ML2 – Probabilities; Bayesian decision theory

Stefano Rovetta

A.y. 2023-2024

Possible scenarios

1 (useful mostly for reasoning):

- Learner = hypothesis space = $\mathscr{H}$ is fixed
- Data are fixed (population)

→ Find correct learners in $\mathscr{H}$ (a *representation* task)

2 (not realizable):

- No data necessary: Probabilities are assumed to be known!

→ Find best hypotesis space $\mathscr{H}$ and optimal learner

3 **(your usual situation)**:

- Data will be **stochastic** but probabilities are **not known**
- Hypothesis space $\mathscr{H}$ fixed, chosen in advance

→ Find learners in $\mathscr{H}$ which are correct **for any possible realization of the data** (a *generalization* task)

Scenario 2:
Complete (probabilistic) knowledge

Probabilities

Probability

An expression of uncertainty.

- **Frequentist probability**
  Probability of an event as a generalization of the frequency of occurrence of that event in infinite repetition of an experiment (**trial**).
- **Subjective probability**
  Probability of an event as a confidence in the fact that the event itself will occur, even in a single experiment.

Probability

Something that may or may not happen is called an **outcome**

Call it ω

All possible outcomes constitute the **sample space** – the set $\Omega = \{\omega_1, \omega_2, \ldots, \omega_N\}$ of all possible outcomes

E.g., results of an experiment, measures of a quantity, weather conditions...

We call **events** all possible subsets of Ω = combinations of possible outcomes:

Event space:

$$\left\{ \emptyset, \{\omega_1\}, \{\omega_2\}, \ldots, \{\omega_N\}, \{\omega_1, \omega_1\}, \{\omega_1, \omega_2\}, \ldots, \{\omega_1, \omega_N\}, \{\omega_2, \omega_2\}, \{\omega_2, \omega_3\}, \ldots, \Omega \right\}$$

$$= \{A_1, A_2, A_3, \ldots, A_{2^N}\}$$

The **power set** of Ω, which has cardinality (size) 2^N

Examples

Tossing a coin

Outcomes: head H or tail $T \longrightarrow \Omega = \{H, T\}$

All events: $\{H\}, \{T\}, \{HT\}, \emptyset$

Only subsets containing one element are possible, the others will not occur:
Events here are **mutually exclusive**

Tossing a coin twice

$\Omega = \{HH, TT, TH, HT\}$

Some possible events:

- "tail only": $A = \{TT\}$
- "at least one head": $A = \{HT, TH, HH\}$
- "no two results are the same": $A = \{HT, TH\}$

Some more examples

Weather forecasts

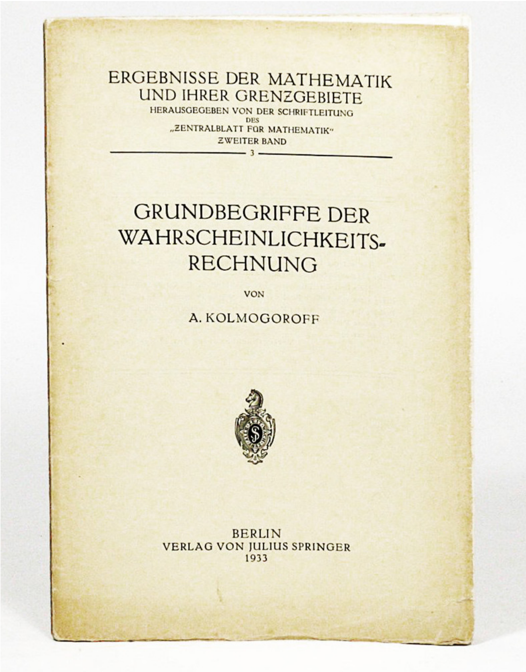
Outcomes: $\Omega = \{S \text{ sunny}, O \text{ overcast}, R \text{ rain }\}$
(mutually exclusive)

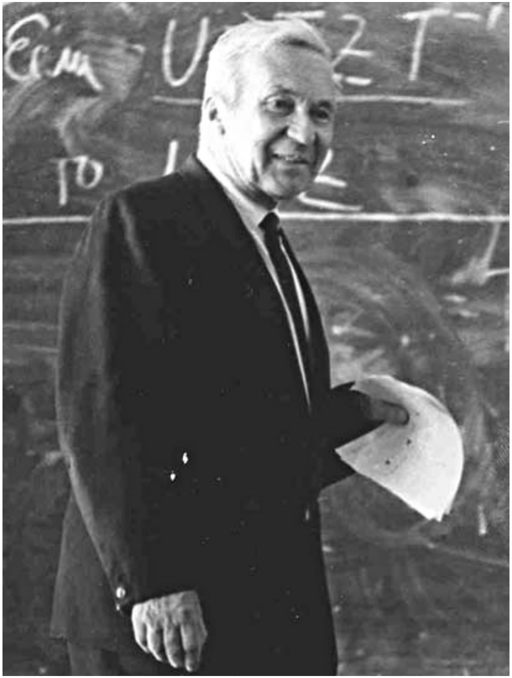
Weather forecasts for the next three days

Outcomes: $\Omega = \{SSS, SSO, SSR, SOS, SOO, SOR, \ldots\}$ ($3^3 = 27$ combinations)

Possible event: today it is Wednesday; will it be sunny on Saturday?
$A = \{SSS, SOS, SRS, OSS, OOS, ORS, RSS, ROS, RRS\}$

Andrej N. Kolmogorov





Set-theoretic concepts apply

Summary of Terminology

Ω	sample space
ω	outcome (point or element)
A	event (subset of Ω)
A^c	complement of A (not A)
$A \bigcup B$	union (A or B)
$A \bigcap B$ or AB	intersection (A and B)
$A - B$	set difference (ω in A but not in B)
$A \subset B$	set inclusion
$\emptyset$	null event (always false)
Ω	true event (always true)

We assign a number to each event

$P(A)$ the probability of event A

Axioms of probability:

1. $P(A) \geq 0$
2. $\sum_{i=1}^{N} P(\omega_i) = 1$, or $P(\Omega) = 1$
3. If A_1 and A_2 are mutually exclusive events
(viewed as sets: if they are **disjoint** = have zero intersection),
then $P(A_1 \textbf{ or } A_2) = P(A_1 \cup A_1) = P(A_1) + P(A_1)$

More properties

(derived from the axioms)

1. $P(A) \leq 1$
2. If A_1 and A_2 are not mutually exclusive,
   then $P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$
3. $P(A_1 \cap A_2) = P(A_1)P(A_2)$ for *independent events*

two events are *independent* if the outcome of one event does not influence the outcome of the second event.
(Opposite of mutually exclusive)

Example

$\Omega^1 = \{1, 2, 3, 4, 5, 6\}$
$P(\omega_i) = P(\omega_j)\ \forall\, i, j$

$\sum_i P(\omega_i) = P(\Omega^1) = 1 \ \Rightarrow\ P(\omega_i) = 1/6\ \forall i$

Suppose we want to know how likely it is that the result is less than 3:
Event $= \{1, 2\} \quad \Rightarrow \quad P(\omega < 3) = P(1 \cup 2) = P(1) + P(2) = 1/3$

Example

$$\begin{aligned}\Omega^2 =& \Omega^1 \times \Omega^1 \\ =& \Big\{ \ \{1,1\},\{1,2\},\{1,3\},\{1,4\},\{1,5\},\{1,6\},\{2,1\},\{2,2\},\{2,3\},\{2,4\},\dots \Big\}\end{aligned}$$

Example (cont.)

We bet on 10. What is the probability of winning?

$$P(10) = P\left(\{4,6\}) \cup \{5,5\} \cup \{6,4\}\right)$$

Outcomes are independent, so $P(\{\omega_i, \omega_j\}) = P(\{\omega_i\} \cap \{\omega_j\}) = P(\omega_i)P(\omega_j)$:

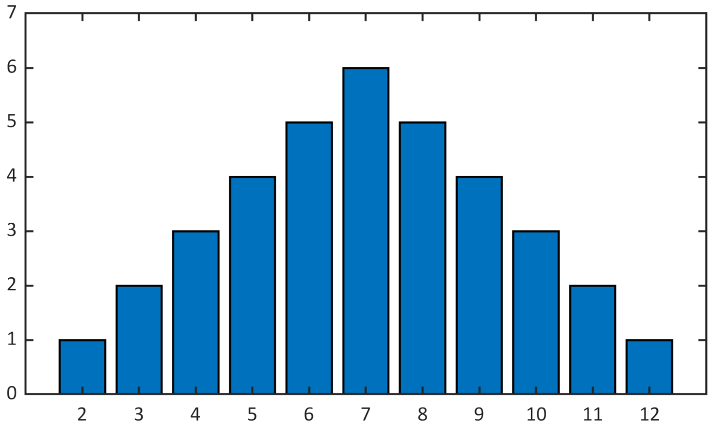
$$P(5 \cap 5) = P(5)\,P(5) = \frac{1}{6}\frac{1}{6} = \frac{1}{36}$$

Probability of any other pair $(6 \cap 4)$, $(4 \cap 6)$... is the same (fair dice)

$$P(10) = \left(P(4)P(6)\right) + \left(P(5)P(5)\right) + \left(P(6)P(4)\right) = 3 \times \frac{1}{36} = \frac{1}{12}$$

$$= \text{num. combinations} \times \left(\frac{1}{6}\right)^{\text{num. dice}}$$

Note that...



...probabilities for different outcomes are not the same!

Continuous sample spaces

Examples: Age of a person, voltage in a circuit, force or torque in a mechanical link, temperature, time of the day…

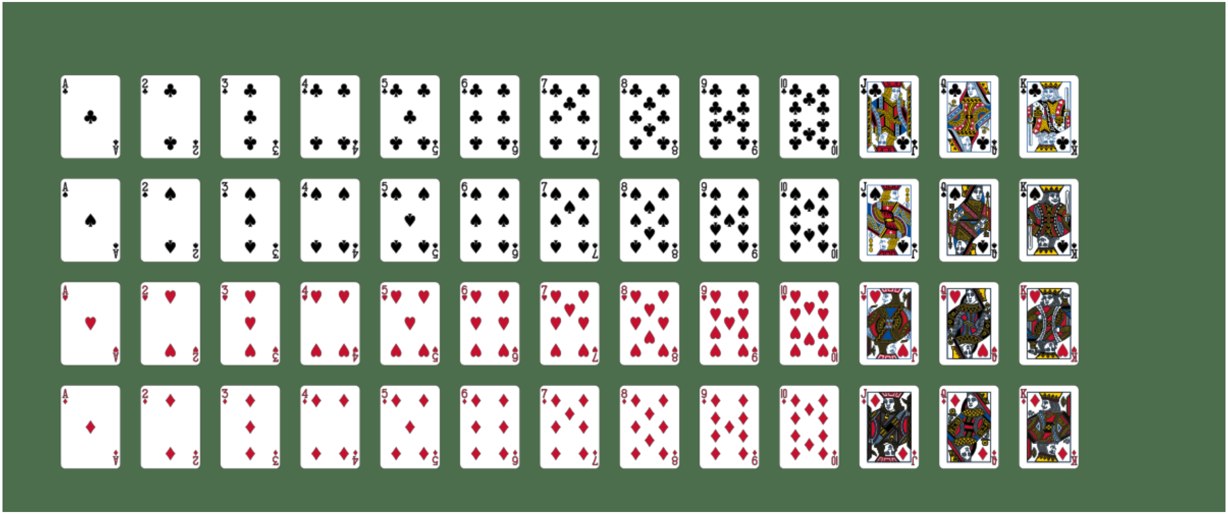
All definitions still apply, at least in non-pathological situations

A trick for visualising probabilities:

- Think of probabilities for discrete sample and event spaces as **counts** and **frequencies**
- Think of probabilities for continuous sample and event spaces as **areas**, **volumes**, **hypervolumes**…

Further example

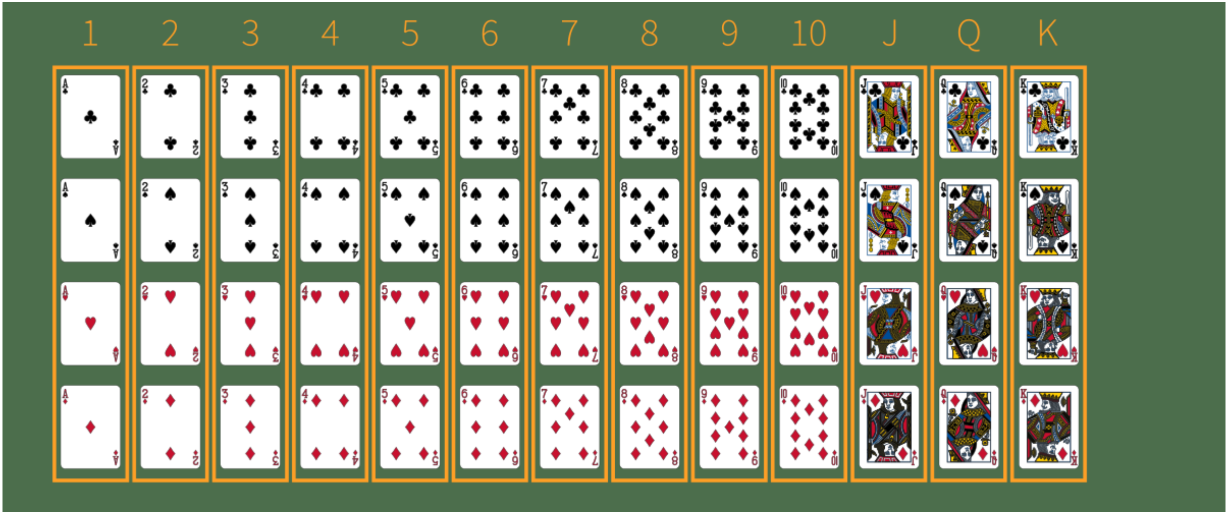
You draw one playing card from a full 52-card deck



- Outcomes: $\omega =$ one specific card
- Sample space Ω: The set of all 52 possible cards
- Event space 2^{Ω}: all possible types of cards (examples follow: compute probabilities for each!)

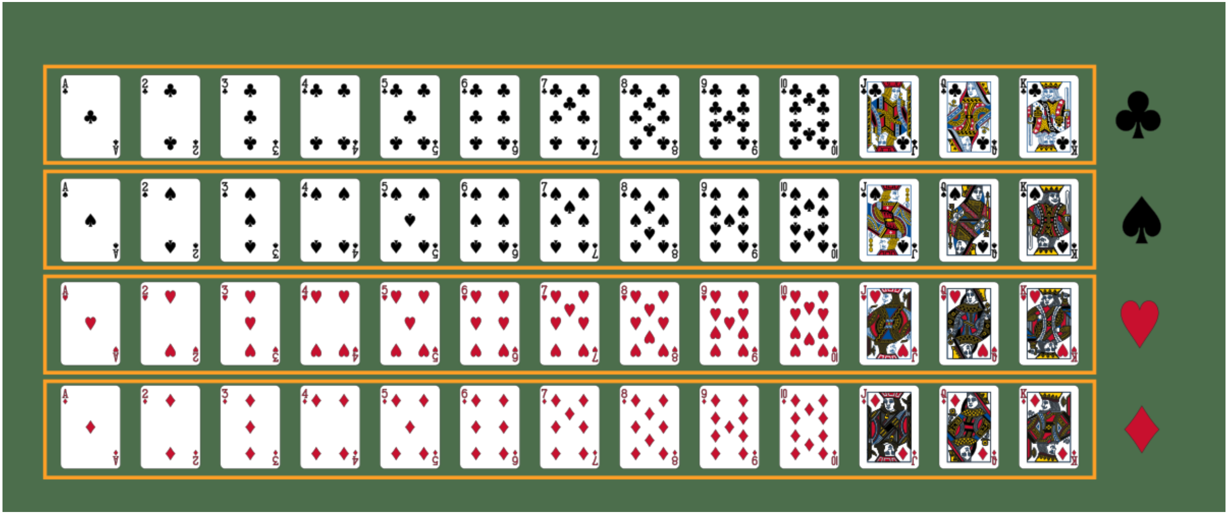
Some possible events

Bet on a specific value



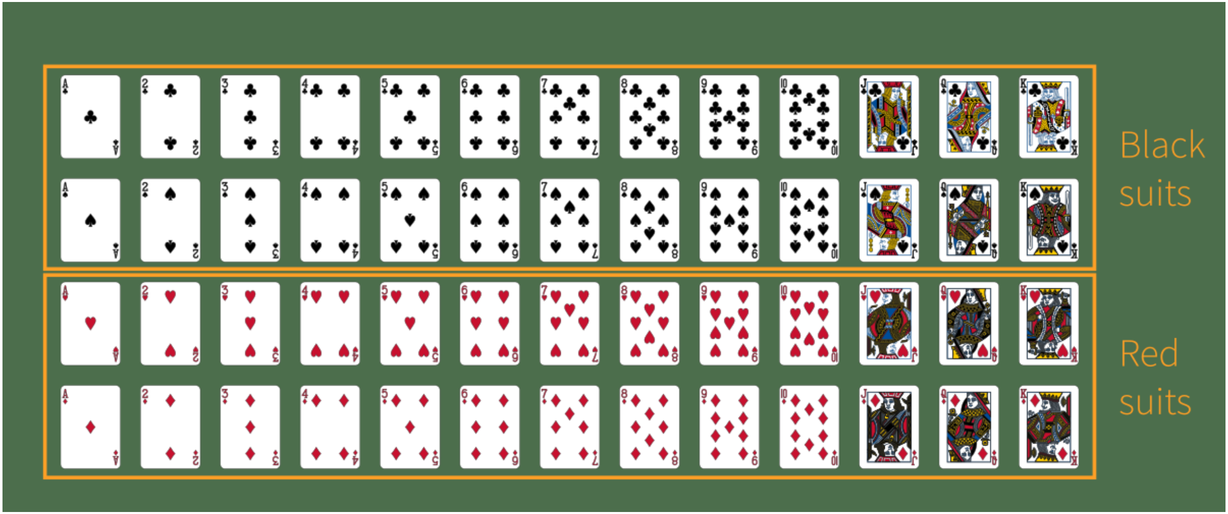
Some possible events

Bet on a specific suit



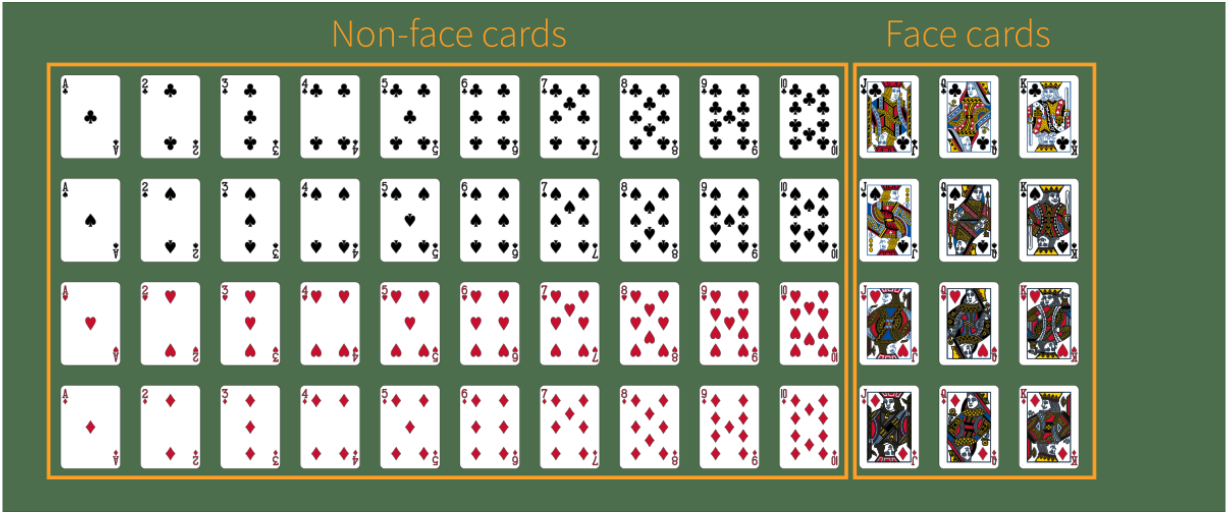
Some possible events

Bet on a specific colour



Some possible events

Bet on whether the card is or is not a face card



Further possible events

The power set is of cardinality $2^{52} \approx 4.5 \times 10^{15} \ldots$
Very many different bets are possible:

- Bet on any face card of a specific colour
- Bet on a card of odd/even value
- Bet on a card whose value is 2 or 7
- ...
- ...and – of course – bet on a specific card (individual outcomes are possible events)

Remark: Probabilities can be seen as quantities generated by functions of events (note that we already write them as such, $P(A)$!)

Random variables

A **random variable** is a numerical variable that doesn't have a fixed value, but changes according to a given probability law.

Example: random number generators in programming
(C/C++ rand() from stdlib.h/cstdlib, Matlab rand, Python numpy.random.rand()...)

Example: voltage measurements at the terminals of a heated resistor

Values may be real or discrete

Characterisation of probability of any type of random variable

Discrete or continuous

Definition

Let x indicate a random variable with values in $\mathscr{X}$.

Given a specific value $\hat{x} \in \mathscr{X}$,

$$P_{\mathscr{X}}(\hat{x}) = \Pr(x \leq \hat{x})$$

is the **cumulative distribution function** or simply **distribution function** of events in $\mathscr{X}$.

Characterisation of probability of discrete random variables

Definition

Let x indicate a random variable with values in a numerable set $\mathscr{X}$,
e.g., an integer number in $\mathscr{X} = \{0, 1\}$.

Given a specific event $\hat{x} \in \mathscr{X}$,

$$F_{\mathscr{X}}(\hat{x}) = \Pr(x = \hat{x})$$

is the **probability mass function** of $\mathscr{X}$.

Gives the finite probability that a random event x has a specific value $\hat{x}$.

Characterisation of probability of continuous random variables

Definition

Let x indicate a random variable with values in a non-numerable set $\mathscr{X}$, e.g., a real number in $\mathscr{X} = [0,1]$.

A function $f_{\mathscr{X}}$ such that

$$P_{\mathscr{X}}(\hat{x}) = \int_{-\infty}^{\hat{x}} f_{\mathscr{X}}(x)\,\mathrm{d}x$$

is the **probability density function** of $\mathscr{X}$.

Gives the **infinitesimal** probability that a random event x has value $\hat{x}$. (Infinitesimal = **null**)

Its **definite integral on a random interval** $[\hat{t}_1, \hat{t}_2] \in \mathscr{E}$ is the **finite** probability that $\hat{t}_1 < e < \hat{t}_2$.

Expectation

An important use of probability functions:
compute the “most likely” or **expected** value of some random X.
In the case of discrete events, it is a weighted sum

$$\mathrm{E}\{X\} = \sum_i \xi_i F_X(\xi_i)$$

ξ_i are the possible values of X
The symbol $\mathrm{E}\{\}$ is the expectation operator

Expectation

For real-valued X:

$$\mathrm{E}\{X\} = \int_{\mathscr{X}} \xi \, f_x(\xi) \, d\xi$$

Conditional probability

$P(E \mid F)$ — The probability of an event E **given** the knowledge that another event F has occurred

Example (dice): $P(\text{'10'}) = 1/12$

But if we know that the first dice is '2', then $P(\text{'10'} \mid \text{first dice is '2'}) = 0$
If we know that the first dice is '5', then $P(\text{'10'} \mid \text{first dice is '5'}) = 1/6$

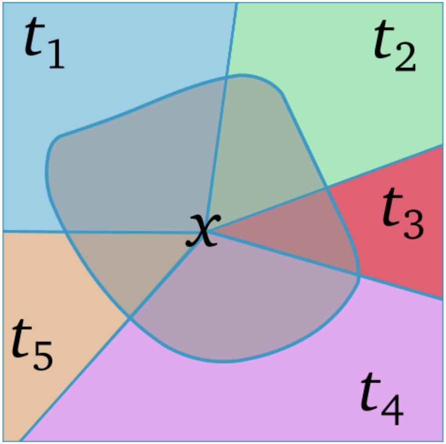
Bayesian probabilities

Bayesian probability jargon

t	is a hypothesis (event)
t_i	is a set of alternative hypotheses, $t \in \{t_i\}$
x	is an experimental observation
$P(t)$	is the **a priori probability** of hypothesis t: probability that t is true before seeing any experimental observation
$P(t\|x)$	is the **a posteriori probability** of hypothesis H after observing X
$P(x\|t)$	is the **likelihood** of observing x when t holds (is verified, is true, is certain)
$P(x)$	is the **marginal probability** of x, the probability of observing x in any case

Total probability theorem

How to compute a marginal probability:



$$P(x) = \sum_i P(x|t_i)P(t_i)$$

Bayes' theorem

$$P(t_i|x) = \frac{P(x|t_i)P(t_i)}{P(x)}$$

Gives the probability of hypothesis t_i after seeing an experimental observation x

Bayes' theorem, alternate form

$$P(t_i|x) = \frac{P(x|t_i)P(t_i)}{\sum_{j=1}^{c} P(x|t_j)P(t_j)}$$

- Uses the total probability theorem
- Does not require $P(x)$
- The denominator is also known as a **partition function**.
  It acts as a **normalizer** that makes the sum of all $P(t_i|x)$ (for $i : 1 \ldots c$) equal one.

Bayesian Decision Theory

Pattern Classification, 2nd Edition

Richard O. Duda, Peter E. Hart, David G. Stork

ISBN: 978-0-471-05669-0
Wiley-Interscience
680 pages
November 2000

The decision problem

- c possible, mutually exclusive events, or "states of nature"

$$\{t_1, t_2, \ldots, t_c\}$$

- s possible actions or "decisions" that we may make

$$\{y_1, y_2, \ldots, y_s\}$$

We want to find a rule that, given any state t, makes the most suitable decision y.
The problem is that t may not be observable!

The decision process

From simplest to most complex:

1. Decisions:
   states of nature $\longrightarrow$ actions

2. Uncertainty in the states of nature:
   probabilities $\longrightarrow$ states of nature $\longrightarrow$ actions

3. Conditional decisions:
   observations $\longrightarrow$ probabilities $\longrightarrow$ states of nature $\longrightarrow$ actions

4. Using Bayes formula:
   observations $\rightarrow$[BAYES]$\rightarrow$ probabilities $\longrightarrow$ states of nature $\longrightarrow$ actions

1: Decisions

states of nature $\longrightarrow$ actions

Rationale

- Given the **state of nature** t, we act consequently and make a decision, or take an action, y.
- So the decision or action is a function of the state of nature, $y(t)$.
- Every decision has a **cost** – a high-cost decision is a “wrong” decision
- We measure this cost using a **loss function** $\lambda(y, t)$
- The loss function evaluates the cost of **each decision**, depending on the **true state of nature**, including our **subjective** considerations (preferences) as well as **objective** elements (actual costs)
- We build a **decision rule**, the function $y(t)$ that given t produces y, attempting to **minimise the loss**.

Example

Buying a pair of shoes. I have two choices:

- (Italian) size 43, cost 200
- Size 42, cost 120

Depending on my shoe size, a possible loss function is:

$$\lambda(y, t) = \lambda\left(\boxed{\text{what I buy}}, \boxed{\text{my actual size}} \right) =$$

$$= \begin{array}{r|c|c|} & \textbf{Buy 43} & \textbf{Buy 42} \\ \hline \textbf{I have size 43} & 200 & 120 + \text{uncomfortable} \\ \textbf{I have size 42} & 200 + \text{uncomfortable} & 120 \end{array}$$

Decision rule that minimises the loss:

- If you have size 42, buy the cheaper pair which is 42
- If you have size 43… depending how you quantify “uncomfortable”, you
  - buy 42 (if you value “uncomfortable” less than 80)
  - or 43 (otherwise).

2: Uncertainty in the states of nature

probabilities $\longrightarrow$ **states of nature** $\longrightarrow$ **actions**

Probabilistic modelling

Usually t is not known with certainty at the time of making a decision.

Example: Insurance
The insurance company must establish the reimbursement policies *before* accidents happen.

Example: Measuring instrument
Due to measurement errors, the reading of an instrument *is not* necessarily the true value of a physical quantity.

$\Rightarrow$ We use probabilities.

$$
\begin{array}{ccccc}
t_1 & t_2 & t_3 & \dots & t_c \\
P(t_1) & P(t_2) & P(t_3) & \dots & P(t_c)
\end{array}
$$

The cost of a decision now cannot be known with certainty

$\Rightarrow$ To evaluate each possible decision, we use the **expected** loss
(average over all possibilities, weighted with the respective probabilities)

For decision y_1:

$$\begin{aligned} R(y_1) &= \lambda(y_1, t_1)P(t_1) + \lambda(y_1, t_2)P(t_2) + \quad \ldots \quad + \lambda(y_1, t_c)P(t_c) \\ &= \sum_{j=1}^{c} \lambda(y_1, t_j)P(t_j) \end{aligned}$$

For a generic decision y_i:

$$R(y_i) = \sum_{j=1}^{c} \lambda(y_i, t_j)P(t_j)$$

$\boxed{R(y_i) \text{ is the } \textbf{risk} \text{ (expected loss) of decision } y_i}$

The decision rule

In the presence of uncertainty, the decision rule is as before

But this time it must minimise the **risk** of the decision

In other words, it must minimise the loss **on average**

3: Conditional decisions

observations $\longrightarrow$ **probabilities** $\longrightarrow$ **states of nature** $\longrightarrow$ **actions**

Unobservable quantities

Many interesting quantities cannot be measured directly.

Example: Disease
A doctor can measure sign and symptoms, but not directly the disease. Given (for instance) fever, this indicates that there *may be* a certain disease (e.g. flu). But flu itself cannot be measured with an instrument.

Example: Stock market trend
There are several indicators that the stock market *may go* in a certain direction in the future (increasing or decreasing), but the true dynamics of the stock market cannot be modelled.

$\Rightarrow$ We use **experimental observations** and **conditional probabilities**.

$$\mathbf{x}$$

$$t_1 \qquad t_2 \qquad t_3 \qquad \ldots \qquad t_c$$

$$P(t_1|\mathbf{x}) \quad P(t_2|\mathbf{x}) \quad P(t_3|\mathbf{x}) \quad \ldots \quad P(t_c|\mathbf{x})$$

Note: observations may be

- a scalar x
  (for instance an individual sensor reading)
- a vector $\mathbf{x}$
  (for instance all the pixels in an image)

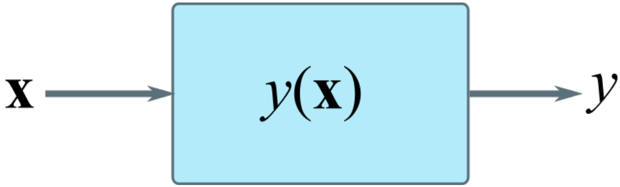
Conditional risk

$$R(y_i \mid \mathbf{x}) = \sum_{j=1}^{c} \lambda(y_i, t_j) P(t_j \mid \mathbf{x})$$

$R(y_i \mid \mathbf{x})$ is the **conditional risk** of decision y_i
when we have the **experimental observation** $\mathbf{x}$

General structure of a decision rule

In the most general case we are considering, we have to design:



A system that, given any observation, outputs the best decision

The decision rule

The decision rule is more or less as before.
But this time, it should minimise the conditional risk **for all possible observations x!**

Of course this may not be possible in all cases.
So the realistic criterion is:

The decision rule must minimise the **average (expected) risk over all possible observations**

So we must take the expectation of the risk over the observations as our criterion to be minimised.

Expected risk

If observations are discrete:

$$R = \sum_{x \in \mathscr{X}} R(\, y(\mathbf{x}) \,|\, \mathbf{x}\,)\, P(\mathbf{x})$$

where $P(\mathbf{x})$ is the probability mass function of experimental observations and $\mathscr{X}$ is the set of all possible inputs (the “input space”).

If observations are continuous:

$$R = \int_{\mathscr{X}} R(\, y(\mathbf{x}) \,|\, \mathbf{x}\,)\, p(\mathbf{x})\, d\mathbf{x}$$

where $p(\mathbf{x})$ is the probability density function of experimental observations.

4: Using Bayes formula

observations $\longrightarrow$[BAYES]$\longrightarrow$ **probabilities** $\longrightarrow$ **states of nature** $\longrightarrow$ **actions**

Bayes decision theory

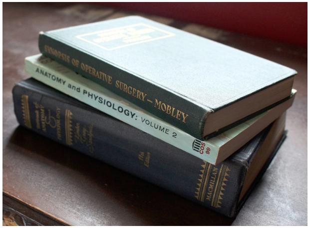
Example: A doctor should diagnose a disease after visiting a patient

He records all observation and measurements into a patient record **x**

Bayes decision theory

Usually a doctor has some information available from his medicine textbooks and from his own experience:

- The incidence of diseases
  $\rightarrow P(t_i)$
- The typical and not-so-typical signs and symptoms of diseases
  $\rightarrow P(\mathbf{x}|t_i)$



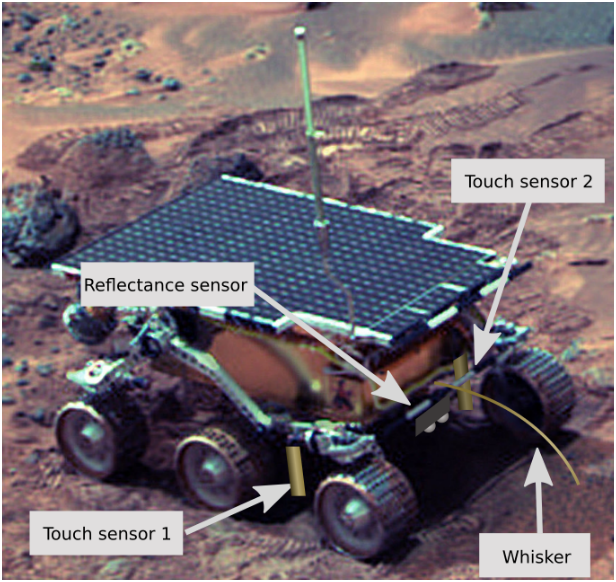
Bayes decision theory

From Bayes' theorem:

$$P(t_i|\mathbf{x}) = \frac{P(\mathbf{x}|t_i)P(t_i)}{\sum_{j=1}^{c} P(\mathbf{x}|t_j)P(t_j)}$$

Numeric example

Autonomous rover for unmanned explorations, equipped with four sensors



Numeric example

cont.

- **Possible states of nature:**
  $t \in \{t_1, t_2, t_3\} = \{\text{'water', 'solid ground', 'sand'}\}$

- **Their a priori probabilities:**
  $P(t_1) = .2$, $P(t_2) = .4$, $P(t_3) = .4$

- **Possible decisions:**
  $y \in \{y_1, y_2\} = \{\texttt{rover:retract}, \texttt{rover:advance}\}$

- **Input observations (readings from sensors):**
  $\mathbf{x} = [\ \texttt{groundTouchSens1}, \texttt{groundTouchSens2}, \texttt{groundOptSens}, \texttt{groundWhisker}\ ]$

- **We now receive an input observation x for which likelihoods are:**
  $P(\mathbf{x}|t_1) = .5$, $P(\mathbf{x}|t_2) = .9$, $P(\mathbf{x}|t_3) = .1$

- NOTE that LIKELIHOODS MAY NOT SUM UP TO 1
  They are not mutually exclusive
  ("in direct competition with each other")

Example

cont.

- **We are given this loss matrix:**

$$\Lambda = \begin{bmatrix} 0.1 & 1.0 & 4.0 \\ 2.0 & 0.1 & 0.1 \end{bmatrix}$$

- **Conditional risk of decision y_1 given observation $\mathbf{x}$:**

$$R(y_1|\mathbf{x}) = \sum_{j=1}^{3} \lambda_{1j} P(t_j|\mathbf{x}) = 0.1 \times 0.5 + 1 \times 0.9 + 4 \times 0.5 = 2.95$$

- **Conditional risk of decision y_2 given observation $\mathbf{x}$:**

$$R(y_2|\mathbf{x}) = \sum_{j=1}^{3} \lambda_{2j} P(t_j|\mathbf{x}) = 2 \times 0.5 + 0.1 \times 0.9 + 0.1 \times 0.5 = 1.14$$

Example

cont.

When we receive input $\mathbf{x}$,

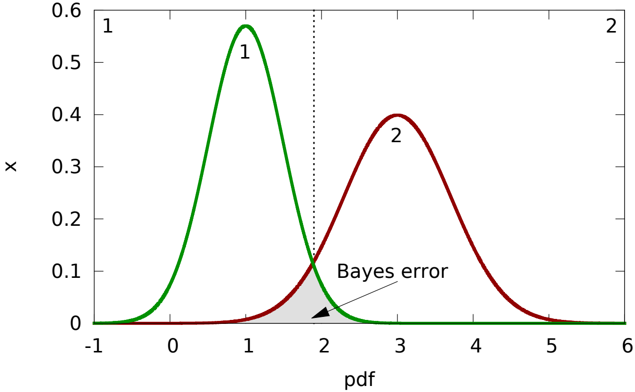
the decision that minimizes the conditional risk is y_2

The Bayes decision criterion: in short

To minimize R, given an input $\mathbf{x}$ the decision rule $y(\mathbf{x})$ should output the decision y that minimizes the ***expected risk*** *R.*

Theoretically optimal criterion (you can't do better than this)

Errors are always possible!



1 and 2 are two example posterior probabilities

The Bayes error, the best possible error probability.

Classification

Classification is a decision problem with:

- $s \equiv c$
- $\{y_1, \ldots, y_s\} \equiv \{t_1, \ldots, t_c\}$

I.e. there is no actual decision to take, we are only recognizing the state of nature (the **class**)

A loss function for classification

zero-one loss:

$$\lambda(y \mid t) = \begin{cases} 0 & \text{if } y = t \\ 1 & \text{if } y \neq t \end{cases}$$

Example: zero-one loss matrix for a three class problem ($c = 3$):

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

What does the 0-1 loss mean?

- All types of errors have the same cost ($= 1$)
- Correct classifications don't have a cost
- $\Rightarrow R$ equals expected probability of error
  (proof: plug zeroes and ones in the definition of conditional cost)

Zero-one loss = minimum-error-rate classification

Designing a classifier

A classifier is a rule $y()$ that receives an observation $\mathbf{x}$
and outputs a **class** $y(\mathbf{x})$.

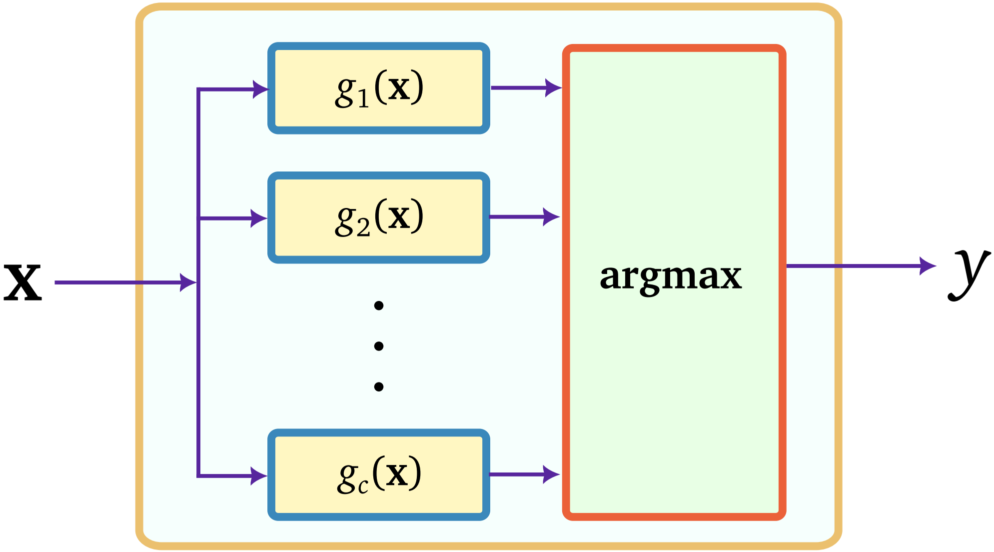
The Bayes decision criterion states that y should minimize $R(y(\mathbf{x}) \mid \mathbf{x})$

A natural idea:

- Build c blocks or "matched filters" g_j, $j : 1 \dots c$
  that compute $g_1(\mathbf{x}) = -R(y = t_1|\mathbf{x}), \dots, g_c(\mathbf{x}) = -R(y = t_c|\mathbf{x})$
- Select $y = t_j$ that has maximum $g_j(\mathbf{x})$

$g_j()$ are called **discriminant functions**

The operation of looking for the location j of the maximum value (the "argument" of the maximum) is called **argmax**



$\mathbf{x}$
$g_1(\mathbf{x})$
$g_2(\mathbf{x})$
$g_c(\mathbf{x})$
argmax
y

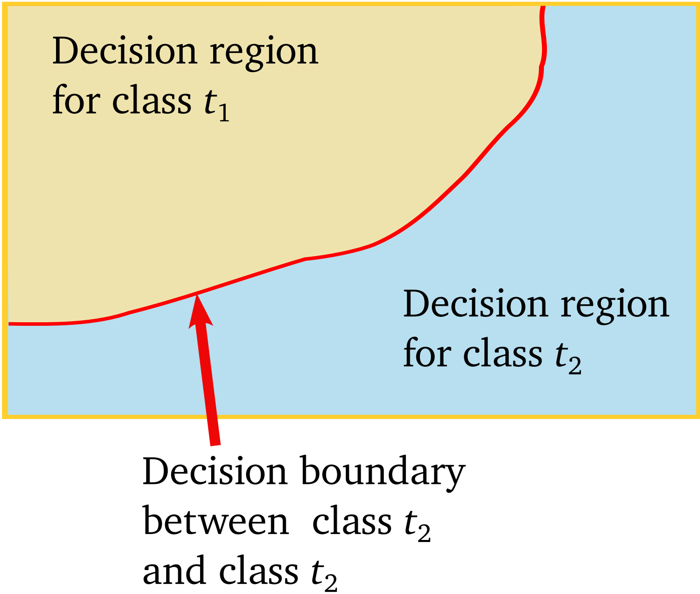
Decision regions

A **decision region** is a subset of the data space with a given minimum-conditional-risk decision (i.e., the decision y is the same for all data in the region)

Decision regions are separated by **decision boundaries** (or **decision surfaces**)

The decision boundary between two regions (say $y = t_j$ and $y = t_k$) are defined by:

$$g_j(\mathbf{x}) = g_k(\mathbf{x})$$


Decision region for class t_1
Decision region for class t_2
Decision boundary between class t_2 and class t_2

Discriminant functions for zero-one loss

In the case of the minimum-error-rate classifier (= using zero-one loss):

$$
\begin{aligned}
g_i(\mathbf{x}) &= -\sum_{j=1, j\neq i}^{c} P(t_j|\mathbf{x}) \\
&= P(t_i|\mathbf{x}) - 1
\end{aligned}
$$

where $P(t_i|\mathbf{x})$ is obtained from Bayes' theorem

Other ways to define discriminant functions

A classifier is defined by the decision boundaries
so the actual functions being compared need not be actually $g_j(\mathbf{x}) = -R(t_j|\mathbf{x})$

They can be any **monotonically increasing transformation** $g_j(\mathbf{x}) = f(-R(t_j|\mathbf{x}))$
that preserves decision boundaries
e.g., $g_j = \log R(t_j|\mathbf{x})$ or $g_j = \frac{1}{1+e^{-R(t_j|\mathbf{x})}}$

This gives us more freedom in building a classifier!

We can use more general **scores** $f(-R(t_j|\mathbf{x}))$

instead of actual conditional risks $-R(t_j|\mathbf{x})$.

Reasonable discriminant functions for zero-one loss

The transformation $f(x) = x + 1$ is monotonic and preserves decision boundaries, so we can avoid the useless -1:

$$
\begin{aligned}
g_i(\mathbf{x}) &= 1 - \sum_{j=1, j \neq i}^{c} P(t_j|\mathbf{x}) \\
&= P(t_i|\mathbf{x})
\end{aligned}
$$

Here we see that if we give the same weight to all errors (0/1 loss) the discriminant functions are simply the probability of each class given the input — so we take the one with maximum probability!

Quite sensible criterion.

A popular classifier: Naive Bayes

Is built using “wrong” discriminant functions based on “naive,” or even “idiot,” assumptions ($\rightarrow$ also “Idiot’s Bayes Classifier”).

- **Recall that:** $\mathbf{x} = [x_1, x_2, x_3, \dots, x_d]$
- Let’s focus on discrete xs
- **So:** $P(t_i|\mathbf{x}) \propto P(\mathbf{x} \mid t_i) P(t_i) = P(x_1, x_2, x_3, \dots, x_d \mid t_i) P(t_i)$
- **In general:** $\Pr(a, b|c) \neq \Pr(a \mid c) \Pr(b \mid c)$
- **Naive assumption:** $\Pr(x_1, \dots, x_d \mid t_i) = \Pr(x_1 \mid t_i) \Pr(x_2 \mid t_i) \cdots \Pr(x_d \mid t_i)$

We are pretending that input variables are **all independent of each other**

Naive Bayes classifier

$$g_i(\mathbf{x}) = P(t_i)\left[P(x_1|t_i) \times P(x_2|t_i) \times \ldots \times P(x_d|t_i)\right]$$
$$= P(t_i)\prod_{j=1}^{d} P(x_j|t_i)$$

How a naive Bayes classifier "learns"

Particularly handy when features are binary (true/false):

To "learn" $P(x_k|t_i)$ we count how often each value of x_k occurs in class t_i in the training set:

$$P(x_k = \text{true}|t_i) = \frac{\text{number of times } x_k = \text{true in class } t_i}{\text{number of observations of class } t_i} = \frac{N_{\text{true},t_i}}{N_{t_i}}$$

$$P(x_k = \text{false}|t_i) = 1 - P(x_k = \text{true}|t_i) = 1 - \frac{N_{\text{true},t_i}}{N_{t_i}}$$

Prior probability of classes:

$$P(t_i) = \frac{\text{number of observations in class } t_i}{\text{number of observations in the training set}} = \frac{N_{t_i}}{N}$$

Not so idiot?

With many features the naive assumption is approximately correct!

Example: Spam detection

- Observations: email messages
- Features:
  - Presence of words typical of spam (from a list)
  - Presence of specific spelling mistakes
  - Mismatch between address shown in links and address actually pointed to
  - Only images and no text
  - Only attachments and little text
  - ...
- Training set: Your “JUNK” email folder
- Target: Your clicks on the “THIS IS SPAM” button