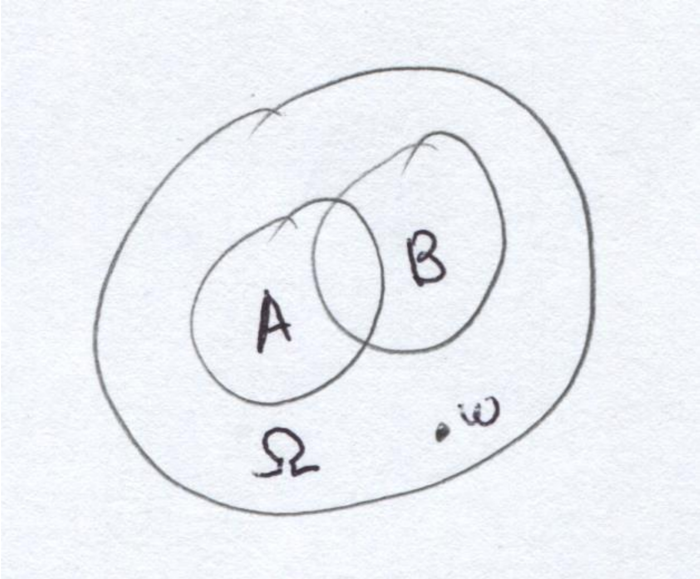
Machine Learning

Probability Theory

Probability space

is a three-tuple (Ω, σ, P) with:

- Ω – the set of elementary events
- σ – algebra
- P – probability measure



σ-algebra over Ω is a system of subsets,
i.e. $\sigma \subseteq \mathcal{P}(\Omega)$ ($\mathcal{P}$ is the power set) with:

- $\Omega \in \sigma$
- $A \in \sigma \quad \Rightarrow \quad \Omega \setminus A \in \sigma$
- $A_i \in \sigma \; i = 1 \dots n \quad \Rightarrow \quad \bigcup_{i=1}^{n} A_i \in \sigma$

σ is closed with respect to the complement and countable conjunction. It follows: $\emptyset \in \sigma$, σ is closed also with respect to the countable disjunction (due to the De Morgan's laws)

Probability space

Examples:

- $\sigma = \{\emptyset, \Omega\}$ (smallest) and $\sigma = \mathcal{P}(\Omega)$ (largest) σ-algebras over Ω
- the minimal σ-algebra over Ω containing a particular subset $A \in \Omega$ is $\sigma = \{\emptyset, A, \Omega \setminus A, \Omega\}$
- Ω is discrete and finite, $\sigma = 2^{\Omega}$
- $\Omega = \mathbb{R}$, the Borel-algebra (contains all intervals among others)
- etc.

Probability measure

$P: \sigma \to [0,1]$ is a „measure" (Π) with the normalizing $P(\Omega) = 1$

σ-additivity: let $A_i \in \sigma$ be pairwise disjoint subsets, i.e. $A_i \cap A_{i'} = \emptyset$, then

$$P\left(\bigcup_i A_i\right) = \sum_i P(A_i)$$

Note: there are sets for which there is no measure.

Examples: the set of irrational numbers, function spaces $\mathbb{R}^\infty$ etc.

Banach-Tarski paradoxon (see Wikipedia ☺):

(For us) practically relevant cases

- The set Ω is „good-natured“, e.g. $\mathbb{R}^n$, discrete finite sets etc.

- $\sigma = \mathcal{P}(\Omega)$, i.e. the algebra is the power set

- We often consider a (composite) „event“ $A \subseteq \Omega$ as the union of elemantary ones

- Probability of an event is

$$P(A) = \sum_{\omega \in A} P(\omega)$$

Random variables

Here a special case – **real-valued** random variables.

A random variable ξ for a probability space (Ω, σ, P) is a mapping $\xi: \Omega \rightarrow \mathbb{R}$, satisfying

$$\{\omega: \xi(\omega) \leq r\} \in \sigma \quad \forall\, r \in \mathbb{R}$$

(always holds for power sets)

Note: elementary events are **not numbers** – they are elements of a general set Ω

Random variables are in contrast numbers, i.e. they can be summed up, subtracted, squared etc.

Distributions

Cummulative distribution function of a random variable ξ :

$$F_\xi(r) = P(\{\omega : \xi(\omega) \le r\})$$

Probability distribution of a **discrete** random variable $\xi : \Omega \to \mathbb{Z}$:

$$p_\xi(r) = P(\{\omega : \xi(\omega) = r\})$$

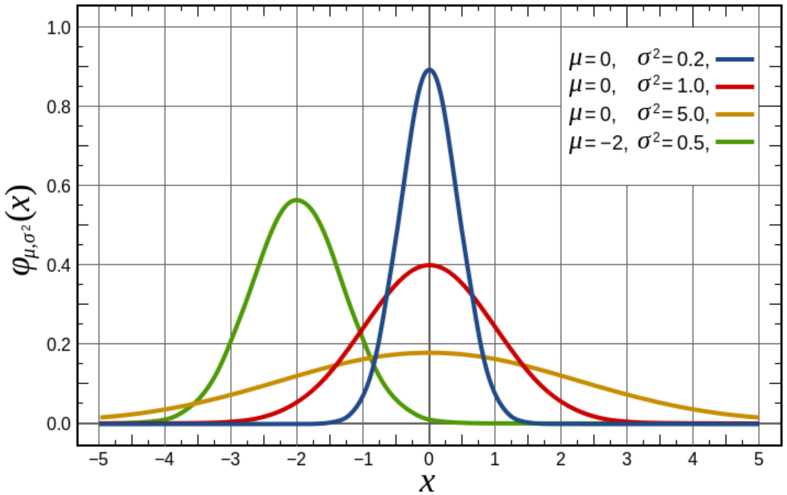
Probability density of a **continuous** random variable $\xi : \Omega \to \mathbb{R}$:

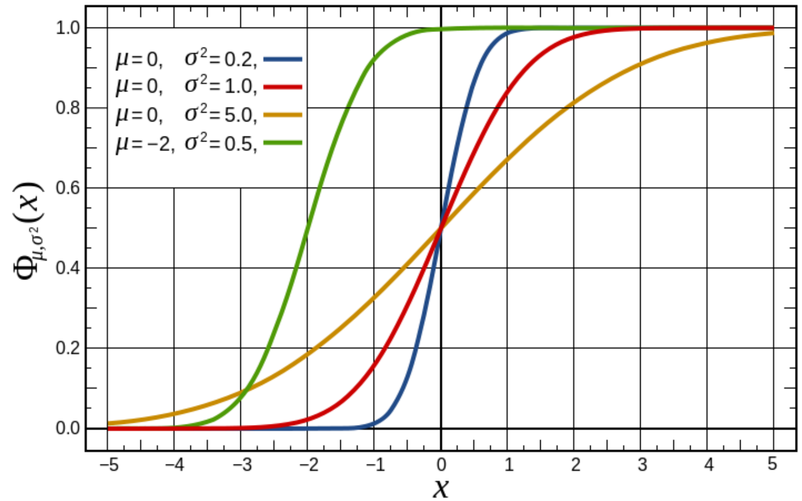
$$p_\xi(r) = \frac{\partial F_\xi(r)}{\partial r}$$

Distributions

Why is it necessary to do it so complex (through the cummulative distribution function)?

Example – a Gaussian





Probability of any particular real value is zero → a „direct“ definition of a „probability distribution“ is senseless ☹

It is indeed possible through the cummulative distribution function.

Mean

A mean (expectation, average ...) of a random variable ξ is

$$\mathbb{E}_P(\xi) = \sum_{\omega\in\Omega} P(\omega)\cdot\xi(\omega) = \sum_r \sum_{\omega:\xi(\omega)=r} P(\omega)\cdot r = \sum_r p_\xi(r)\cdot r$$

Arithmetic mean is a special case:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{n} x_i = \sum_r p_\xi(r)\cdot r$$

with

$$x \equiv r \quad \text{and} \quad p_\xi(r) = \frac{1}{N}$$

(uniform probability distribution).

Mean

The probability of an event $A \in \Omega$ can be expressed as the mean value of a corresponding „indicator“-variable

$$P(A) = \sum_{\omega \in A} P(\omega) = \sum_{\omega \in \Omega} P(\omega) \cdot \xi(\omega)$$

with

$$\xi(\omega) = \begin{cases} 1 & \text{if} \quad \omega \in A \\ 0 & \text{otherwise} \end{cases}$$

Often, the set of elementary events can be associated with a random variable (just enumerate all $\omega \in \Omega$).

Then one can speak about a “probability distribution over Ω“ (instead of the probability measure).

Example 1 – numbers of a die

The set of elementary events: $\Omega = \{a, b, c, d, e, f\}$

Probability measure: $P(\{a\}) = \frac{1}{6}, P(\{c, f\}) = \frac{1}{3}$ ...

Random variable (number of a die): $\xi(a) = 1, \xi(b) = 2 \dots \xi(f) = 6$

Cummulative distribution: $F_\xi(3) = \frac{1}{2}, F_\xi(4.5) = \frac{2}{3}$ ...

Probability distribution: $p_\xi(1) = p_\xi(2) \dots p_\xi(6) = \frac{1}{6}$

Mean value: $\mathbb{E}_P(\xi) = 3.5$

Another random variable (squared number of a die)

$$\xi'(a) = 1, \xi'(b) = 4 \ \dots \ \xi'(f) = 36$$

Mean value: $\mathbb{E}_P(\xi) = 15\frac{1}{6}$

Note: $\mathbb{E}_P(\xi') \neq \mathbb{E}_P^2(\xi)$

Example 2 – two independent dice numbers

The set of elementary events (6x6 faces):

$$\Omega = \{a, b, c, d, e, f\} \times \{a, b, c, d, e, f\}$$

Probability measure: $P(\{ab\}) = \frac{1}{36}, P(\{cd, fa\}) = \frac{1}{18}$ ...

Two random variables:

1) The number of the first die: $\xi_1(ab) = 1, \xi_1(ac) = 1, \xi_1(ef) = 5$ ...

2) The number of the second die: $\xi_2(ab) = 2, \xi_2(ac) = 3, \xi_2(ef) = 6$ ...

Probability distributions:

$$p_{\xi_1}(1) = p_{\xi_1}(2) = \cdots = p_{\xi_1}(6) = \frac{1}{6}$$

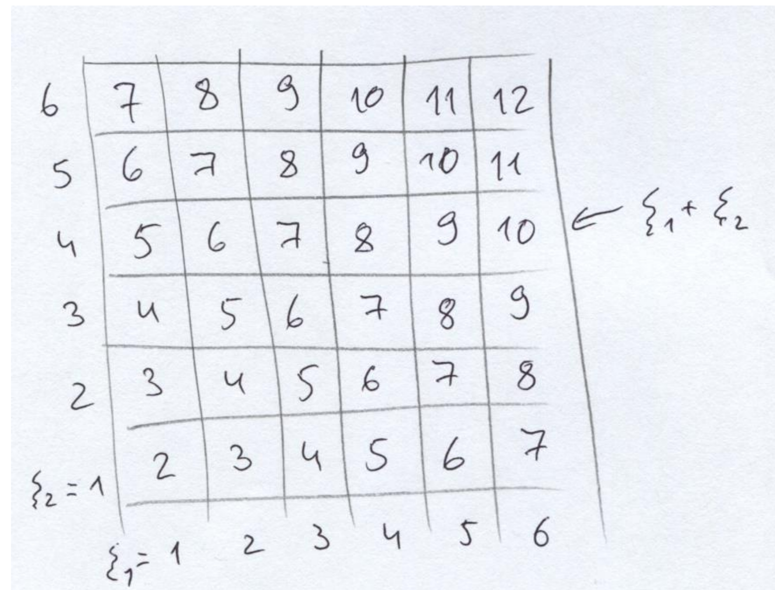
$$p_{\xi_2}(1) = p_{\xi_2}(2) = \cdots = p_{\xi_2}(6) = \frac{1}{6}$$

Example 2 – two independent dice numbers

Consider the new random variable: $\xi = \xi_1 + \xi_2$

The probability distribution p_ξ is not uniform anymore ☺

$$p_\xi \propto (1,2,3,4,5,6,5,4,3,2,1)$$



Mean value is $\mathbb{E}_P(\xi) = 7$

In general for mean values:

$$\mathbb{E}_P(\xi_1 + \xi_2) = \sum_{\omega \in \Omega} P(\omega) \cdot (\xi_1(\omega) + \xi_2(\omega)) = \mathbb{E}_P(\xi_1) + \mathbb{E}_P(\xi_2)$$

Random variables of higher dimension

Analogously: Let $\xi\colon\Omega \to \mathbb{R}^n$ be a mapping ($n = 2$ for simplicity), with $\xi = (\xi_1, \xi_2)$, $\xi_1\colon\Omega \to \mathbb{R}$ and $\xi_2\colon\Omega \to \mathbb{R}$

Cummulative distribution function:

$$F_\xi(r,s) = P(\{\omega\colon\xi_1(\omega) \le r\} \cap \{\omega\colon\xi_2(\omega) \le s\})$$

Joint probability distribution (discrete):

$$p_{\xi=(\xi_1,\xi_2)}(r,s) = P(\{\omega\colon\xi_1(\omega) = r\} \cap \{\omega\colon\xi_2(\omega) = s\})$$

Joint probability density (continuous):

$$p_{\xi=(\xi_1,\xi_2)}(r,s) = \frac{\partial^2 F_\xi(r,s)}{\partial r\,\partial s}$$

Independence

Two events $A \in \sigma$ and $B \in \sigma$ are **independent**, if

$$P(A \cap B) = P(A) \cdot P(B)$$

Interesting: Events A and $\bar{B} = \Omega \setminus B$ are independent, if A and B are independent ☺

Two random variables are independent, if

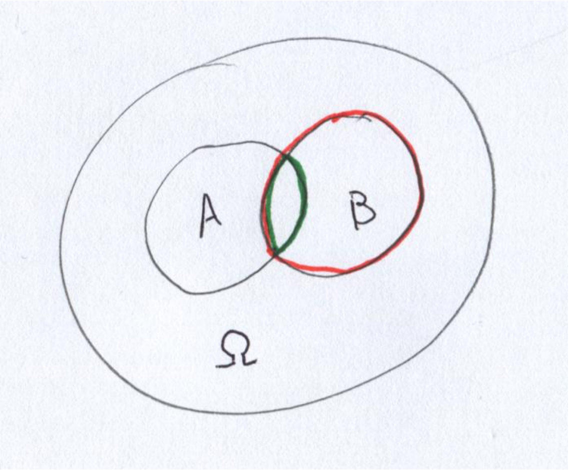
$$F_{\xi=(\xi_1,\xi_2)}(r,s) = F_{\xi_1}(r) \cdot F_{\xi_2}(s) \quad \forall\, r,s$$

It follows (example for continuous ξ):

$$p_\xi(r,s) = \frac{\partial^2 F_\xi(r,s)}{\partial r \partial s} = \frac{\partial F_{\xi_1}(r)}{\partial r} \cdot \frac{\partial F_{\xi_2}(s)}{\partial s} = p_{\xi_1}(r) \cdot p_{\xi_2}(s)$$

Conditional probability:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$



Independence (almost equivalent): A and B are independent, if

$$P(A \mid B) = P(A) \quad \text{and/or} \quad P(B \mid A) = P(B)$$

Bayes‘ Theorem (formula, rule)

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

Further definitions (for random variables)

Shorthand: $p(x, y) \equiv p_\xi(x, y)$

Marginal probability distribution:

$$p(x) = \sum_y p(x, y)$$

Conditional probability distribution:

$$p(x|y) = \frac{p(x, y)}{p(y)}$$

Note: $\sum_x p(x|y) = 1$

Independent probability distribution:

$$p(x, y) = p(x) \cdot p(y)$$

Example

Let the probability to be taken ill be

$$p(ill) = 0.02$$

Let the conditional probability to have a temperature in that case is

$$p(temp|ill) = 0.9$$

However, one may have a temperature without any illness, i.e.

$$p\left(temp\middle|\overline{ill}\right) = 0.05$$

What is the probability to be taken ill provided that one has a temperature?

Example

Bayes' rule:

$$p(ill|temp) = \frac{p(temp|ill) \cdot p(ill)}{p(temp)} =$$

(marginal probability in the denominator)

$$= \frac{p(temp|ill) \cdot p(ill)}{p(temp|ill) \cdot p(ill) + p\left(temp\middle|\overline{ill}\right) \cdot p(\overline{ill})} =$$

$$= \frac{0.9 \cdot 0.02}{0.9 \cdot 0.02 + 0.05 \cdot 0.98} \approx 0.27$$

– not so high as expected ☺, the reason – very low **prior** probability to be taken ill

Further topics

The model

Let two random variables be given:

- The first one is typically discrete (i.e. $k \in K$) and is called “class”
- The second one is often continuous ($x \in X$) and is called “observation”

Let the joint probability distribution $p(x, k)$ be “given”.

As k is discrete it is often specified by $p(x, k) = p(k) \cdot p(x|k)$

The recognition task: given x, estimate k.

Usual problems (questions):

- How to estimate k from x ?
- The joint probability is not always explicitly specified.
- The set K is sometimes huge.

Further topics

The learning task:

Often (almost always) the probability distribution is known up to free parameters. How to choose them (learn from examples)?

Next classes:

1. Recognition, Bayessian Decision Theory

2. Probabilistic learning, Maximum-Likelihood principle

3. Discriminative models, recognition and learning

…