Knowledge Discovery and Data Mining 1 (VO) (707.003)

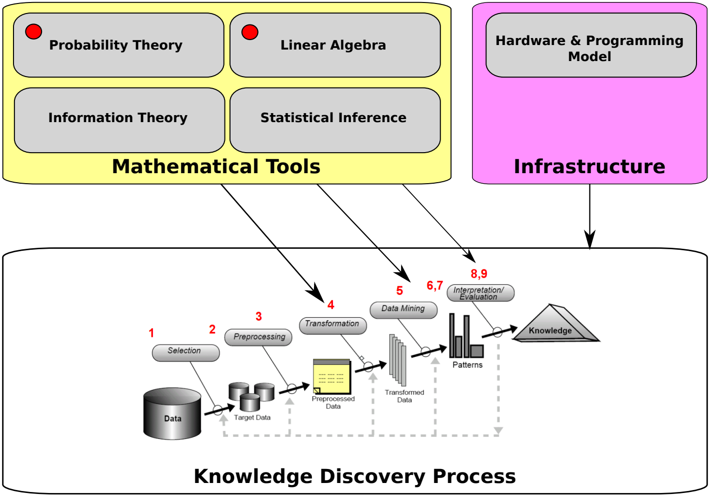
Review of Probability Theory

Denis Helic

KTI, TU Graz

Oct 9, 2014

Big picture: KDDM



Outline

1. Introduction
2. Conditional Probability and Independence
3. Random Variables
4. Discrete Random Variables
5. Continuous Random Variables

Random experiments

- In random experiments we can not predict the output in advance
- We can observe some "regularity" if we repeat the experiment a large number of times
- E.g. when tossing a coin we can not predict the result of a single toss
- If we toss many times we get an average of 50% of "heads" (fair coin)
- Probability theory is a mathematical theory which describes such phenomena

The state space

- It is the set of all possible outcomes of the experiment
- We denote the state space by Ω
- Coin toss: $\Omega = \{t, h\}$
- Two successive coin tosses: $\Omega = \{tt, th, ht, hh\}$
- Dice roll: $\Omega = \{1, 2, 3, 4, 5, 6\}$
- The lifetime of a light-bulb: $\Omega = \mathbb{R}_+$

The events

- An event is a property that either holds or does not hold after the experiment is done
- Mathematically, an event is a subset of Ω
- We denote the events by capital letters: $A, B, C, \ldots$
- E.g. rolling at least one heads in two successive coin tosses
- $A = \{th, ht, hh\}$

The events

Some basic properties of events

If A and B are two events, then:

- The contrary event of A is the complement set A^c
- The event "A or B" is the union $A \cup B$
- The event "A and B" is the intersection $A \cap B$

The events

Some basic properties of events

If A and B are two events, then:

- The sure event is Ω
- The impossible event is the empty set $\emptyset$
- An elementary (atomic) event is a subset of Ω containing a single element, e.g. $\{\omega\}$

The events

- We denote by $\mathcal{A}$ the family of all events
- Very often $\mathcal{A} = 2^{\Omega}$, the set of all subsets of Ω
- The family $\mathcal{A}$ should be closed under the operations from above
- If $A, B \in \mathcal{A}$, then we must have: $A^c \in \mathcal{A}$, $A \cap B \in \mathcal{A}$, $A \cup B \in \mathcal{A}$
- Also: $\Omega \in \mathcal{A}$ and $\emptyset \in \mathcal{A}$

The probability

- With each event we associate a number $P(A)$ called the probability of A
- $P(A)$ is between 0 and 1
- "Frequency" interpretation
- $P(A)$ is a limit of the "frequency" with which A is realized
- $P(A) =$ limit of $\frac{f(A)}{n}$ as n goes to positive infinity

The probability

Basic properties of the probabilities

(i) $0 \leq P(A) \leq 1$

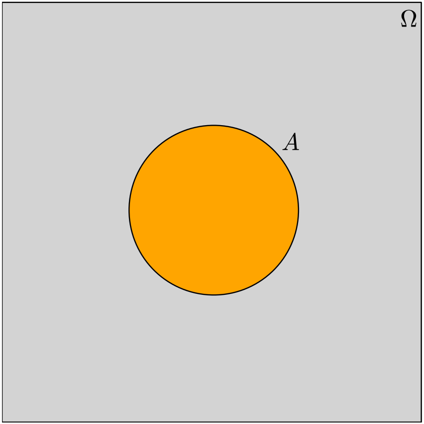
(ii) $P(\Omega) = 1$

(iii) $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \emptyset$

The probability

- The model is a triple $(\Omega, \mathcal{A}, P)$
- Ω is the state space
- $\mathcal{A}$ is the collection of all events
- $P(A)$ is the collection of all probabilities for $A \in \mathcal{A}$
- P is a mapping from $\mathcal{A}$ into $[0, 1]$ which satisfies at least properties (ii) and (iii) (*Kolmogorov axioms*)

Venn diagrams



Probability measure

- The probability $P(A)$ of the event A is the area of the set in the diagram
- The area of Ω is 1
- E.g. radius of the event A is $r = 0.2$
- $P(A) = 0.1257$

Properties of probability measures

Properties

If P is a probability measure on $(\Omega, \mathcal{A})$, then:

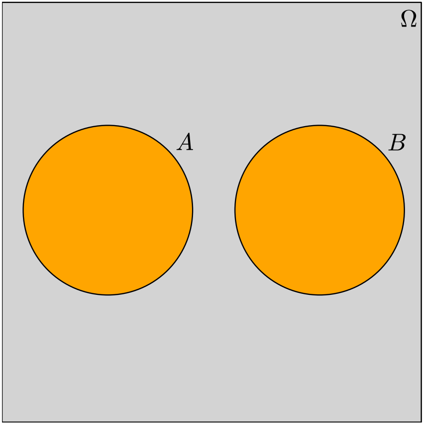
(i) $P(\emptyset) = 0$

(ii) For every finite sequence A_n of pairwise disjoint (whenever $i \neq j$, $A_i \cap A_j = \emptyset$) elements of $\mathcal{A}$ we have:

$$P(\cup_{n=1}^{m}) = \sum_{n=1}^{m} P(A_n)$$

- Property (ii) is called *additivity*

Additivity



Probability measure

- The probability of $P(A \cup B)$ of the event $A \cup B$ is the sum of areas of the sets A and B in the diagram
- $P(A) = 0.1257$
- $P(A) = 0.1257$
- $P(A \cup B) = 0.2514$

Properties of probability measures

Properties

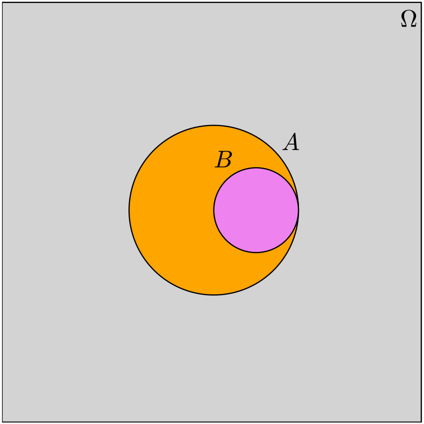
If P is a probability measure on $(\Omega, \mathcal{A})$, then:

(i) For $A, B \in \mathcal{A}$, $A \subset B \implies P(A) \leq P(B)$

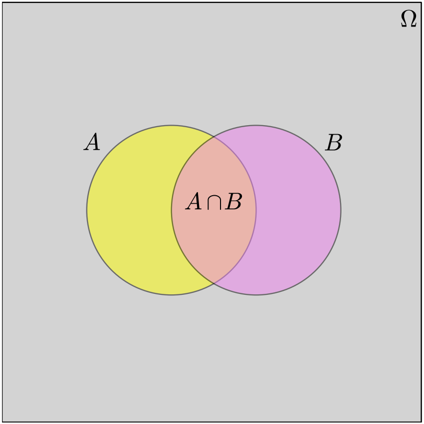
(ii) For $A, B \in \mathcal{A}$, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

(iii) For $A \in \mathcal{A}$, $P(A) = 1 - P(A^c)$

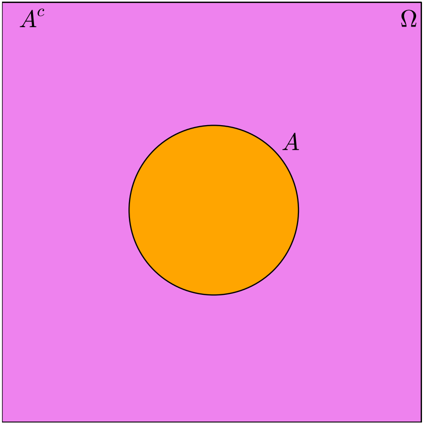
Subsets



Union



Complement



Conditional probability

- We always have the triple $(\Omega, \mathcal{A}, P)$
- Typically we suppress $(\Omega, \mathcal{A})$ and talk only about P
- Nevertheless, they are always present!
- Conditional probability and independence are crucial for application of probability theory in data mining!

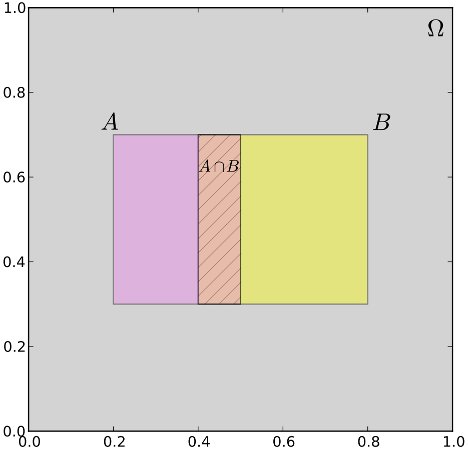
Conditional probability

Definition

If $P(B) > 0$ then we define the conditional probability of A given B:

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

Conditional probability



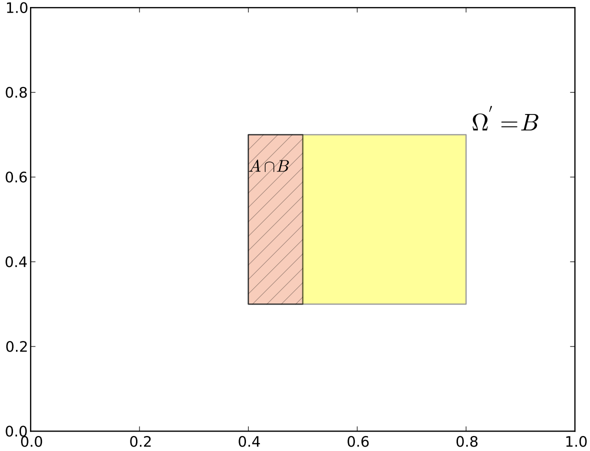
Conditional probability

- $P(B) = 0.16$
- $P(A) = 0.12$
- $P(A \cap B) = 0.04$
- $P(A|B) = \frac{P(A \cap B)}{P(B)} = 0.25$

Conditional probability

- One intuitive explanation is that B occurred first and then we ask what is the probability that now A occurs as well
- Time dimension
- Another intuitive explanation is that our knowledge about the world increased
- We have more information and know that B already occurred
- Technically, B restricts the state space (makes it smaller)

Conditional probability



Conditional probability: example

- We throw two dies
- Event $A = \{\text{snake eyes}\}$
- Event $B = \{\text{double}\}$
- $\Omega = \{(1,1),(1,2),\dots,(6,5),(6,6)\}$
- $A = \{(1,1)\}$, $B = \{(1,1),(2,2),\dots,(6,6)\}$

Conditional probability: example

- $P(A) = \frac{1}{36}$
- $P(B) = \sum_{i=1}^{6} \frac{1}{36}$ by final additivity and because events are pairwise disjoint $\implies P(B) = \frac{1}{6}$
- $A \cap B = A$
- $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{36}}{\frac{1}{6}} = \frac{1}{6}$
- $A = \{(1,1)\}$, $B = \{(1,1), (2,2), \ldots, (6,6)\}$

Conditional probability: example

- We have two boxes: red and blue
- Each box contains apples and oranges
- We first pick box at random
- Then we pick a fruit from that box again at random
- We are interested in the conditional probabilities of picking a specific fruit if a specific box was selected

Conditional probability: example

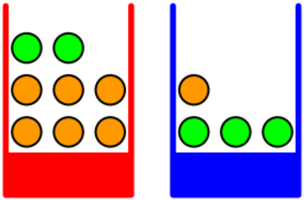


Figure: From the book "Pattern Recognition and Machine Learning" by Bishop

Conditional probability: example

- $P(A|B) = \frac{3}{4}$
- $P(O|B) = \frac{1}{4}$
- $P(A|R) = \frac{1}{4}$
- $P(O|R) = \frac{3}{4}$

Independence

Definition

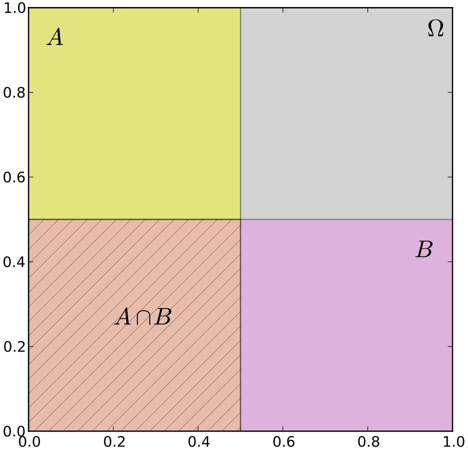
Event A and B are independent if:

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

Independence

- Events A and B are not related to each other
- We flip coin once: event A
- Second flip is the event B
- The outcome of the second flip is not dependent on the outcome of the first flip
- Intuitively, A and B are independent

Independence



Independence

- $P(A) = 0.5$
- $P(B) = 0.5$
- $P(A \cap B) = 0.25$
- $P(A)P(B) = 0.25$

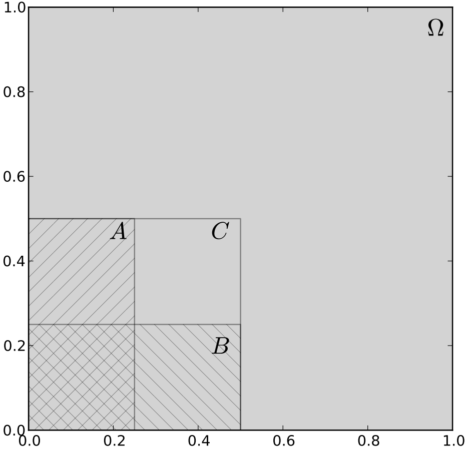
Conditional independence

Definition

Suppose $P(C) > 0$. Event A and B are conditionally independent given C if:

$$P(A \cap B|C) = P(A|C)P(B|C)$$

Conditional independence



Conditional independence

- $P(A) = \frac{1}{8}$, $P(B) = \frac{1}{8}$, $P(C) = \frac{1}{4}$
- $P(A|C) = \frac{1}{2}$, $P(B|C) = \frac{1}{2}$
- $P(A \cap B|C) = \frac{1}{4}$, $P(A|C)P(B|C) = \frac{1}{4}$
- $P(A \cap B) = \frac{1}{16}$, $P(A)P(B) = \frac{1}{64}$

Conditional independence

Remark

(i) Independence $\not\Rightarrow$ conditional independence

(ii) Conditional independence $\not\Rightarrow$ independence

Remark

Suppose $P(B) > 0$. Events A and B are independent if and only if $P(A|B) = P(A)$.

Independence: Example

- Pick a card at random from a deck of 52 cards
- $A = \{\text{the card is a heart}\}$, $B = \{\text{the card is Queen}\}$
- $P(i) = \frac{1}{52}$
- By additivity, $P(A) = \frac{13}{52}$, $P(B) = \frac{4}{52}$
- $P(A \cap B) = \frac{1}{52}$ (Queen heart), $P(A)P(B) = \frac{1}{52}$
- A and B are independent

Bayes rule

Remark

Suppose $P(A) > 0$ and $P(B) > 0$. Then,

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

Theorem

Suppose $P(A) > 0$ *and* $P(B) > 0$. *Then,*

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

Bayes rule

- One of the most important concepts in statistical inference
- Bayesian statistics
- You start with a probabilistic model with parameters B
- You observe data A and you are interested in the probability of parameters given the data

Chain & partition rule

Theorem

If $A_1, A_2, \dots, A_n$ *are events and* $P(A_1 \cap \dots \cap A_{n-1}) > 0$*, then*

$$P(A_1 \cap \dots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2)\dots P(A_n|A_1 \cap \dots \cap A_{n-1})$$

Definition

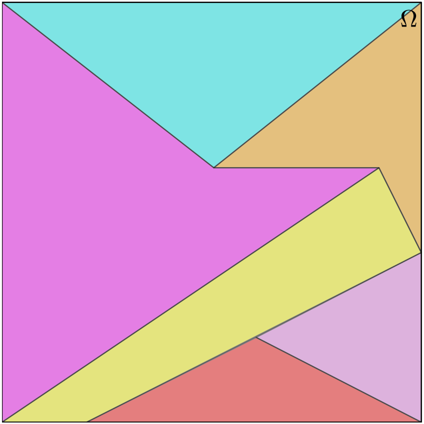
A partition of Ω is a finite or countable collection (B_n) if $B_n \in \mathcal{A}$ and:

(i) $P(B_n) > 0, \forall n$

(ii) $B_i \cap B_j = \emptyset, \forall i \neq j$ (pairwise disjoint)

(iii) $\cup_i B_i = \Omega$

Partition rule



Partition rule

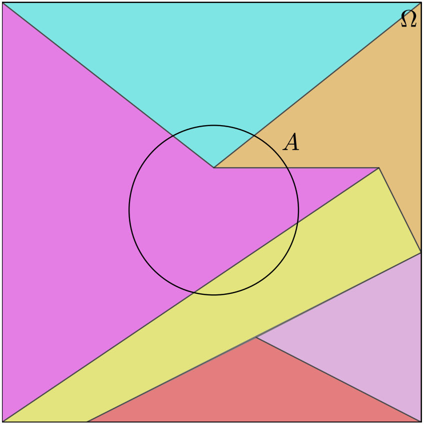
Theorem

Let B_n, $n \geq 1$ a finite or countable partition of Ω. Then if $A \in \mathcal{A}$:

$$P(A) = \sum_n P(A|B_n)P(B_n)$$

$$P(A) = \sum_n P(A \cap B_n)$$

Partition rule



Bayes rule revisited

Theorem

Let B_n, $n \geq 1$ *a finite or countable partition of* Ω *and suppose* $P(A) > 0$. *Then*

$$P(B_n|A) = \frac{P(A|B_n)P(B_n)}{\sum_m P(A|B_m)P(B_m)}$$

Bayes rule: example

Medical tests

Donated blood is screened for AIDS. Suppose that if the blood is HIV positive the test will be positive in 99% of cases. The test has also 5% false positive rating. In this age group one in ten thousand people are HIV positive.

Suppose that a person is screened as positive. What is the probability that this person has AIDS?

Bayes rule: example

- $P(A) =$

Bayes rule: example

- $P(A) = 0.0001$
- $P(A^c) =$

Bayes rule: example

- $P(A) = 0.0001$
- $P(A^c) = 0.9999$
- $P(P|A) =$

Bayes rule: example

- $P(A) = 0.0001$
- $P(A^c) = 0.9999$
- $P(P|A) = 0.99$
- $P(N|A) =$

Bayes rule: example

- $P(A) = 0.0001$
- $P(A^c) = 0.9999$
- $P(P|A) = 0.99$
- $P(N|A) = 0.01$
- $P(P|A^c) =$

Bayes rule: example

- $P(A) = 0.0001$
- $P(A^c) = 0.9999$
- $P(P|A) = 0.99$
- $P(N|A) = 0.01$
- $P(P|A^c) = 0.05$
- $P(N|A^c) =$

Bayes rule: example

- $P(A) = 0.0001$
- $P(A^c) = 0.9999$
- $P(P|A) = 0.99$
- $P(N|A) = 0.01$
- $P(P|A^c) = 0.05$
- $P(N|A^c) = 0.95$
- $P(A|P) = ?$

Bayes rule: example

$$
\begin{aligned}
P(A|P) &= \frac{P(P|A)P(A)}{P(P)} \\
&= \frac{P(P|A)P(A)}{P(P|A)P(A) + P(P|A^c)P(A^c)} \\
&= 0.00198
\end{aligned}
$$

Bayes rule: example

- The disease is so rare that the number of false positives outnumbers the people who have the disease
- E.g. what can we expect in a population of 1 million
- 100 will have the disease and 99 will be correctly diagnosed
- 999,900 will not have the disease but 49,995(!) will be falsely diagnosed
- If your test is positive the likelihood that you have the disease is $\frac{99}{99+49995} = 0.00198$

Bayes rule: example

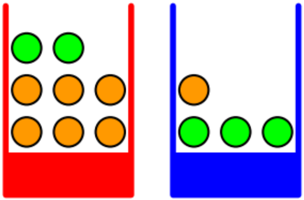


Figure: From the book "Pattern Recognition and Machine Learning" by Bishop

Bayes rule: example

- $P(R) = \frac{2}{5}$
- $P(B) = \frac{3}{5}$
- $P(A|B) = \frac{3}{4}$
- $P(O|B) = \frac{1}{4}$
- $P(A|R) = \frac{1}{4}$
- $P(O|R) = \frac{3}{4}$

Bayes rule: example

Fruits

We select orange. What is the probability the box was red? $P(R|O) = ?$

$$
\begin{aligned}
P(R|O) &= \frac{P(O|R)P(R)}{P(O)} \\
&= \frac{P(O|R)P(R)}{P(O|R)P(R) + P(O|B)P(B)} \\
&= \frac{2}{3}
\end{aligned}
$$

Random variables

- We use random variables to refer to "random quantities"
- E.g. we flip a coin 5 times and are interested in the number of heads
- E.g. the lifetime of the bulb
- E.g. the number of occurrences of a word in a text document

Random variables

Definition

Given a probability measure space $(\Omega, \mathcal{A}, P)$ a random variable is a **function** $X : \Omega \to \mathbb{R}$ such that $\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{A}, \forall x \in \mathbb{R}$

Remark

The condition is a technicality ensuring that the set is measurable. We just need to know that a random variable is a function that maps events onto numbers.

Random variables: example

- We transmit 10 data packets over a communication channel
- Events are of the form $(S, S, S, F, F, S, S, S, S, S)$
- State space Ω contains all possible 2^{10} sequences
- What is the probability that we will observe n successful transmissions
- We associate a r.v. with the number of successful transmissions
- The r.v. takes on the values $0, 1, \dots, 10$

Discrete random variables

Definition

A r.v. X is discrete if $X(\Omega)$ is countable (finite or countably infinite).

Remark

- E.g. $X(\Omega) = \{x_1, x_2, \dots\}$
- Ω is countable $\implies$ $X(\Omega)$ is countable and X is discrete
- These r.v. are called **discrete random variables**

Discrete random variables

- A discrete r.v. is characterized by its **probability mass function** (PMF)

$$p_X(x) = P(X = x)$$

$$p_X(x) = \sum_{\{\omega : X(\omega) = x\}} P(\{\omega\})$$

- We will shorten the notation and write $p(x)$

Discrete random variables: example

Die rolls

Let X be the sum of two die rolls.

- $\Omega = \{(1,1), (1,2), \dots, (6,5), (6,6)\}$
- $X(\Omega) = \{2, 3, \dots, 12\}$
- E.g. $X((1,2)) = 3, X((2,1)) = 3, X((3,5)) = 8, \dots$
- E.g. $p(3) =?$

$$p(3) = \sum_{\{\omega : X(\omega) = 3\}} P(\{\omega\}) = P((1,2)) + P((2,1)) = \frac{2}{36}$$

Discrete random variables: example

- E.g. $p(4) = ?$

$$p(4) = \sum_{\{\omega : X(\omega)=4\}} P(\{\omega\}) = P((1,3)) + P((3,1)) + P((2,2)) = \frac{3}{36}$$

$$p(x) = \frac{x-1}{36}, 2 \leq x \leq 7$$

Joint probability mass function

- We can introduce multiple r.v. on the same probability measure space $(\Omega, \mathcal{A}, P)$
- Let X and Y be r.v. on that space, then the probability that X and Y take on values x and y is given by:

$$P(\{\omega \in \Omega | X(\omega) = x, Y(\omega) = y\})$$

- Shortly, we write:

$$P(X = x, Y = y)$$

Joint probability mass function

- We define joint PMF as:

$$p_{XY}(x, y) = P(X = x, Y = y)$$

- Shortly, we write:

$$p(x, y)$$

Joint PMF: example

Text classification

Suppose we have a collection of documents that are either about China or Japan (document topics). We model a word occurrence as an event ω in a probability space. Let $\Omega = \{\text{all word occurrences}\}$. Let X be a r.v. that maps those occurrences to an enumeration of words and let Y be a r.v. that maps an occurrence to an enumeration of topics (either China or Japan). What is the joint PMF $p(x, y)$.

Document	Class
Chinese Beijing Chinese	China
Chinese Chinese Shanghai	China
Chinese Macao	China
Tokyo Japan Chinese	Japan

Joint PMF: example

Word	Class
Chinese	China
Beijing	China
Chinese	China
Chinese	China
Chinese	China
Shanghai	China
Chinese	China
Macao	China
Tokyo	Japan
Japan	Japan
Chinese	Japan

Joint PMF: example

Y \ X	Chinese	Beijing	Shanghai	Macao	Tokyo	Japan
China	5	1	1	1	0	0
Japan	1	0	0	0	1	1

Joint PMF: example

Y \ X	Chinese	Beijing	Shanghai	Macao	Tokyo	Japan
China	$^5/_{11}$	$^1/_{11}$	$^1/_{11}$	$^1/_{11}$	0	0
Japan	$^1/_{11}$	0	0	0	$^1/_{11}$	$^1/_{11}$

Marginal PMF

- $p(x)$ and $p(y)$ are called marginal probability mass functions

Remark

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

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

Marginal PMF: example

Y \ X	Chinese	Beijing	Shanghai	Macao	Tokyo	Japan	**p(y)**
China	$5/11$	$1/11$	$1/11$	$1/11$	0	0	$\mathbf{8/11}$
Japan	$1/11$	0	0	0	$1/11$	$1/11$	$\mathbf{3/11}$
p(x)	$\mathbf{6/11}$	$\mathbf{1/11}$	$\mathbf{1/11}$	$\mathbf{1/11}$	$\mathbf{1/11}$	$\mathbf{1/11}$	

Conditional PMF

Definition

A conditional probability mass function is defined as:

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

- Again, we can easily establish connection to the underlaying probability space and events

Conditional PMF: example

X / $p(x\|C)$	Chinese	Beijing	Shanghai	Macao	Tokyo	Japan
$p(x\|China)$	$5/8$	$1/8$	$1/8$	$1/8$	0	0

X / $p(x\|J)$	Chinese	Beijing	Shanghai	Macao	Tokyo	Japan
$p(x\|Japan)$	$1/3$	0	0	0	$1/3$	$1/3$

Independence

Definition

Two r.v. X and Y are independent if:

$$p(x, y) = p(x)p(y), \forall x, y \in \mathbb{R}$$

- All rules are equivalent to the rules for events, we just work with PMFs instead.

Joint PMF

- In general we can have many r.v. defined on the same probability measure space Ω
- $X_1, \ldots, X_n$
- We define the joint PMF as:

$$p(x_1, \ldots, x_n) = P(X_1 = x_1, \ldots, X_n = x_n)$$

Common discrete random variables

- Certain random variables commonly appear in nature and applications
- **Bernoulli** random variable
- **Binomial** random variable
- **Geometric** random variable
- **Poisson** random variable
- **Power-law** random variable

Common discrete random variables

- IPython Notebook examples
- http://kti.tugraz.at/staff/denis/courses/kddm1/pmf.ipynb

Command Line

ipython notebook --pylab=inline pmf.ipynb

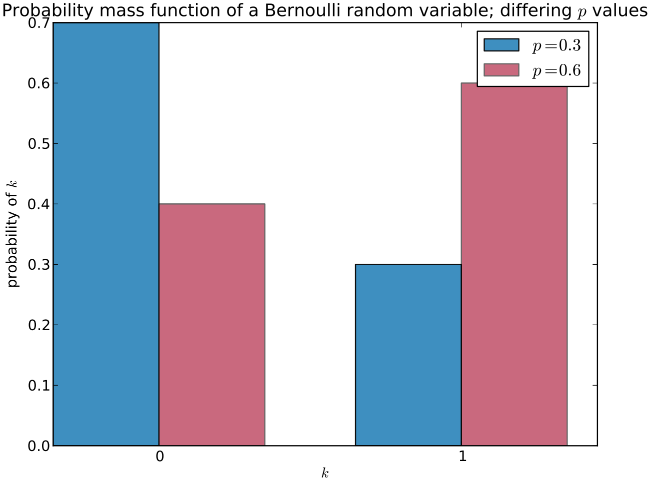
Bernoulli random variable

PMF

$$p(x) = \begin{cases} 1 - p \text{ if } x = 0 \\ p \text{ if } x = 1 \end{cases}$$

- Bernoulli r.v. with parameter p
- Models situations with two outcomes
- E.g. we start a task on a cluster node. Does the node fail ($X = 0$) or successfully finish the task ($X = 1$)?

Bernoulli random variable



Binomial random variable

- Suppose $X_1, \ldots, X_n$ are independent and identical Bernoulli r.v.
- The Binomial r.v. with parameters (p, n) is

$$Y = X_1 + \cdots + X_n$$

- Models the number of successes in n Bernoulli trials

Binomial random variable

Cluster nodes

We start tasks on n cluster nodes. How many nodes successfully finish their task?

- Probability of a single cluster configuration with k successes.

$$p(\omega) = (1 - p)^{n-k} p^k$$

Binomial random variable

- How many successful configurations exist?

$$p(k) = N(k)(1-p)^{n-k}p^k$$

$$N(k) = \binom{n}{k} = \frac{n!}{(n-k)!k!}$$

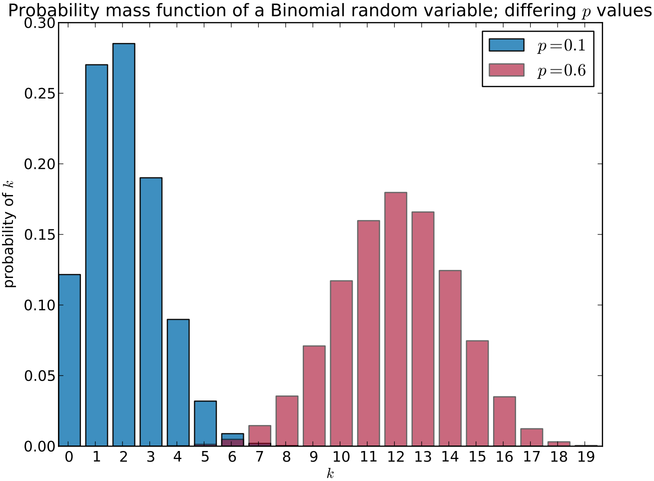
Binomial random variable

PMF

$$p(k) = \binom{n}{k}(1-p)^{n-k}p^k$$

- E.g. how many heads we get in n coin flips
- E.g. how many packets we transmit over n communication channels

Binomial random variable



Power-law (Zipf) random variable

- Power-law distribution is a very commonly occurring distribution
- Word occurences in natural language
- Friendships in a social network
- Links on the web
- PageRank, etc.

Power-law (Zipf) random variable

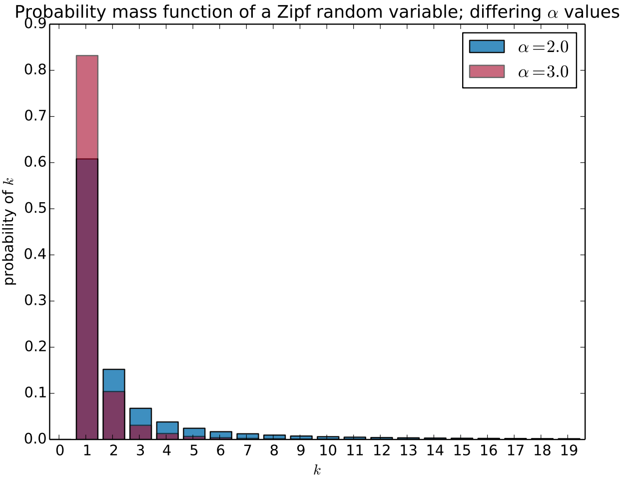
PMF

$$p(k) = \frac{k^{-\alpha}}{\zeta(\alpha)}$$

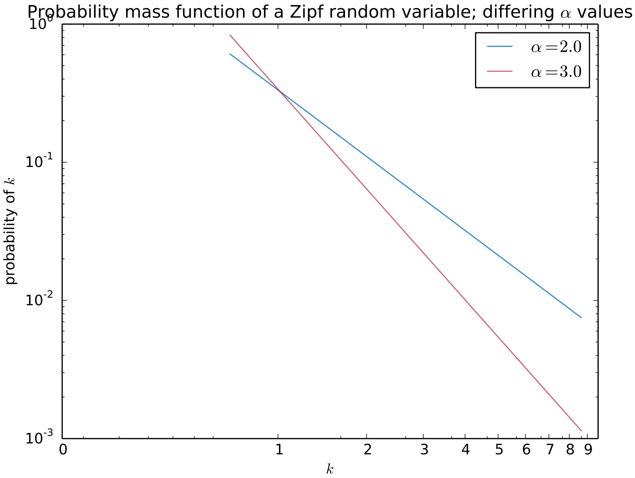
- $k \in \mathbb{N}$, $k \geq 1$, $\alpha > 1$
- $\zeta(\alpha)$ is the Riemann zeta function

$$\zeta(\alpha) = \sum_{k=1}^{\infty} k^{-\alpha}$$

Power-law (Zipf) random variable



Power-law (Zipf) random variable



Expectation

Definition

The expectation of a discrete r.v. X with PMF p is

$$E[X] = \sum_{x \in X(\Omega)} x p(x)$$

when this sum is "well-defined", otherwise the expectation does not exist.

Remark

(i) "Well-defined": it could be infinite, but it should not alternate between $-\infty$ and ∞

(ii) Expectation is the average value of a r.v.

Expectation: example

Gambling game

We play repeatedly a gambling game. Each time we play we either win 10€ or lose 10€. What are our average winnings?

- Let w_k be our winning for game k. Then the average winning in n games is:

$$W = \frac{w_1 + \cdots + w_n}{n}$$

Expectation: example

- Let n_W be the number of wins and n_L the number of losses. Then,

$$W = \frac{10n_W - 10n_L}{n} = 10\frac{n_W}{n} - 10\frac{n_L}{n}$$

- If we approximate $P(\{win\}) \approx \frac{n_W}{n}$ and $P(\{loss\}) \approx \frac{n_L}{n}$. Then,

$$W = 10P(\{win\}) - 10P(\{loss\}) = \sum_{x \in X(\Omega)} xp(x)$$

Linearity of expectation

Theorem

Suppose X and Y are discrete r.v. such that $E[X] < \infty$ and $E[Y] < \infty$. Then,

- $E[aX] = aE[X], \forall a \in \mathbb{R}$
- $E[X + Y] = E[X] + E[Y]$

Variance

Definition

The variance $\sigma^2(X)$, $var(X)$ of a discrete r.v. X is the expectation of the r.v. $(X - E[X])^2$

$$var(X) = E[(X - E[X])^2]$$

Remark

Variance indicates how close X typically is to $E[X]$

$$var(X) = E[X^2] - (E[X])^2$$

Covariance

Definition

The covariance $cov(X, Y)$ of two discrete r.v. X and Y is the expectation of the r.v. $(X - E[X])(Y - E[Y])$

$$cov(X, Y) = E[(X - E[X])(Y - E[Y])]$$

$$cov(X, Y) = E[XY] - E[X]E[Y]$$

Covariance

Remark

Covariance measures how much two r.v. change together, i.e. do X and Y tend to be small together, or is X large when Y is small (or vice versa), or do they change independently of each other

- If greater values of X correspond with greater values of Y, and same holds for small values then $cov(X, Y) > 0$
- In the opposite case $cov(X, Y) < 0$

Covariance

- If X and Y are independent then $cov(X, Y) = 0$
- This follows because $E[XY] = E[X]E[Y]$ in the case of independence
- Is the opposite true?
- If $cov(X, Y) = 0$ are X and Y independent?

Covariance

Covariance and independence

Suppose X takes on values $\{-2, -1, 1, 2\}$ with equal probability. Suppose $Y = X^2$.

$$cov(X, Y) = E[XY] - E[X]E[Y] = E[X^3] = 0$$

- Clearly X and Y are not independent
- They are linearly independent, but not independent in general

Covariance

Remark

(i) Independence of X and Y $\implies$ $cov(X, Y) = 0$

(ii) $cov(X, Y) = 0$ $\not\Longrightarrow$ Independence of X and Y

Continuous random variables

Definition

A r.v. X is continuous (general) if $X(\Omega)$ is uncountable.

- A general description is given by $P(X \in (-\infty, x])$
- We define the cumulative distribution function (CDF) of X as:

$$F_X(x) = P(X \in (-\infty, x])$$

- We will write shortly $F(x)$ and $P(X \leq x)$

Cumulative distribution function (CDF)

Definition

A cumulative distribution function (CDF) is a function $F : \mathbb{R} \to \mathbb{R}$ such that

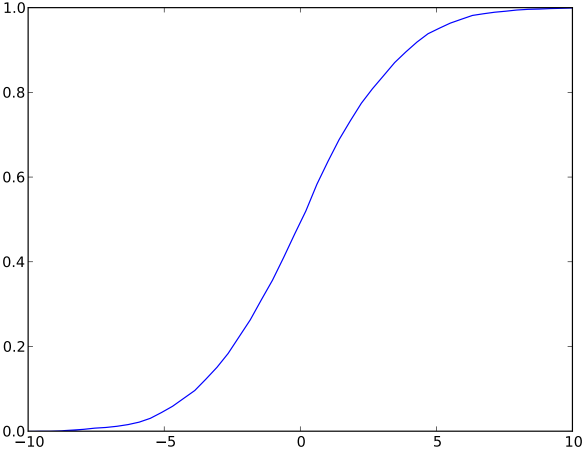
(i) F is non-decreasing ($x \leq y \implies F(x) \leq F(y)$)

(ii) F is right-continuous ($\lim\limits_{x \searrow a} = F(a)$)

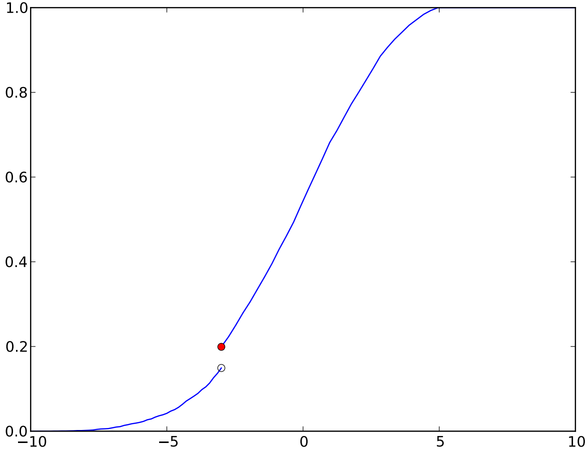
(iii) $\lim\limits_{x \to \infty} F(x) = 1$

(iv) $\lim\limits_{x \to -\infty} F(x) = 0$

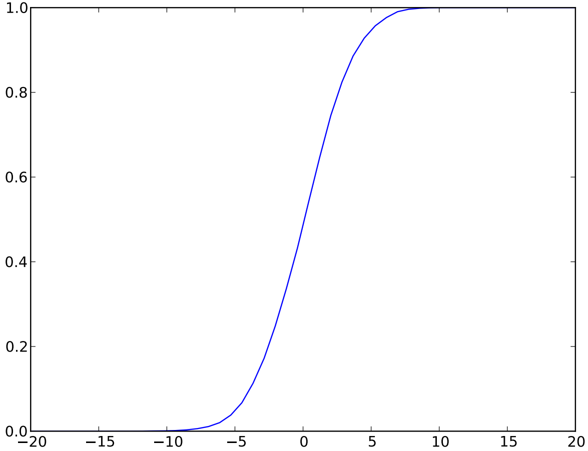
CDF: non-decreasing



CDF: right-continuous



CDF: infinity limits



Probability density function (PDF)

- Suppose that CDF is continuous and differentiable

Definition

A probability density function (PDF) of a r.v. X is defined as:

$$f(x) = \frac{dF(x)}{dx}$$

Definition

A joint PDF of two r.v. X and Y is defined as:

$$f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y}$$

Probability density function (PDF)

Definition

Suppose X and Y are two r.v. defined on the same probability measure space. Conditional PDF of X given Y is defined as:

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

Expectation

Definition

Expectation $E[X]$ of a r.v. X with a PDF $f(x)$ is defined as:

$$E[X] = \int_{-\infty}^{\infty} xp(x)dx$$

- In a similar way we define variance and covariance for a joint PDF

Common continuous random variables

- Certain random variables commonly appear in nature and applications
- **Exponential** random variable
- **Normal (Gaussian)** random variable
- **Power-law** random variable

Common continuous random variables

- IPython Notebook examples
- http://kti.tugraz.at/staff/denis/courses/kddm1/pdf.ipynb

Command Line

```
ipython notebook --pylab=inline pdf.ipynb
```

Normal (Gaussian) random variable

- Normal distribution is a very commonly occurring distribution
- Continuous approximate to the binomial for large n and p not too close to neither 0 nor 1
- Continuous approximate to the Poisson dist. with $n\lambda$ large
- Measurement errors
- Student grades
- Measures of sizes of living organisms

Normal random variable

PDF

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

- μ is the mean (expectation) and σ^2 is the variance of a normally distributed r.v.

CDF

$$F(x) = \Phi(\frac{x-\mu}{\sigma}), \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{x'^2}{2}} dx'$$

Normal random variable