Machine Learning

Probability Theory - A short recap

Prof. Matthias Hein

Machine Learning Group
Department of Mathematics and Computer Science
Saarland University, Saarbrücken, Germany

Lecture 2, 23.10.2013

Roadmap for today

- Discrete and continuous probability (... a bit of measure theory)
- Random variables
- Joint density, marginal density and the transformation law
- Expectation, variance, covariance, correlation, quantiles
- Independence, conditional probability, conditional independence
- Basic notions from statistics

Probability

History

- 17th century: development through the studies of games,
- 1933: axiomatic formulation of probability theory by Kolmogorov.

Interpretation of probability

- **Frequentist interpretation:**

$$\mathrm{P}(A) = \lim_{n \to \infty} \frac{n_A}{n},$$

  where n_A is the number of times A happens when we have done n trials.

- **Bayesian interpretation:**
  Probabilities are quantifying "rational belief". In the Bayesian framework we can answer:

  What is the probability that the sun rises tomorrow ?

Discrete probability

- **Atomic/elementary events:** set of possible outcomes which cannot be further divided $\Omega = \{\omega_1, \ldots, \omega_n\}$ (e.g. coin, $\Omega = \{H, T\}$).
- **Elementary event:** A singleton $\{\omega_r\}$ of Ω is called **elementary event**.
- **Events:** set of possible events the powerset 2^Ω (for the coin: $\{\emptyset, H, T, \{H, T\}\}$).
- **Probability measure:** a function $P : 2^\Omega \to [0, 1]$, such that
  - $\mathrm{P}(\emptyset) = 0$ and $\mathrm{P}(\Omega) = 1$,
  - $\sum_{\omega_i \in \Omega} \mathrm{P}(\omega_i) = 1$,
  - $A \in 2^\Omega \quad \Longrightarrow \quad \mathrm{P}(A) = \sum_{\omega_i \in A} \mathrm{P}(\omega_i)$.
- **Additivity rule of probabilities:**
  Let $A, B \in 2^\Omega$ then $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$.

Example: Binomial distribution

- An experiment with two outcomes $\mathcal{Y} = \{0, 1\}$ is called **Bernoulli** trial - determined by $p = \mathrm{P}(Y = 1)$.
- binomial distribution models n repeated Bernoulli trials where the outcomes are independent, e.g. a coin toss. Denote by X the number of times we observe $Y = 1$ (order does not matter).
- The **binomial probability measure** $\mathrm{Bin}(n, p)$ is then defined as,

$$\mathrm{P}\Big(X = k\Big) = \binom{n}{k} p^k (1 - p)^{n-k},$$

with the binomial coefficient $\binom{n}{k} = \frac{n!}{k!(n-k)!}$.

- Set of events: $\Omega = \{0, \ldots, n\}$ and one can check,

$$\mathrm{P}(\Omega) = \sum_{k=0}^{n} \mathrm{P}\Big(X = k\Big) = \sum_{k=0}^{n} \binom{n}{k} p^k (1 - p)^{n-k} = (1 - p + p)^n = 1.$$

General probability measures

Can we do the same if Ω is uncountable e.g. $\Omega = \mathbb{R}$?

One can show that if one tries to assign probabilities to **any** subset of 2^Ω one can get inconsistencies.

Banach-Tarski paradox

One can cut a ball of volume 1 into disjoint pieces and reassemble it so that one gets two balls of volume 1.

$\Longrightarrow$ it makes no sense to assign to every set a volume.

$\Longrightarrow$ we have to replace the powerset 2^Ω with something smaller.

$\Longrightarrow$ leads to the definition of a σ-Algebra.

Problem is resolved by Kolmogorov by a rigorous definition of measure theory

Definition

Let 2^Ω be the **power set**, the set of all subsets of Ω, and $\mathcal{A} \subset 2^\Omega$. $\mathcal{A}$ is a **σ-algebra** if the following conditions hold:

1. $\emptyset \in \mathcal{A}$ and $\Omega \in \mathcal{A}$,
2. if $A \in \mathcal{A}$, then also the complement A^c is contained in A,
3. $\mathcal{A}$ is closed under **countable** unions and intersections, that is if $A_1, A_2, \ldots$ is a sequence of events in $\mathcal{A}$, then
   - $\cup_{i=1}^\infty A_i \in \mathcal{A}$
   - $\cap_{i=1}^\infty A_i \in \mathcal{A}$.

- The pair $(\Omega, \mathcal{A})$ is called a **measure space**.
- All sets in the σ-algebra are called **measurable**.
- Probabilities will be only assigned to measurable sets.

Borel σ-Algebra

Definition

Let $C \subset 2^\Omega$. The σ-algebra generated by C is the smallest σ-algebra containing C. The **Borel σ-algebra** $\mathcal{B}$ in $\mathbb{R}^d$ is the σ-algebra generated by the open sets in $\mathbb{R}^d$.

Lebesgue Measure on $\mathbb{R}^d$

- We consider the Borel σ-Algebra $\mathcal{B}$ on $\mathbb{R}^d$
- The Lebesgue measure $\mu : \mathcal{B} \to \mathbb{R}_+$ is now just the usual measure of volume. For the one-dimensional case, we have

$$\mu(]a, b[) = b - a,$$

- A set $A \in \mathcal{B}$ has **measure zero** if $\mu(A) = 0$. Any countable set of points has measure zero.

Warning: The Lebesgue measure works acctually on its own σ-algebra but the difference is for our purposes neglectable.

Definition

A **probability measure** defined on a σ-algebra $\mathcal{A}$ of Ω is a function $\mathrm{P} : \mathcal{A} \to [0,1]$ that satisfies:

1. $\mathrm{P}(\Omega) = 1$,
2. For every countable sequence $(A_n)_{n \geq 1}$ of elements of $\mathcal{A}$, pairwise disjoint (that is $A_m \cap A_n = \emptyset$ whenever $m \neq n$), one has

$$\mathrm{P}\Big(\bigcup_{n=1}^{\infty} A_n\Big) = \sum_{n=1}^{\infty} \mathrm{P}(A_n).$$

The second property is called **countably additive**. $\mathrm{P}(A)$ is called the probability of A.

Probability on continuous spaces

In the case $\Omega = \mathbb{R}^d$ we will work with measures which have a density with respect to the **Lebesgue measure**.

Definition

Let $\mathcal{B}$ be the Borel σ-algebra in $\mathbb{R}^d$. A probability measure P on $(\mathbb{R}^d, \mathcal{B})$ has a **density** p if p is a non-negative (Borel measurable) function on $\mathbb{R}^d$ satisfying

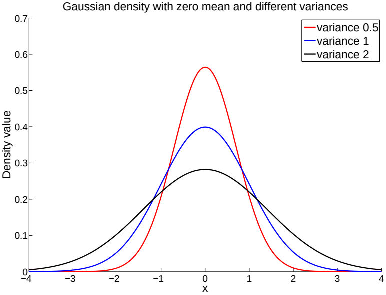
$$\mathrm{P}(A) = \int_A p(x)dx = \int_A p(x_1, \ldots, x_d)\, dx_1 \ldots dx_d\,,$$

for all $A \in \mathcal{B}$.

- This implies: $1 = \mathrm{P}(\mathbb{R}^d) = \int_{\mathbb{R}^d} p(x)dx$.
- In the following we always abbreviate, $dx = dx_1 \ldots dx_d$.
- **Not** all probability measures on $\mathbb{R}^d$ have a density.

The **Gaussian distribution** or normal distribution on $\mathbb{R}$ has two parameters μ (mean) and σ^2 (variance). The associated probability measure is denoted by $\mathcal{N}(\mu, \sigma^2)$. The density (with respect to the Lebesgue measure) is given as

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



Multivariate Gaussian $\mathcal{N}(\mu, C)$
Parameters: $\mu \in \mathbb{R}^d$ and $C \in \mathbb{R}^{d\times d}$ (positive-definite),

$$p(x) = \frac{1}{(2\pi)^{\frac{d}{2}}|\det C|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^T C^{-1}(x-\mu)}$$
$$= \frac{1}{(2\pi)^{\frac{d}{2}}|\det C|^{\frac{1}{2}}} e^{-\frac{1}{2}\left\langle x-\mu, C^{-1}(x-\mu)\right\rangle},$$

- A Gaussian density is uniquely determined by the mean and the covariance matrix C,
- Special case $C = \sigma^2 \mathbb{1}$,

$$p(x) = \frac{1}{(2\pi\sigma^2)^{\frac{d}{2}}} e^{-\frac{1}{2\sigma^2}\|x-\mu\|^2}.$$

Uniqueness of the density

f, f' agree **almost everywhere** (up to sets of measure zero) if

$$\int_{\mathbb{R}^d} \mathbb{1}_{f(x) \neq f'(x)} \, dx = 0.$$

Theorem

A non-negative (Borel measurable) function p on $\mathbb{R}^d$ is the density of a probability measure P on $\mathbb{R}^d$ if and only if

$$\int_{\mathbb{R}^d} p(x) dx = 1.$$

- *Any other positive Borel measurable function p' which agrees with p almost everywhere induces the same probability measure.*
- *Conversely, a probability measure on $\mathbb{R}^d$ determines its density (if it exists) up to sets of Lebesgue measure zero*

$$p(x) = \lim_{r \to 0} \frac{\mathrm{P}(B(x,r))}{\mathrm{vol}(B(x,r))}, \qquad \textit{almost everywhere.}$$

Distribution function:

- The **(cumulative) distribution** function of a probability measure P on $(\mathbb{R}, \mathcal{B})$ is the function

$$F(x) = \mathrm{P}\Big((-\infty, x]\Big).$$

If the distribution function F is sufficiently differentiable, then

$$p(x) = \frac{\partial F}{\partial x}\Big|_x.$$

- The distribution function of P on $(\mathbb{R}^d, \mathcal{B})$ is the function

$$F(x_1, \ldots, x_d) = \mathrm{P}\Big(\prod_{i=1}^{d}(-\infty, x_i]\Big).$$

If the distribution function F is sufficiently differentiable, then

$$p(x_1, \ldots, x_d) = \frac{\partial^d F}{\partial x_1 \ldots \partial x_d}\Big|_{x_1, \ldots, x_d}.$$

Quantile

Quantiles: Quantiles are only defined for distributions on $\mathbb{Z}$ and $\mathbb{R}$.

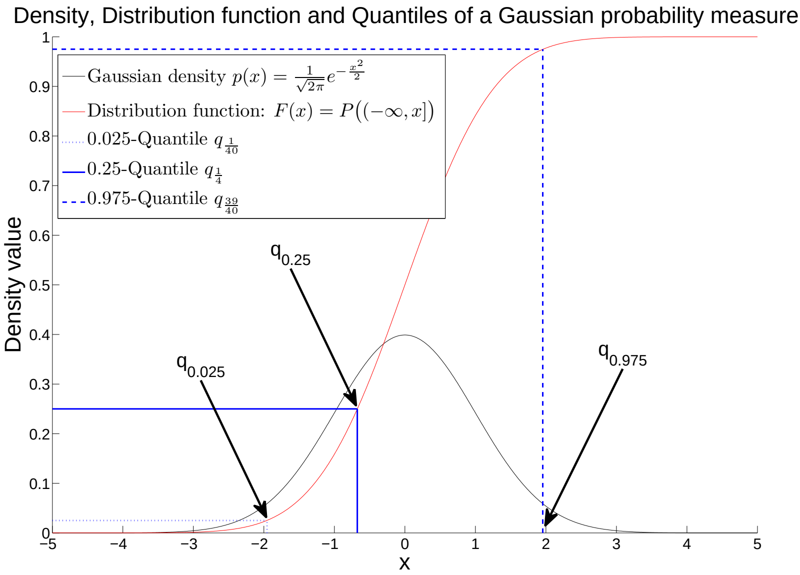
Definition

The **α-quantile** of a probability measure on $\mathbb{Z}$ or $\mathbb{R}$ is the real number q_α such that

$$F(q_\alpha) = \mathrm{P}(]-\infty, q_\alpha]) = \alpha.$$

The **median** is the $\frac{1}{2}$-quantile.

- Median and mean agree if the distributions are symmetric,
- The median is more robust to changes of the probability measure.



Density, Distribution function and Quantiles of a Gaussian probability measure
Gaussian density $p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$
Distribution function: $F(x) = P\big((-\infty, x]\big)$
0.025-Quantile $q_{\frac{1}{40}}$
0.25-Quantile $q_{\frac{1}{4}}$
0.975-Quantile $q_{\frac{39}{40}}$
$q_{0.25}$
$q_{0.025}$
$q_{0.975}$
Density value
x

Measurable functions and random variables

Definition

Let $(\Omega, \mathcal{A})$ and $(\Gamma, \mathcal{B})$ be spaces with a σ-Algebra (**measurable space**). A function $X : \Omega \to \Gamma$ is called **measurable** if

$$X^{-1}(B) \in \mathcal{A}, \text{ for all } B \in \mathcal{B}.$$

If in addition there is a probability measure defined on $(\Omega, \mathcal{A})$ then $X : \Omega \to \Gamma$ is called **random variable**. The **probability measure** or **law** P_X of X is defined for any $B \subset \Gamma$ as

$$\mathrm{P}_X(B) = \mathrm{P}(X^{-1}(B)) = \mathrm{P}(\{\omega \,|\, X(\omega) \in B\}).$$

- Random variables are denoted by capital letters e.g. X, Y, Z in order to distinguish them from normal variables x, y, z.
- If the target space Γ is $\mathbb{R}^d$ or $\mathbb{Z}$, we speak of $\mathbb{R}^d$-valued resp. $\mathbb{Z}$-valued random variables.
- A **random variable** X is a variable with random values and can be identified with the probability measure P_X (omission of sample space

Examples of random variables

- Coin toss: $\Omega = \{H, T\}$, $\mathrm{P}(H) = p$. Define $Z : \{H, T\} \to \mathbb{Z}$ by

$$Z = \begin{cases} 1 \text{ if } H, \\ 0 \text{ if } T. \end{cases}$$

  Z is a random variable with Bernoulli-distribution:

$$\mathrm{P}_Z(Z = 1) = \mathrm{P}(Z^{-1}(1)) = \mathrm{P}(H) = p, \text{ and similar } \mathrm{P}_Z(Z = 0) = 1 - p.$$

- Repeat the coin toss independently n times and denote by X the number of times we observe head (order does not matter). Let Ω be the set of all sequences of n variables with the alphabet $\{H, T\}$, then $|\Omega| = 2^n$. X is a random variable $X : \Omega \to \mathbb{Z}$ with distribution

$$\mathrm{P}_X\Big(X = k\Big) = \mathrm{P}(X^{-1}(k)) = \binom{n}{k} p^k (1 - p)^{n-k}.$$

  Example: Let $n = 3$, then $X^{-1}(2) = \{HHT, HTH, THH\}$.

Joint density and marginals

Theorem

Let $X = (X_1, X_2)$ be a $\mathbb{R}^2$-valued random variable with density p_X on $\mathbb{R}^2$. Then the densities p_{X_1} of X_1 and p_{X_2} of X_2 are given as

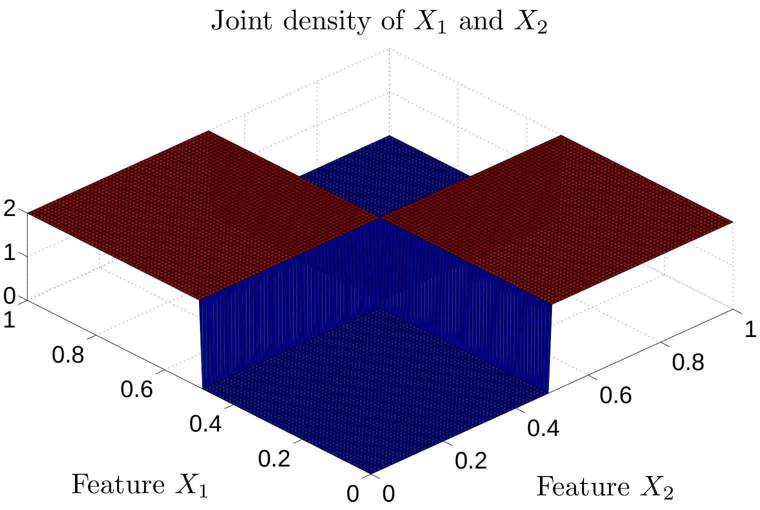
$$p_{X_1}(x_1) = \int_{\mathbb{R}} p_X(x_1, x_2) dx_2, \qquad p_{X_2}(x_2) = \int_{\mathbb{R}} p_X(x_1, x_2) dx_1.$$

- The law of $X = (X_1, X_2)$ is called the **joint measure** of X_1 and X_2 with density $p_X(x_1, x_2)$,
- p_{X_1} and p_{X_2} are called **marginal densities** of X and are associated to the probability measures of X_1 respectively X_2.

$\Longrightarrow$ the joint measure can in general not be reconstructed from the knowledge of the marginal densities (only if X_1 and X_2 are independent).

$\Longrightarrow$ the concept can be directly generalized to random variables $X = (X_1, \ldots, X_d)$ taking values in $\mathbb{R}^d$.

Joint Measure and marginals II



What are the marginal densities of X_1 and X_2 ?

What is the density of a function of a random variable X ?

Theorem

Let $X = (X_1, \ldots, X_d)$ have joint density p_X. Let $g : \mathbb{R}^d \to \mathbb{R}^d$ be continuously differentiable and injective, with non-vanishing Jacobian. Then $Y = g(X)$ has density

$$p_Y(y) = \begin{cases} p_X(g^{-1}(y))\,|\det J_{g^{-1}}(y)| & \text{if } y \text{ is in the range of } g, \\ 0 & \text{otherwise.} \end{cases}$$

- The Jacobian J_g of a function $g : \mathbb{R}^d \to \mathbb{R}^d$ at x is the $d \times d$- matrix

$$J_g(x)_{ij} = \frac{\partial g_i}{\partial x_j}\Big|_x .$$

- This formula follows directly from the rule for changing coordinates for a multidimensional integral.
- result allows to generate samples from complicated densities from simple ones.

Example: samples from an exponential distribution

How to generate samples from an exponential distribution ?

$p_\lambda(y) = \lambda \exp(-\lambda\, y)$, for $y \geq 0$, $p_\lambda(y) = 0$ if $y < 0$ with $\lambda > 0$

- **available:** samples from uniform distribution (Matlab/C: `rand`) on $[0, 1]$.
- we need a function $g : [0, 1] \to \mathbb{R}$ (resp. g^{-1}) such that

$$p_\lambda(y) = \lambda \exp(-\lambda\, y) = p_X(g^{-1}(y)) \left| \frac{\partial g^{-1}}{\partial y} \right| = \left| \frac{\partial g^{-1}}{\partial y} \right|.$$

General case: complicated differential equation.

This case: $g^{-1}(y) = \exp(-\lambda\, y)$ fulfills that ! $\implies g(x) = -\frac{\log(x)}{\lambda}$

- X_i samples from the uniform distribution on $[0, 1]$,
- $Y_i = g(X_i) = -\frac{\log(X_i)}{\lambda}$ are samples from the exponential distribution.

Expectation and variance

Definition

The **expected value** or **expectation** $\mathbb{E}[X]$ of a $\mathbb{R}^d$-valued random variable X is defined as

$$(\mathbb{E}[X])_i = \int_{\mathbb{R}^d} x_i \; p(x) \, dx \; = \int_{\mathbb{R}^d} x_i \; p(x_1, \ldots, x_d) \, dx_1 \ldots dx_d,$$

and for a discrete random variable X taking values in $\mathbb{Z}$ it is defined as,

$$\mathbb{E}[X] = \sum_{n=-\infty}^{\infty} n \, \mathrm{P}(X = n).$$

The **variance** $\mathrm{Var}[X]$ (also $\sigma^2(X)$) of an $\mathbb{Z}$- or $\mathbb{R}$-valued random variable X is defined as

$$\mathrm{Var}[X] = \mathbb{E}[(X - \mathbb{E}X)^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.$$

The standard deviation of X is $\sigma(X) = \sqrt{\mathrm{Var}[X]}$.

Expectation of functions of random variables

We can also define the expectation of functions of random variables.

$$\mathbb{E}[f(X)] = \int_{\mathbb{R}^d} f(x)\, p(x)\, dx \;=\; \int_{\mathbb{R}^d} f(x_1, \ldots, x_d)\, p(x_1, \ldots, x_d)\, dx_1 \ldots, dx_d.$$

Probability via expectation

Let $\mathbb{1}_A$ be the indicator function of the set A, that is

$$\mathbb{1}_A(x) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{else} \end{cases}.$$

then the probability of any set can be written as an expectation,

$$\mathbb{E}[\mathbb{1}_A] = \mathrm{P}(A).$$

Definition

The **covariance** $\mathrm{Cov}(X,Y)$ of two $\mathbb{R}$-valued random variables X and Y is defined as,

$$\mathrm{Cov}(X,Y) = \mathbb{E}[(X-\mathbb{E}X)(Y-\mathbb{E}Y)] = \mathbb{E}[X\,Y] - \mathbb{E}[X]\,\mathbb{E}[Y].$$

The **correlation** $\mathrm{Corr}(X,Y)$ of two $\mathbb{R}$-valued random variables X and Y is then defined as,

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Cov}(X,X)\,\mathrm{Cov}(Y,Y)}} = \frac{\mathrm{Cov}(X,Y)}{\sigma(X)\sigma(Y)}.$$

The covariance matrix C of an $\mathbb{R}^d$-valued random variable X is given as $C_{ij} = \mathrm{Cov}(X_i, X_j)$ or in matrix form

$$C = \mathbb{E}[(X-\mathbb{E}[X])(X-\mathbb{E}[X])^T].$$

Properties of covariance and correlation:

- The expectation and variance have the following properties $\forall a, b \in \mathbb{R}$,

$$\begin{aligned}
\mathbb{E}[aX + bY] &= a\,\mathbb{E}[X] + b\mathbb{E}[Y], \\
\mathrm{Var}[aX + b] &= a^2\ \mathrm{Var}[X], \\
\mathrm{Var}[X + Y] &= \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\ \mathrm{Cov}(X, Y).
\end{aligned}$$

- One has: $-1 \leq \mathrm{Corr}(X, Y) \leq 1$.
- Correlation is a measure of **linear dependence**.
  If X and Y are linearly dependent, that is $Y = aX + b$ with $a, b \in \mathbb{R}$, then

$$\mathrm{Corr}(X, Y) = \mathrm{Corr}(X, aX + b) = \frac{a}{|a|} = \begin{cases} 1, & \text{if } a > 0, \\ 0, & \text{if } a = 0, \\ -1, & \text{if } a < 0. \end{cases}$$

  Thus linearly dependent random variables achieve maximal correlation.

Independence

Definition

- Two events $A, B \in \mathcal{A}$ are **independent** if
$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B).$$
- Let A, B be events and $\mathrm{P}(B) > 0$. The **conditional probability** of A given B is defined as
$$\mathrm{P}(A|B) = \frac{P(A \cap B)}{\mathrm{P}(B)}.$$

Theorem

Suppose $\mathrm{P}(B) > 0$*.*

- *A and B are independent if and only if $\mathrm{P}(A|B) = \mathrm{P}(A)$.*
- *The operation $A \to \mathrm{P}(A|B)$ from $\mathcal{A} \to [0,1]$ defines a new probability measure on $\mathcal{A}$, called the* ***conditional probability measure given*** *B.*

Note: generally $\mathrm{P}(A|B) \neq \mathrm{P}(B|A)$.

Definition

A collection of events (E_n) is called a **partition** of Ω if $E_n \in \mathcal{A}$ for each n, they are pairwise disjoint, $E_n \cap E_m = \emptyset$ for $m \neq n$, $\mathrm{P}(E_n) > 0$ for each n, and $\cup_n E_n = \Omega$.

Law of total probability

Theorem

Let $(E_n)_{n \geq 1}$ be a finite or countable partition of Ω. Then if $A \in \mathcal{A}$,

$$\mathrm{P}(A) = \sum_n \mathrm{P}(A|E_n)\mathrm{P}(E_n).$$

Theorem (Bayes theorem)

Let (E_n) be a finite or countable partition of Ω, and suppose $\mathrm{P}(A) > 0$. Then

$$\mathrm{P}(E_n|A) = \frac{\mathrm{P}(A|E_n)\mathrm{P}(E_n)}{\sum_m \mathrm{P}(A|E_m)\mathrm{P}(E_m)}.$$

- We frequently use the Bayes theorem as follows. Let A, B two events,

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

which basically follows from the definition of conditional probability.

Independence of random variables

The concept of independence:

- Two random variables X, Y are maximally dependent if $Y = f(X)$ with f one-to-one.
- Two random variables are independent if knowledge about one variable does not tell you anything about the other one (successive coin tosses).

Definition

Two random variables X, Y with values in the measurable spaces $(E, \mathcal{E})$ and $(F, \mathcal{F})$ are **independent** if

$$\mathrm{P}(X \in A, Y \in B) = \mathrm{P}(X \in A)\mathrm{P}(Y \in B), \quad \text{for all } A \in \mathcal{E},\ B \in \mathcal{F}.$$

- in the following we restrict ourselves to random variables with density or probabilities on discrete spaces.

Independence of random variables II

Proposition

Let X, Y be $\mathbb{R}$-valued random variables with joint-density $p_{X \times Y}$ and marginal densities p_X and p_Y, then X and Y are **independent** *if*

$$p_{X \times Y}(x, y) = p_X(x)\, p_Y(y), \quad \forall x, y \in \mathbb{R}.$$

The **conditional density** $p(x|Y = y)$ *of X given $Y = y$ is defined as,*

$$p(x|y) = \frac{p(x, y)}{p(y)}, \quad \forall\, y \text{ with } p(y) > 0.$$

Let X, Y be $\mathbb{Z}$-valued random variables. X and Y are **independent** *if,*

$$\mathrm{P}_{X \times Y}(X = i, Y = j) = \mathrm{P}_X(i)\, \mathrm{P}_Y(j), \quad \forall i, j \in \mathbb{Z}.$$

The **conditional probability** $\mathrm{P}(X = i|Y = j)$ *of X given $Y = j$ is,*

$$\mathrm{P}(X = i|Y = j) = \frac{\mathrm{P}_{X \times Y}(X = i, Y = j)}{\mathrm{P}(Y = j)}, \quad \forall\, j \text{ with } \mathrm{P}(Y = j) > 0.$$

Problems with continuous probabilities:
The conditional probability $\mathrm{P}(X = x | Y = y)$ of X given $Y = y$ is undefined since the event $Y = y$ has probability mass $\mathrm{P}(Y = y) = 0$.

Underlying argumentation:

$$\begin{aligned}
\mathrm{P}(X \in [x, x + \Delta x],\ Y \in [y, y + \Delta y]) &\approx\ p_{X \times Y}(x, y)\, \Delta x\, \Delta y \\
\mathrm{P}(Y \in [y, y + \Delta y]) &\approx\ p_Y(y)\, \Delta y. \\
\mathrm{P}(X \in [x, x + \Delta x] \,|\, Y \in [y, y + \Delta y]) &\approx\ \frac{p_{X \times Y}(x, y)}{p_Y(y)}\, \Delta x.
\end{aligned}$$

Dividing by Δx and taking the limit $\Delta x \to 0$ and $\Delta y \to 0$ yields the result.

Notation: From now on we discard the subscript at $p_X(x)$.

Conditional density function

- the conditional density function defines a density for $X|Y = y$ by varying the set $\{Y = y\}$
- extend the definition of the conditional density also to values y with $p(y) = 0$ by assigning an arbitrary value.

The **conditional probability density** of $X|Y = y$ is defined as

$$p(x|y) = \begin{cases} \frac{p(x,y)}{p(y)} & \text{if } p(y) > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Let X, Y be random variables on $\mathcal{X}, \mathcal{Y}$, then

$$\int_{\mathcal{X}} p(x|y)\, dx = 1, \qquad \int_{\mathcal{Y}} p(x|y)\, p(y)\, dy = p(x).$$

Conditional expectation

Definition

Let X, Y be two $\mathbb{R}$-valued random variables. The **conditional expectation** $\mathbb{E}[X|Y = y]$ of X given $Y = y$ is defined for y with $p(y) > 0$ as the quantity

$$\mathbb{E}[X|Y = y] = \int_{\mathbb{R}} x\, p(x|y)\, dx.$$

The **conditional expectation** $\mathbb{E}[X|Y]$ of X given Y is a random variable $h(Y)$ with values

$$h(y) = \mathbb{E}[X|Y = y].$$

Proposition

Important properties of the conditional expectation are:

- $\mathbb{E}[X|Y] = \mathbb{E}[X]$, *if* X *and* Y *are* **independent**,
- $\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$ *(sometimes called the* **"tower property"***)*,
- $\mathbb{E}[f(Y)g(X)|Y] = f(Y)\mathbb{E}[g(X)|Y]$.