Probability Theory

Dr. Joschka Boedecker
AG Maschinelles Lernen
Albert-Ludwigs-Universität Freiburg

jboedeck@informatik.uni-freiburg.de

Acknowledgement
Slides courtesy of Martin Riedmiller

Probabilities

probabilistic statements subsume different effects due to:

- convenience: declaring all conditions, exceptions, assumptions would be too complicated.
  Example: "I will be in lecture if I go to bed early enough the day before and I do not become ill and my car does not have a breakdown and ..."
  or simply: I will be in lecture with probability of 0.87
- lack of information: relevant information is missing for a precise statement.
  Example: weather forcasting
- intrinsic randomness: non-deterministic processes.
  Example: appearance of photons in a physical process

Probabilities (cont.)

- intuitively, probabilities give the expected relative frequency of an event
- mathematically, probabilities are defined by axioms (Kolmogorov axioms). We assume a set of possible outcomes Ω. An event A is a subset of Ω
  - the probability of an event A, $P(A)$ is a welldefined non-negative number: $P(A) \geq 0$
  - the certain event Ω has probability 1: $P(\Omega) = 1$
  - for two disjoint events A and B: $P(A \cup B) = P(A) + P(B)$

  P is called probability distribution
- important conclusions (can be derived from the above axioms):
  $P(\emptyset) = 0$
  $P(\neg A) = 1 - P(A)$
  if $A \subseteq B$ follows $P(A) \leq P(B)$
  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Probabilities (cont.)

- example: rolling the dice $\Omega = \{1, 2, 3, 4, 5, 6\}$
  Probability distribution (optimal dice):
  $P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = \frac{1}{6}$
  probabilities of events, e.g.:
  $P(\{1\}) = \frac{1}{6}$
  $P(\{1, 2\}) = P(\{1\}) + P(\{2\}) = \frac{1}{3}$
  $P(\{1, 2\} \cup \{2, 3\}) = \frac{1}{2}$ Probability distribution (manipulated dice):
  $P(1) = P(2) = P(3) = 0.13, P(4) = P(5) = 0.17, P(6) = 0.27$
- typically, the actual probability distribution is not known in advance, it has to be estimated

Joint events

- for pairs of events A, B, the joint probability expresses the probability of both events occuring at same time: $P(A, B)$
  example:
  $P($"Bayern München is losing", "Werder Bremen is winning"$) = 0.3$
- Definition: for two events the conditional probability of $A|B$ is defined as the probability of event A if we consider only cases in which event B occurs. In formulas:

$$P(A|B) = \frac{P(A, B)}{P(B)}, P(B) \neq 0$$

- with the above, we also have

$$P(A, B) = P(A|B)P(B) = P(B|A)P(A)$$

- example: $P($"caries"$|$"toothaches"$) = 0.8$
  $P($"toothaches"$|$"caries"$) = 0.3$

Joint events (cont.)

- a contigency table makes clear the relationship between joint probabilities and conditional probabilities:

	B	$\neg B$	
A	$P(A, B)$	$P(A, \neg B)$	$P(A)$
$\neg A$	$P(\neg A, B)$	$P(\neg A, \neg B)$	$P(\neg A)$
	$P(B)$	$P(\neg B)$	

marginals
joint prob

with $P(A) = P(A, B) + P(A, \neg B)$,
$P(\neg A) = P(\neg A, B) + P(\neg A, \neg B)$,
$P(B) = P(A, B) + P(\neg A, B)$,
$P(\neg B) = P(A, \neg B) + P(\neg A, \neg B)$

conditional probability = joint probability / marginal probability

Joint events (Example)

- example of a contigency table: cars and drivers

	red	*blue*	*other*	
male	0.05	0.15	0.35	0.55
female	0.2	0.05	0.2	0.45
	0.25	0.2	0.55	1

marginals

joint prob

e.g: I observed a blue car. How likely is the driver female?
How to express that in probabilistic terms?
$P('female'|'blue') = \frac{P('female','blue')}{P('blue')}$
How to access these values?
$P('female',' blue')$: from table
$P('blue') = P('blue',' male') + P('blue', female') = 0.2$ ('Marginalisation')
Therefore, $P('female'|'blue') = \frac{0.05}{0.2} = 0.25$
$\Rightarrow$ joint probabilty table allows to answer arbitrary questions about domain.

Marginalisation

- Let $B_1, ...B_n$ disjoint events with $\cup_i B_i = \Omega$. Then
  $P(A) = \sum_i P(A, B_i)$
  This process is called marginalisation.

Productrule and chainrule

- from definition of conditional probability:

$$P(A, B) = P(A|B)P(B) = P(B|A)P(A)$$

- repeated application: chainrule:

$$\begin{aligned} & P(A_1, \ldots, A_n) = P(A_n, \ldots, A_1) \\ = \quad & P(A_n|A_{n-1}, \ldots, A_1)\, P(A_{n-1}, \ldots, A_1) \\ = \quad & P(A_n|A_{n-1}, \ldots, A_1)\, P(A_{n-1}|A_{n-2}, \ldots, A_1)\, P(A_{n-2}, \ldots, A_1) \\ = \quad & \ldots \\ = \quad & \Pi_{i=1}^{n} P(A_i|A_1, \ldots, A_{i-1}) \end{aligned}$$

Conditional Probabilities

- conditionals:
  Example: if someone is taking a shower, he gets wet (by causality)
  $P(\text{"wet"} \,|\, \text{"taking a shower"}) = 1$
  while:
  $P(\text{"taking a shower"} \,|\, \text{"wet"}) = 0.4$
  because a person also gets wet if it is raining
- causality and conditionals:
  causality typically causes conditional probabilities close to 1:
  $P(\text{"wet"} \,|\, \text{"taking a shower"}) = 1$, e.g.
  $P(\text{"score a goal"} \,|\, \text{"shoot strong"}) = 0.92$ ('vague causality': if you shoot strong, you very likely score a goal').
  Offers the possibility to express vagueness in reasoning.
  you cannot conclude causality from large conditional probabilities:
  $P(\text{"being rich"} \,|\, \text{"owning an airplane"}) \approx 1$
  but: owning an airplane is not the reason for being rich

Bayes rule

- from the definition of conditional distributions:

$$P(A|B)P(B) = P(A, B) = P(B|A)P(A)$$

Hence:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

is known as Bayes rule.

- example:

$$P(\text{“taking a shower”}\,|\,\text{“wet”}) = P(\text{“wet”}\,|\,\text{“taking a shower”})\frac{P(\text{“taking a shower”})}{P(\text{“wet”})}$$

$$P(\text{reason}|\text{observation}) = P(\text{observation}|\text{reason})\frac{P(\text{reason})}{P(\text{observation})}$$

Bayes rule (cont)

- often this is useful in diagnosis situations, since $P(\text{observation}|\text{reason})$ might be easily determined.
- often delivers suprising results

Bayes rule - Example

- if patient has meningitis, then very often a stiff neck is observed
  $P(S|M) = 0.8$ (can be easily determined by counting)
- observation: 'I have a stiff neck! Do I have meningitis?' (is it reasonable to be afraid?)
  $P(M|S) = ?$
- we need to now: $P(M) = 0.0001$ (one of 10000 people has meningitis) and $P(S) = 0.1$ (one out of 10 people has a stiff neck).
- then:
  $$P(M|S) = \frac{P(S|M)P(M)}{P(S)} = \frac{0.8 \times 0.0001}{0.1} = 0.0008$$
- Keep cool. Not very likely

Independence

- two events A and B are called independent, if

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

- independence means: we cannot make conclusions about A if we know B and vice versa. Follows: $P(A|B) = P(A)$, $P(B|A) = P(B)$
- example of independent events: roll-outs of two dices
- example of dependent events: $A =$'car is blue', $B =$'driver is male'
  $\rightarrow$ (from example)
  $P('blue')\,P('male') = 0.2 \cdot 0.55 = 0.11 \neq P('blue','male') = 0.15$

Random variables

- random variables describe the outcome of a random experiment in terms of a (real) number
- a random experiment is a experiment that can (in principle) be repeated several times under the same conditions
- discrete and continuous random variables
- probability distributions for discrete random variables can be represented in tables:

  Example: random variable X (rolling a dice):

X	1	2	3	4	5	6
$P(X)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

- probability distributions for continuous random variables need another form of representation

Continuous random variables

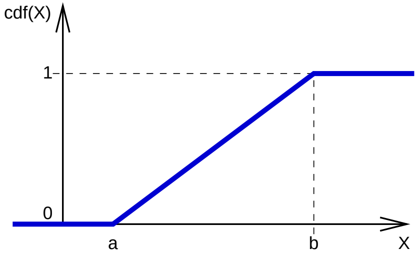
- problem: infinitely many outcomes
- considering intervals instead of single real numbers: $P(a < X \leq b)$
- cumulative distribution functions (cdf):
  A function $F : \mathbb{R} \rightarrow [0, 1]$ is called cumulative distribution function of a random variable X if for all $c \in \mathbb{R}$ hold:

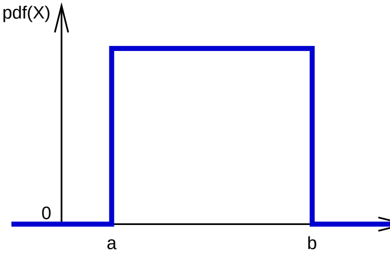
  $$P(X \leq c) = F(c)$$

- Knowing F, we can calculate $P(a < X \leq b)$ for all intervals from a to b
- F is monotonically increasing, $\lim_{x\rightarrow-\infty} F(x) = 0$, $\lim_{x\rightarrow\infty} F(x) = 1$
- if exists, the derivative of F is called a probability density function (pdf). It yields large values in the areas of large probability and small values in the areas with small probability. But: the value of a pdf cannot be interpreted as a probability!

Continuous random variables (cont.)

- example: a continuous random variable that can take any value between *a* and *b* and does not prefer any value over another one (uniform distribution):



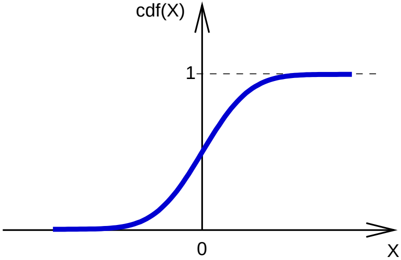


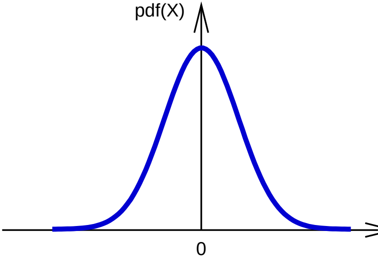
Gaussian distribution

- the Gaussian/Normal distribution is a very important probability distribution. Its pdf is:

$$pdf(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

$\mu \in \mathbb{R}$ and $\sigma^2 > 0$ are parameters of the distribution.
The cdf exists but cannot be expressed in a simple form
μ controls the position of the distribution, σ^2 the spread of the distribution





Statistical inference

- determining the probability distribution of a random variable (estimation)
- collecting outcome of repeated random experiments (data sample)
- adapt a generic probability distribution to the data. example:
  - Bernoulli-distribution (possible outcomes: 1 or 0) with success parameter p (=probability of outcome '1')
  - Gaussian distribution with parameters μ and σ^2
  - uniform distribution with parameters a and b
- maximum-likelihood approach:

$$\underset{\text{parameters}}{\textit{maximize}}\ P(\text{data sample}|\text{distribution})$$

Statistical inference (cont.)

- maximum likelihood with Bernoulli-distribution:
- assume: coin toss with a twisted coin. How likely is it to observe head?
- repeat several experiments, to get a sample of observations, e.g.: 'head', 'head', 'number', 'head', 'number', 'head', 'head', 'head', 'number', 'number', ...
  You observe k times 'head' and n times 'number' Probabilisitic model: 'head' occurs with (unknown) probability p, 'number' with probability $1-p$
- maximize the likelihood, e.g. for the above sample:

$$\underset{p}{maximize}\ p \cdot p \cdot (1-p) \cdot p \cdot (1-p) \cdot p \cdot p \cdot p \cdot (1-p) \cdot (1-p) \cdot \dots = p^k (1-p)^n$$

Statistical inference (cont.)

$$\underset{p}{maximize}\ p \cdot p \cdot (1-p) \cdot p \cdot (1-p) \cdot p \cdot p \cdot p \cdot (1-p) \cdot (1-p) \cdot \dots = p^k (1-p)^n$$

Trick 1: Taking logarithm of function does not change position of minima
rules: $\log(a \cdot b) = \log(a) + \log(b), log(a^b) = b\, log(a)$

Trick 2: Minimizing -log() instead of maximizing log()

This yields:

$$\underset{p}{minimize}\ -\log(p^k (1-p)^n) = -k \log p - n \log(1-p)$$

calculating partial derivatives w.r.t p and zeroing: $p = \frac{k}{k+n}$
$\Rightarrow$ The relative frequency of observations is used as estimator for p

Statistical inference (cont.)

- maximum likelihood with Gaussian distribution:
- given: data sample $\{x^{(1)}, \dots, x^{(p)}\}$
- task: determine optimal values for μ and σ^2
  assume independence of the observed data:

  $$P(\text{data sample}|\text{distribution}) = P(x^{(1)}|\text{distribution}) \cdot \dots \cdot P(x^{(p)}|\text{distribution})$$

  replacing probability by density:

  $$P(\text{data sample}|\text{distribution}) \propto \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x^{(1)}-\mu)^2}{\sigma^2}} \cdot \dots \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x^{(p)}-\mu)^2}{\sigma^2}}$$

  performing log transformation:

  $$\sum_{i=1}^{p} \Big(\log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2}\frac{(x^{(i)}-\mu)^2}{\sigma^2} \Big)$$

Statistical inference (cont.)

- minimizing negative log likelihood instead of maximizing log likelihood:

$$\underset{\mu,\sigma^2}{minimize} \; -\sum_{i=1}^{p} \big(\log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2} \frac{(x^{(i)} - \mu)^2}{\sigma^2} \big)$$

- transforming into:

$$\underset{\mu,\sigma^2}{minimize} \; \frac{p}{2} \log(\sigma^2) + \frac{p}{2} \log(2\pi) + \frac{1}{\sigma^2} \big(\frac{1}{2} \sum_{i=1}^{p} (x^{(i)} - \mu)^2 \big) \underbrace{\big(\frac{1}{2} \sum_{i=1}^{p} (x^{(i)} - \mu)^2 \big)}_{\text{sq. error term}}$$

- observation: maximizing likelihood w.r.t. μ is equivalent to minimizing squared error term w.r.t. μ

Statistical inference (cont.)

- extension: regression case, μ depends on input pattern and some parameters
- given: pairs of input patterns and target values $(\vec{x}^{(1)}, d^{(1)}), \dots, (\vec{x}^{(p)}, d^{(p)})$, a parameterized function f depending on some parameters $\vec{w}$
- task: estimate $\vec{w}$ and σ^2 so that $d^{(i)} - f(\vec{x}^{(i)}; \vec{w})$ fits a Gaussian distribution in best way
- maximum likelihood principle:

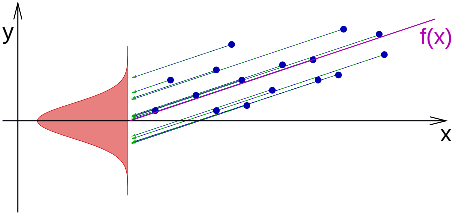
$$\underset{\vec{w},\sigma^2}{maximize} \ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(d^{(1)}-f(\vec{x}^{(1)};\vec{w}))^2}{\sigma^2}} \cdot \dots \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(d^{(p)}-f(\vec{x}^{(p)};\vec{w}))^2}{\sigma^2}}$$

Statistical inference (cont.)

- minimizing negative log likelihood:

$$\underset{\vec{w},\sigma^2}{minimize}\ \frac{p}{2}\log(\sigma^2)+\frac{p}{2}\log(2\pi)+\frac{1}{\sigma^2}\Big(\frac{1}{2}\sum_{i=1}^{p}(d^{(i)}-f(\vec{x}^{(i)};\vec{w}))^2\Big)\ \underbrace{\Big(\frac{1}{2}\sum_{i=1}^{p}(d^{(i)}-f(\vec{x}^{(i)};}_{\text{sq. error term}}$$

- f could be, e.g., a linear function or a multi layer perceptron



- minimizing the squared error term can be interpreted as maximizing the data likelihood $P(\text{trainingdata}|\text{modelparameters})$

Probability and machine learning

	machine learning	statistics
unsupervised learning	we want to create a model of observed patterns	estimating the probability distribution $P(\text{patterns})$
classification	guessing the class from an input pattern	estimating $P(\text{class}\|\text{input pattern})$
regression	predicting the output from input pattern	estimating $P(\text{output}\|\text{input pattern})$

- probabilities allow to precisely describe the relationships in a certain domain, e.g. distribution of the input data, distribution of outputs conditioned on inputs, ...
- ML principles like minimizing squared error can be interpreted in a stochastic sense

References

- Norbert Henze: Stochastik für Einsteiger
- Chris Bishop: Neural Networks for Pattern Recognition