdots & x_d \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} = \begin{bmatrix} x_1^2 + x_2^2 + \cdots + x_d^2 \end{bmatrix} = \|x\|^2$$

$$xx^T = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} \begin{bmatrix} x_1 & x_2 & \cdots & x_d \end{bmatrix} = \begin{bmatrix} x_1^2 & x_1 x_2 & \cdots & x_1 x_d \\ x_2 x_1 & x_2^2 & \cdots & x_2 x_d \\ \vdots & \vdots & \ddots & \vdots \\ x_d x_1 & x_d x_2 & \cdots & x_n^2 \end{bmatrix}$$

Associative but not commutative

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

$$\begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 3 & 0 \\ 1 & 0 \end{pmatrix}$$

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

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

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

$$\mathbf{x}^T\mathbf{x} = 4 \text{ Then, } 4^4$$

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?

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

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

$$\begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 0 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

$$= \begin{pmatrix} x_1 & 2x_1 + 3x_2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = x_1^2 + 2x_1x_2 + 3x_2^2$$

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 $M = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 3 & 4 & 5 \end{pmatrix}$?

$$[x_1 \;\; x_2 \;\; x_3] \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 3 & 4 & 5 \end{pmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

$$= [x_1 + 3x_3 \quad 2x_2 + 4x_3 \quad 5x_3] \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = x_1^2 + 3x_1x_3 + 2x_2^2 + 4x_2x_3 + 5x_3^2$$

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

$$\begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = x_1^2 + 2x_1x_2 + 3x_2^2$$

$$ax_1^2 + bx_1x_2 + cx_1x_2 + dx_2^2 = x_1^2 + 2x_1x_2 + 3x_2^2$$

$$(a = 1, d = 3, b + c = 2)$$

$$\begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} 1 & b \\ 2-b & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

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**: $M = \mathrm{diag}(m_1, m_2, \ldots, m_d)$

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

Determinant of a square matrix

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

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

(3)(2) – (1)(1) = 5

Determinant of a 3x3 matrix

$$\begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = a\begin{vmatrix} e & f \\ h & i \end{vmatrix} - b\begin{vmatrix} d & f \\ g & i \end{vmatrix} + c\begin{vmatrix} d & e \\ g & h \end{vmatrix}$$

this can be expanded out to give

$$\begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = a(ei - fh) - b(di - fg) + c(dh - eg)$$

$$= aei + bfg + cdh - ceg - bdi - afh.$$

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 2x2 matrix:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad-bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

determinant

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 $A = \text{diag}(a_1, \ldots, a_d)$?

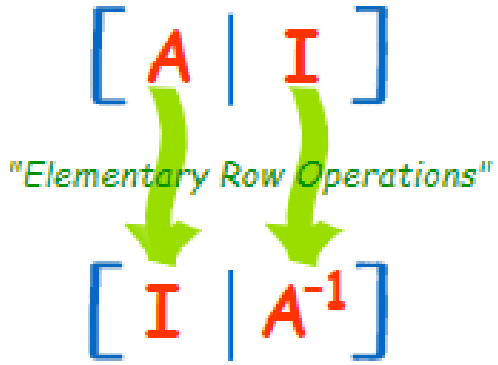
  = diag($1/a_1$, …, $1/a_d$)

$$A = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 4 \end{bmatrix}. \qquad A^{-1} = \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{4} \end{bmatrix}.$$

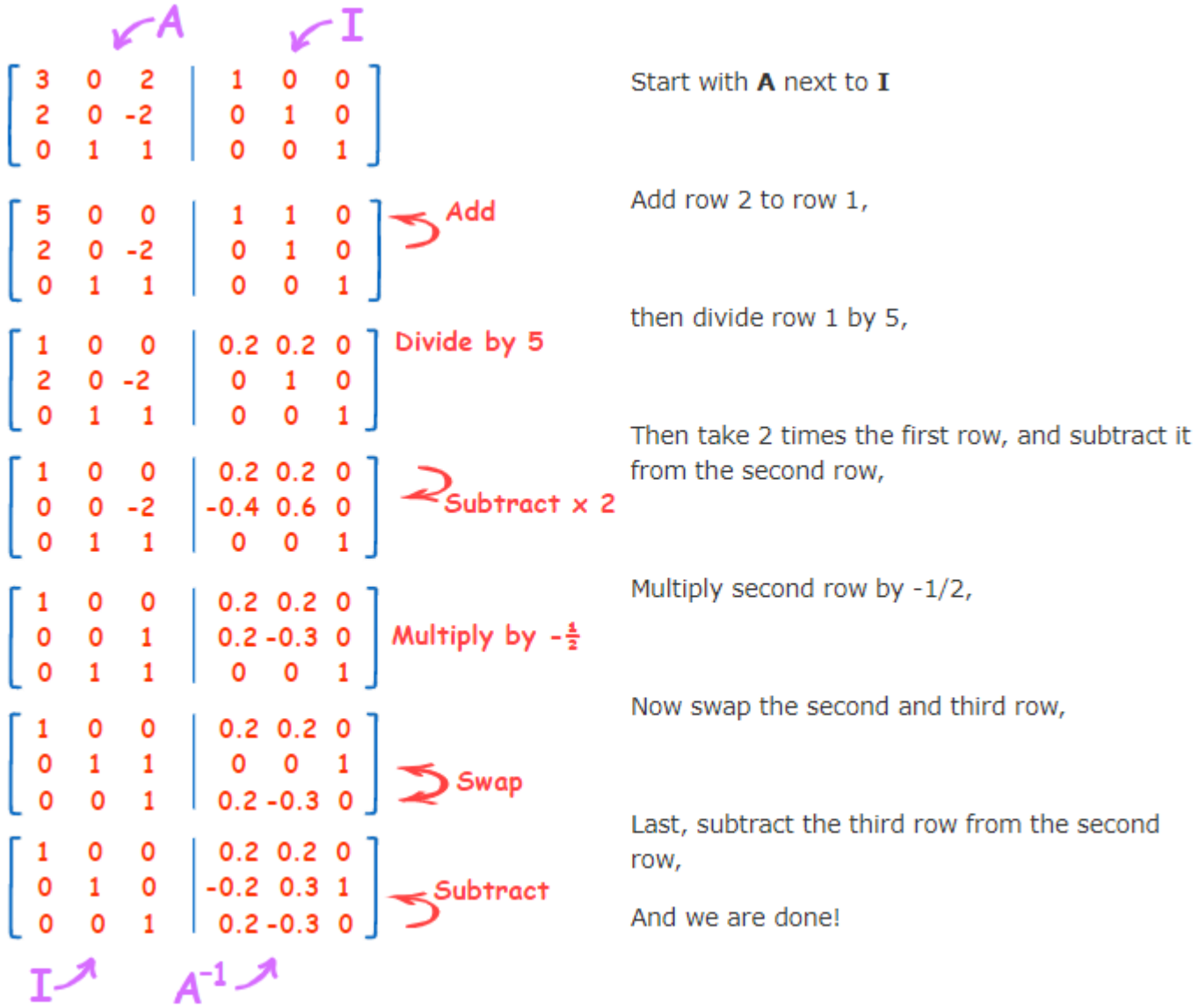
Appendix: Inverse of larger matrices

Gauss-Jordan method

- Play around with the rows (adding, multiplying or swapping) until we make Matrix **A** into the Identity Matrix



Appendix: Example



Start with **A** next to **I**

Add row 2 to row 1,

then divide row 1 by 5,

Then take 2 times the first row, and subtract it from the second row,

Multiply second row by -1/2,

Now swap the second and third row,

Last, subtract the third row from the second row,

And we are done!

And matrix **A** has been made into an Identity Matrix ...

Appendix: Determinant of a n x n matrix

The Leibniz formula for the determinant of an $n \times n$ matrix A is

$$\det(A) = \sum_{\sigma \in S_n} \left(\operatorname{sgn}(\sigma) \prod_{i=1}^{n} a_{i,\sigma_i} \right).$$

- Here the sum is computed over all permutations σ of the set {1, 2, …, n}. A permutation is a function that reorders this set of integers. The value in the i^{th} position after the reordering σ is denoted by σ_i.
- For example, for n = 3, the original sequence 1, 2, 3 might be reordered to σ = [2, 3, 1], with $\sigma_1 = 2$, $\sigma_2 = 3$, and $\sigma_3 = 1$. The set of all such permutations (also known as the symmetric group on n elements) is denoted by S_n.
- For each permutation σ, sgn(σ) denotes the signature of σ, a value that is +1 whenever the reordering given by σ can be achieved by successively interchanging two entries an even number of times, and −1 whenever it can be achieved by an odd number of such interchanges.

In any of the $n!$ summands, the term

$$\prod_{i=1}^{n} a_{i,\sigma_i}$$

is notation for the product of the entries at positions (i, σ_i), where i ranges from 1 to n:

$$a_{1,\sigma_1} \cdot a_{2,\sigma_2} \cdots a_{n,\sigma_n}.$$

Appendix: Determinant of a n x n matrix

For example, the determinant of a 3 × 3 matrix A (n = 3) is

$$\sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} a_{i,\sigma_i}$$

$$= \operatorname{sgn}([1,2,3]) \prod_{i=1}^{n} a_{i,[1,2,3]_i} + \operatorname{sgn}([1,3,2]) \prod_{i=1}^{n} a_{i,[1,3,2]_i} + \operatorname{sgn}([2,1,3]) \prod_{i=1}^{n} a_{i,[2,1,3]_i} +$$

$$\operatorname{sgn}([2,3,1]) \prod_{i=1}^{n} a_{i,[2,3,1]_i} + \operatorname{sgn}([3,1,2]) \prod_{i=1}^{n} a_{i,[3,1,2]_i} + \operatorname{sgn}([3,2,1]) \prod_{i=1}^{n} a_{i,[3,2,1]_i}$$

$$= \prod_{i=1}^{n} a_{i,[1,2,3]_i} - \prod_{i=1}^{n} a_{i,[1,3,2]_i} - \prod_{i=1}^{n} a_{i,[2,1,3]_i} + \prod_{i=1}^{n} a_{i,[2,3,1]_i} + \prod_{i=1}^{n} a_{i,[3,1,2]_i} - \prod_{i=1}^{n} a_{i,[3,2,1]_i}$$

$$= a_{1,1}a_{2,2}a_{3,3} - a_{1,1}a_{2,3}a_{3,2} - a_{1,2}a_{2,1}a_{3,3} + a_{1,2}a_{2,3}a_{3,1} + a_{1,3}a_{2,1}a_{3,2} - a_{1,3}a_{2,2}a_{3,1}.$$

Generative Models 1

DSE 210

Machine learning versus Algorithms

In both fields, the goal is to develop

procedures that exhibit a desired input-output behavior.

- **Algorithms**: the input-output mapping can be precisely defined.
  Input: Graph G.
  Output: MST of G.
- **Machine learning**: the mapping cannot easily be made precise.
  Input: Picture of an animal.
  Output: Name of the animal.

Instead, we simply provide examples of (input,output) pairs and ask the machine to *learn* a suitable mapping itself.

Inputs and outputs

Basic terminology:

- The input space, $\mathcal{X}$.
  E.g. 32×32 RGB images of animals.
- The output space, $\mathcal{Y}$.
  E.g. Names of 100 animals.

x:

y: "bear"

After seeing a bunch of examples (x, y), pick a mapping

$$f : \mathcal{X} \to \mathcal{Y}$$

that accurately replicates the input-output pattern of the examples.

Learning problems are often categorized according to the type of *output space*: (1) discrete, (2) continuous, (3) probability values, or (4) more general structures.

Discrete output space: classification

Binary classification:

- Spam detection
  $\mathcal{X} = \{\text{email messages}\}$
  $\mathcal{Y} = \{\text{spam}, \text{not spam}\}$
- Credit card fraud detection
  $\mathcal{X} = \{\text{descriptions of credit card transactions}\}$
  $\mathcal{Y} = \{\text{fraudulent}, \text{legitimate}\}$

Multiclass classification:

- Animal recognition
  $\mathcal{X} = \{\text{animal pictures}\}$
  $\mathcal{Y} = \{\text{dog}, \text{cat}, \text{giraffe}, \ldots\}$
- News article classification
  $\mathcal{X} = \{\text{news articles}\}$
  $\mathcal{Y} = \{\text{politics}, \text{business}, \text{sports}, \ldots\}$

Continuous output space: regression

- A parent's concerns
  How cold will it be tomorrow morning?
  $\mathcal{Y} = [-273, \infty)$
- For the asthmatic
  Predict tomorrow's air quality (max over the whole day)
  $\mathcal{Y} = [0, \infty)$ $(< 100$: okay, > 200: dangerous)
- Insurance company calculations
  In how many years will this person die?
  $\mathcal{Y} = [0, 200]$

What are suitable predictor variables ($\mathcal{X}$) in each case?

Conditional probability functions

Here $\mathcal{Y} = [0, 1]$ represents probabilities.

- Dating service
  What is the probability these two people will go on a date if introduced to each other?
  If we modeled this as a classification problem, the binary answer would basically always be "no". The goal is to find matches that are slightly less unlikely than others.
- Credit card transactions
  What is the probability that this transaction is fraudulent?
  The probability is important, because – in combination with the amount of the transaction – it determines the overall risk and thus the right course of action.

Structured output spaces

The output space consists of structured objects, like sequences or trees.

Dating service

Input: description of a person
Output: rank-ordered list of all possible matches

$\mathcal{Y}$ = space of all permutations

Example:
$x = Tom$
$y = (Nancy, Mary, Chloe, \ldots)$

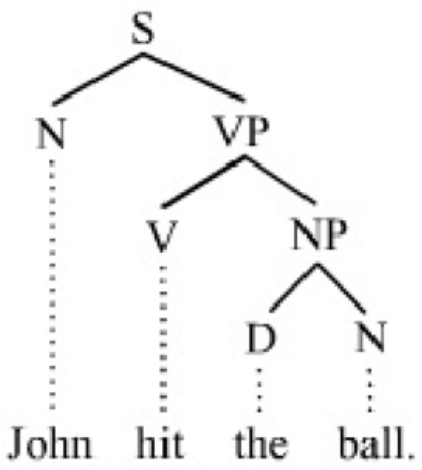
Language processing

Input: English sentence
Output: parse tree showing grammatical structure

$\mathcal{Y}$ = space of all trees

Example:
$x =$ "John hit the ball"

$y =$



A basic classifier: nearest neighbor

Given a labeled training set $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$.

Example: the MNIST data set of handwritten digits.

To classify a new instance x:

- Find its nearest neighbor amongst the $x^{(i)}$
- Return $y^{(i)}$

The data space

We need to choose a distance function.

Each image is 28×28 grayscale.
One option: Treat images as 784-dimensional vectors, and use Euclidean (ℓ_2) distance:

$$\|x - x'\| = \sqrt{\sum_{i=1}^{784} (x_i - x_i')^2}.$$

Summary:

- Data space $\mathcal{X} = \mathbb{R}^{784}$ with ℓ_2 distance
- Label space $\mathcal{Y} = \{0, 1, \ldots, 9\}$

Performance on MNIST

Training set of 60,000 points.

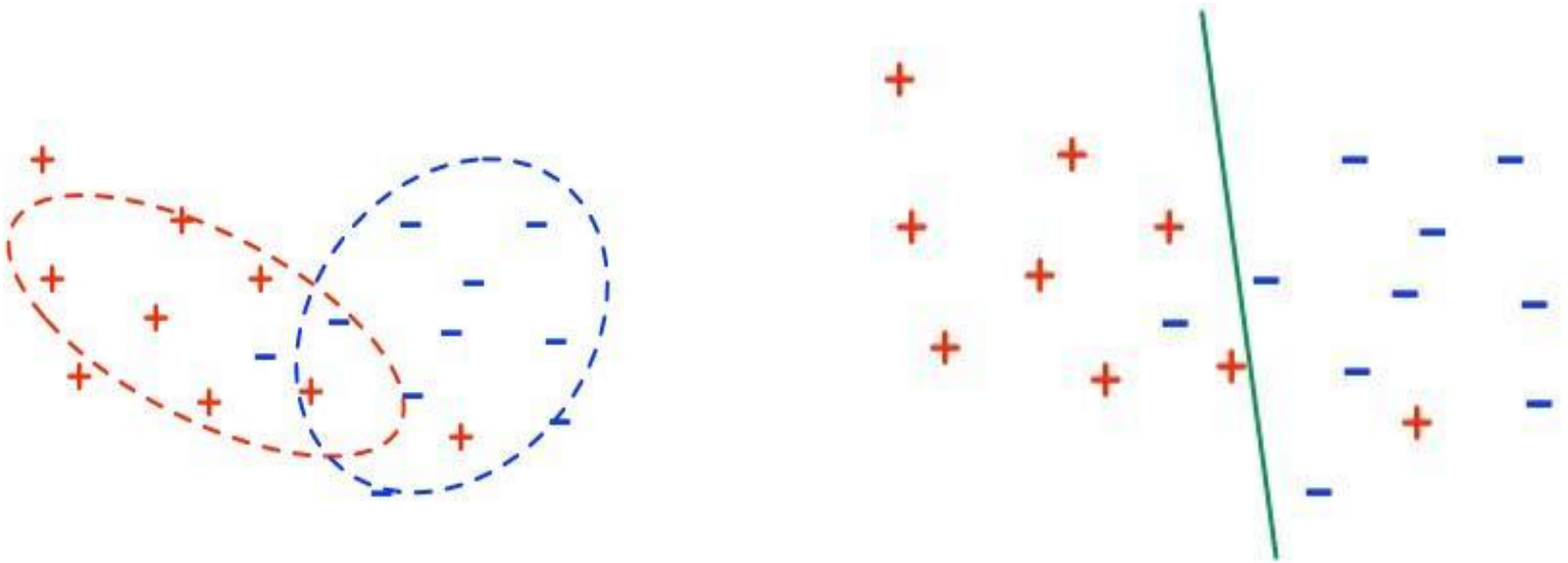
- What is the error rate on training points? **Zero.**
  In general, **training error** is an overly optimistic predictor of future performance.
- A better gauge: separate test set of 10,000 points.
  Test error = fraction of test points incorrectly classified.
- What test error would we expect for a random classifier? **90%.**
- Test error of nearest neighbor: **3.09%.**

Examples of errors:

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

Quick Quiz

- Probability of a patient having liver disease is 0.1, and the probability of an incoming patient being alcoholic is 0.05. If the probability of being alcoholic given that the person has liver disease is 0.07, then what is the probability of an incoming patient having liver disease given that he is alcoholic?

  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

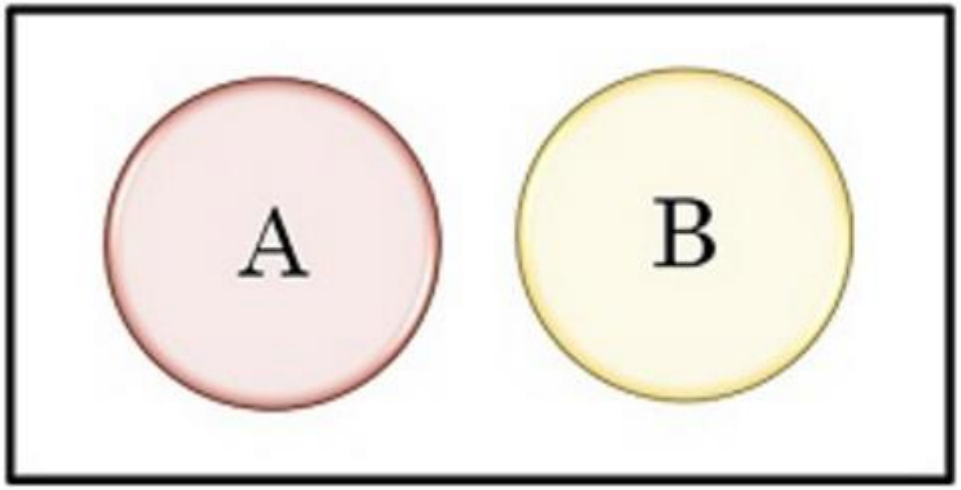
- Disjoint events cannot happen at the same time.
- e.g.: when tossing a coin, the result can either be heads or tails but cannot be both.

Independent

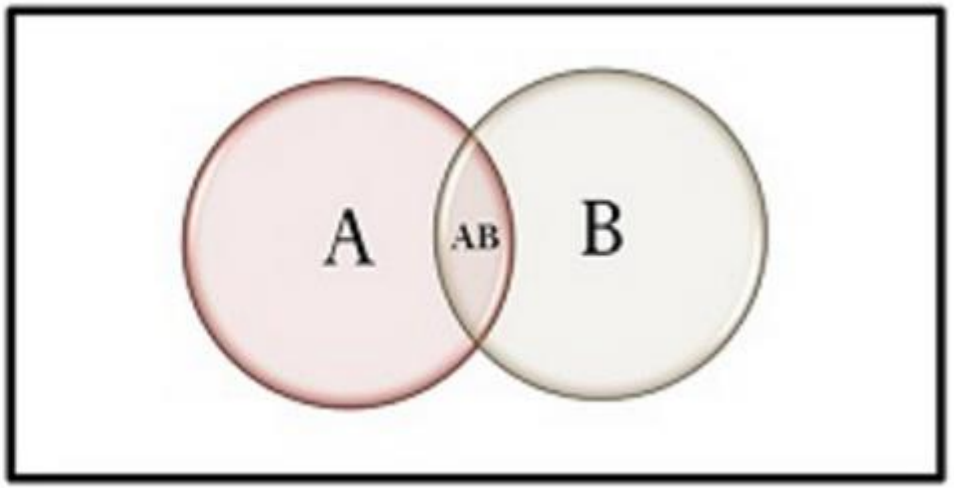
- Occurrence of one event does not influence the other(s).
- e.g.: when tossing two coins, the result of one flip does not affect the result of the other.

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 ,

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

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

Generative models

An unknown underlying distribution D over $X \times Y$.

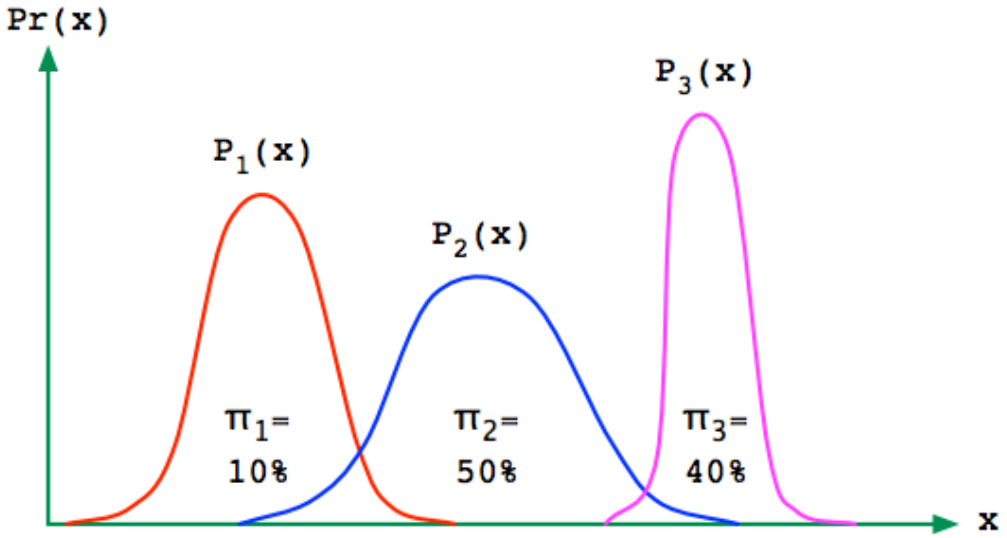
Generating a point (x , y) in two steps:

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

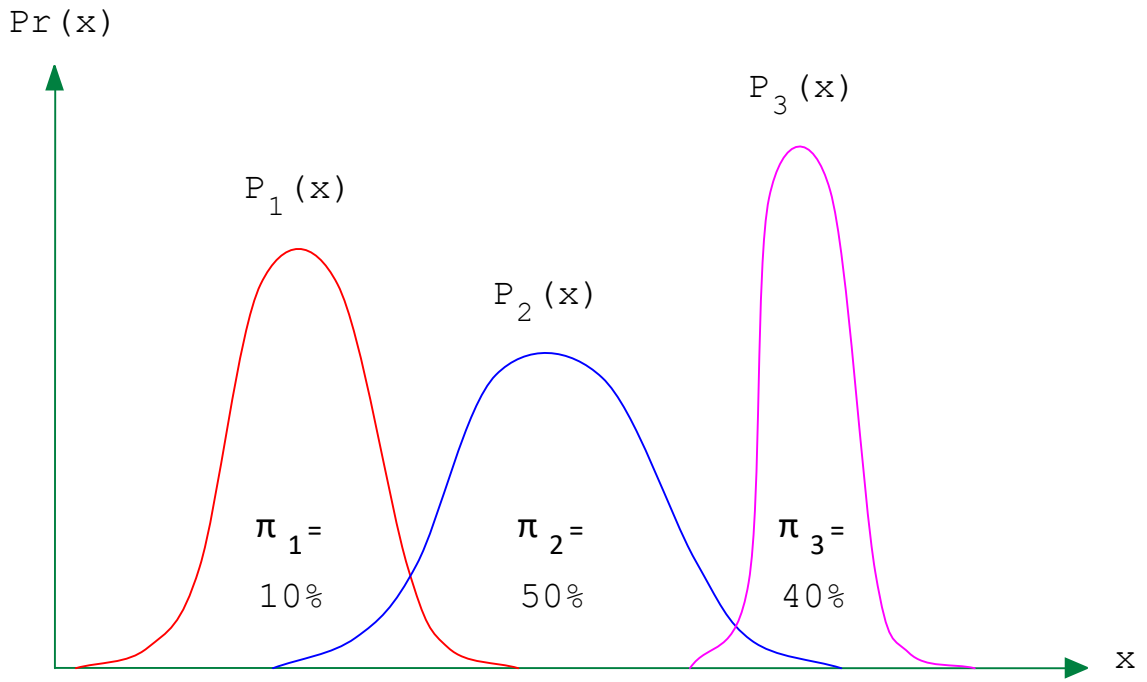
Y = {1, 2, 3}



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 $Y = \{1, 2, \ldots, k\}$, 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 R^p can be hard.

Solution: approximate each P_j with a simple, parametric distribution.

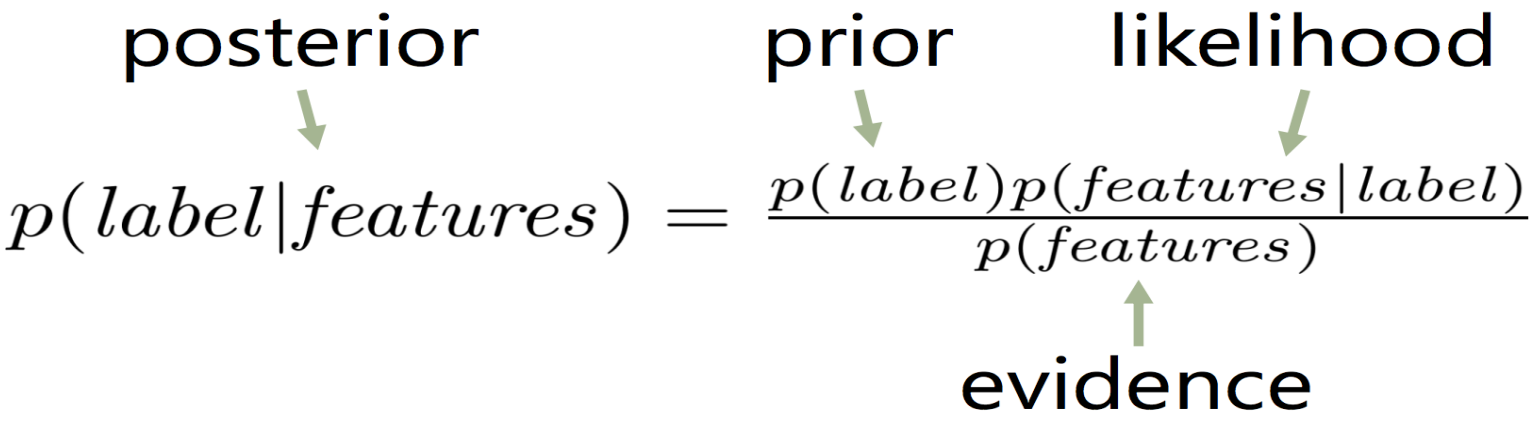
Some options:

- 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* |*data*))
2. Makes a conditional independence assumption that features are independent given the label.

$P(feature_i , feature_j | label) = P(feature_i | label) \cdot P(feature_i | label)$



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

Naive Bayes

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

Binarized MNIST:

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

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

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

Smoothed estimate of coin bias

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 = # of instances of class j

n_{ji} = # of instances of class j 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!

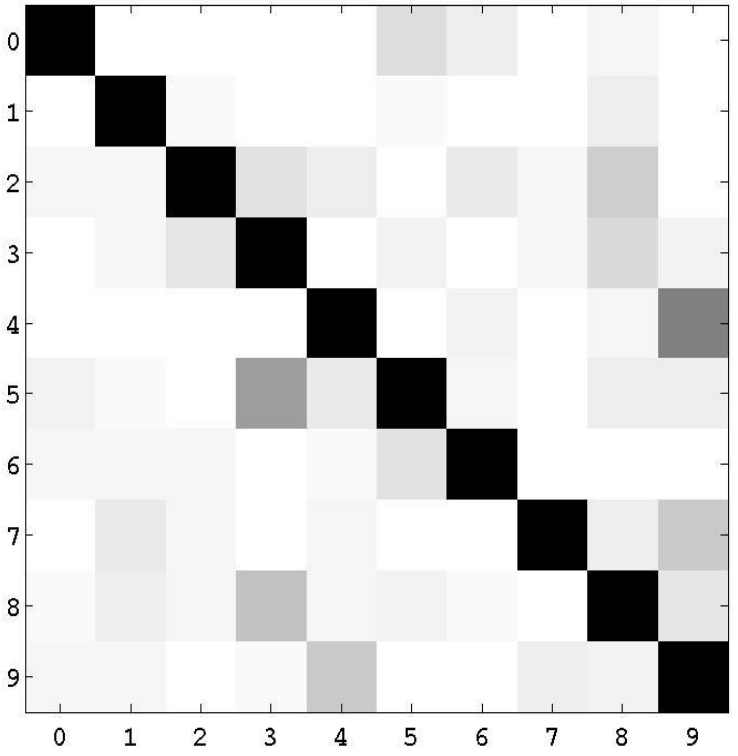
$$w_i^{(j)} = log(p_{ji}) - log(1 - p_{ji}), b^{(j)} = log(\pi_j) + \sum_{i}^{p} log(1 - p_{ji})$$

Example: MNIST

Result of training: mean vectors for each class.

Test error rate: 15.54%.

Visualization of the “confusion matrix” →



Other types of data

How would you handle data:

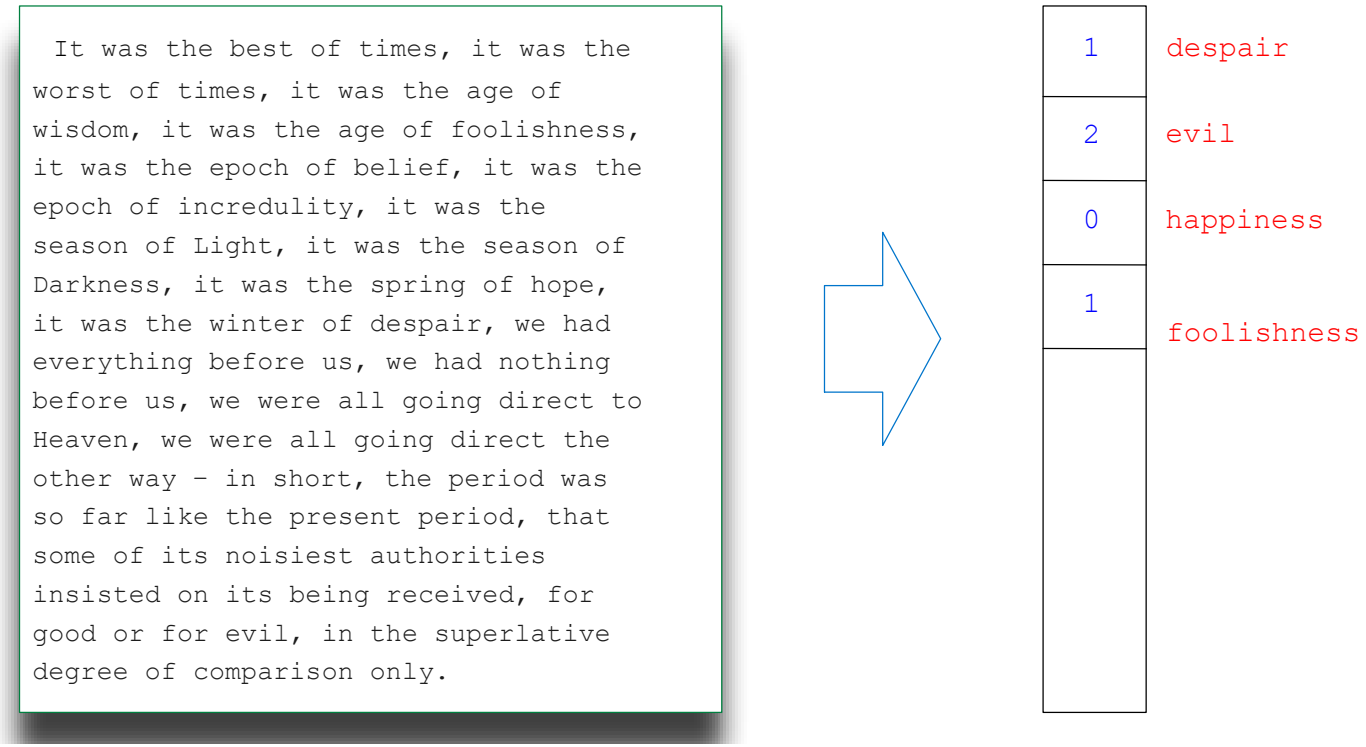
- Whose features take on more than two discrete values (such as ten possible colors)?
- Whose features are real-valued?
- Whose features are positive integers?
- Whose features are mixed: some real, some Boolean, etc?

How would you handle "missing data": situations in which data points occasionally (or regularly) have missing entries?

- At train time: **???**
- At test time: **???**

Handling text data

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



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

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

where x_i = # of times the ith word appears in the document.

A standard distribution over such document-vectors x: the **multinomial**.

Multinomial naive Bayes

Multinomial distribution over a vocabulary V:

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

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

For naive Bayes: one multinomial distribution per class.

- Class probabilities $\pi_1, \ldots, \pi_k$
- Multinomials $p^{(1)} = (p_{11}, \ldots, p_{1|V|}), \ldots, p^{(k)} = (p_{k1}, \ldots, p_{k|V|})$

Classify document x as

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

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

Bernoulli vs Multinomial

Naive Bayes using Bernoulli Distribution

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

Naive Bayes using Multinomial Distribution

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

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 classification

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

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