Introduction to

Information Theory and Its Applications

Radim Bělohlávek

Dept. Computer Science
Palacký University, Olomouc
radim.belohlavek@acm.org

Outline

This is a preliminary version of a text providing introduction to Information Theory.

Information Theory: What and Why

- **information**: one of key terms in our society:
  “INFORMATION IS MONEY”
  and popular keywords like “information/knowledge society”

- information is a **central topic in computer science**:
  - storage and retrieval of information: data structures, algorithms,
  - search engines: Google etc.
  - information security (secure communication, banking, national security, etc.)
  - knowledge representation (artificial intelligence)
  - communication of information: technical means of passing information (mobile technology, etc.)

- various **meaning** of information
  - information as something new, possibly valuable
  - information vs. data (sometimes confused)

- **theories** of information

  - various attempts to formalize the notion of information (semantic information, information logics)
  - the most important so far is **classical information theory** and its extensions

- basic features of classical information theory
  - invented by Claude Shannon (American engineer and mathematician, with IBM), seminal paper Shannon C. E.: "A Mathematical Theory of Communicatio", Bell System Technical Journal, 27, pp. 379-423 & 623-656, July & October, 1948.
  -
  - Wikipedia: "Information theory is the mathematical theory of data communication and storage, generally considered to have been founded in 1948 by Claude E. Shannon. The central paradigm of classic information theory is the engineering problem of the transmission of information over a noisy channel. The most fundamental results of this theory are Shannon's source coding theorem, which establishes that on average the number of bits needed to represent the result of an uncertain event is given by the entropy; and Shannon's noisy-channel coding theorem, which states that reliable communication is possible over noisy channels

provided that the rate of communication is below a certain threshold called the channel capacity. The channel capacity is achieved with appropriate encoding and decoding systems.

Information theory is closely associated with a collection of pure and applied disciplines that have been carried out under a variety of banners in different parts of the world over the past half century or more: adaptive systems, anticipatory systems, artificial intelligence, complex systems, complexity science, cybernetics, informatics, machine learning, along with systems sciences of many descriptions. Information theory is a broad and deep mathematical theory, with equally broad and deep applications, chief among them coding theory.

Coding theory is concerned with finding explicit methods, called codes, of increasing the efficiency and fidelity of data communication over a noisy channel up near the limit that Shannon proved is all but possible. These codes can be roughly subdivided into data compression and error-correction codes. It took many years to find the good codes whose existence Shannon proved. A third class of codes are cryptographic ciphers; concepts from coding theory and information theory are much used in cryptography and cryptanalysis; see the article on deciban for an interesting historical application.

Information theory is also used in information retrieval, intelligence gathering, gambling, statistics, and even musical composition.''

- information understood as decrease of uncertainty:
  - ∗ suppose we an measure our uncertainty about a state X of some system of our interest; $U(X)$ . . . our uncertainty
  - ∗ suppose some action A leads from state X_1 to state X_2
  - ∗ information (carried) by A is defined by $I(A) = U(X_1) - U(X_2)$ (**informati of uncertainty**)
  - ∗ example: tossing a dice, X . . . possible outcomes, A . . . message saying that a result is a number ≥ 5
- applications in communication theory (bounds to channels capacity), coding and cryptography, decision making, learning
- has extensions (uncertainty-based information, based on measures of uncertainty different from probabilistic)

part I

PRELIMINARIES FROM PROBABILITY

The concept of a probability space

Introduction Probability theory studies situations (experiments, observations) the outcomes of which are uncertain. The set of all possible outcomes is called a sample space and is denoted by Ω, the outcomes $\omega \in \Omega$ are called elementary events. For instance, the experiment might by throwing a die. There are six outcomes, $\omega_1, \dots, \omega_6$, where ω_i is "the result is i". In probability theory, we also deal with events which are sets of elementary events. For instance, event $A = \{\omega_2, \omega_4, \omega_6\}$ might be described by "the result is an even number". We are interested only in a certain collection $\mathcal{B}$ of events, i.e. $\mathcal{B} \subseteq 2^\Omega$. Then, a function P, called a probability measure, assigns to each event $A \in \mathcal{B}$ a number $P(A)$, called a probability of event A. We require $0 \leq P(A) \leq 1$; if $P(A) = 1$, A is called a certain event; if $P(A) = 0$, A is called an impossible event. For instance, with a fair die we might have $P(\{\omega_2, \omega_4, \omega_6\}) = 1/2$, $P(\{\omega_2\}) = 1/6$, etc. The triplet $\langle \Omega, \mathcal{B}, P \rangle$ is called a probability space (and it is precisely defined).

The abstract notion of a probability space is due to A. N. Kolmogorov (1903–1987, Soviet mathematician, one the most influential mathematicians).

Recall:

- σ-**algebra** (σ-fields, Borel field) of sets over a set Ω is a subset $\mathcal{B} \subseteq 2^\Omega$ which satisfies

  - $\Omega \in \mathcal{B}$,
  - if $A \in \mathcal{B}$ then $\overline{A} \in \mathcal{B}$ (closedness under complements),
  - if $A_i \in \mathcal{B}$ for $i \in \mathbf{N}$ then $\bigcup_{i\in\mathbf{N}} A_i \in \mathcal{B}$ (closedness under countable unions).

- If $\mathcal{B}$ is a σ-algebra over Ω, then if $A, B \in \mathcal{B}$ then $A \cup B \in \mathcal{B}$, $A \cap B \in \mathcal{B}$, $A - B \in \mathcal{B}$, $\emptyset \in \mathcal{B}$ (exercise).

- some useful facts:

- For any system $\mathcal{F} \subseteq 2^{\Omega}$ of subsets of Ω there exists a minimal σ-algebra $\mathcal{B}(\mathcal{F})$ over Ω containing $\mathcal{F}$. Proof: 2^{Ω} itself is a σ-algebra over Ω; an intersection of σ-algebras $\mathcal{B}_i$ over Ω is a σ-algebra over Ω; therefore, $\mathcal{B}(\mathcal{F}) = \bigcap\{\mathcal{B} \mid \mathcal{B}$ is a σ-algebra containing $\mathcal{F}\}$.

- facts from measure theory (not required)

Definition A **probability space** is a triplet $\langle \Omega, \mathcal{B}, P\rangle$ such that

- Ω is a non-empty set called **sample space** (výběrový prostor, prostor elementárních jevu); $\omega \in \Omega$ are called **elementary events**;

- $\mathcal{B}$ is a σ-algebra over Ω; $A \in \mathcal{B}$ are called **events**;

- P is a function, called a **probability measure**, assigning to each $A \in \mathcal{B}$ a real number $P(A) \geq 0$, such that
  - $P(\Omega) = 1$ (probability of a certain event is 1),
  - $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$ for any sequence $A_1, A_2, \dots$ of pairwise disjoint events, i.e. $A_i \cap A_j = \emptyset$ for $i \neq j$ (σ-additivity).

Lemma (properties of probability space) If $\langle \Omega, \mathcal{B}, P \rangle$ is a probability space, we have (for $A, B \in \mathcal{B}$)

- $0 \leq P(A) \leq 1$ for any $A \in \mathcal{B}$,
- $A \subseteq B$ implies $P(A) \leq P(B)$,
- $P(\emptyset) = 0$,
- $P(\overline{A}) = 1 - P(A)$,
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Proof Exercise.

Example (1) Let $\Omega = \{\omega_1, \dots, \omega_n\}$ be finite, $\mathcal{B} = 2^{\Omega}$. Then if $\langle \Omega, \mathcal{B}, P \rangle$ is a probability space, P is uniquely given by its restriction to $\{\omega_1\}, \dots, \{\omega_n\}$.

Indeed, due to σ-additivity, for $A = \{\omega_{i_1}, \ldots, \omega_{i_k}\}$ we have $P(A) = P(\{\omega_{i_1}\}) + \cdots + P(\{\omega_{i_k}\})$. In this case, we denote

$$P(\{\omega_i\}) \text{ also by } P(\omega_i) \text{ or } p_i.$$

This means that in this case, a probability measure is completely given by a vector $\langle p_1, \ldots, p_n \rangle$ of reals $p_i \geq 0$ such that $\sum_{i=1}^{n} p_i = 1$.

(2) The same holds true if Ω is countably infinite (but we have to deal with infinite sums; note that if $\sum_{i=1}^{\infty} p_i = 1$ then this sum does not depend on the order of $p_1, p_2, \ldots$, i.e. for any bijection $\sigma : \mathbf{N} \to \mathbf{N}$ we have $\sum_{i=1}^{\infty} p_{\sigma(i)} = 1$).

Spaces from Example (1) and (2) are called **discrete probability spaces**. For Ω finite or countable, a function $p : \Omega \to [0, 1]$ satisfying

$$\sum_{i=1}^{\infty} p(\omega_i) = 1$$

is called a **probability distribution**.

Therefore:

Theorem There exists a bijective correspondence between discrete probability spaces and probability distributions.

Examples of discrete probability spaces: see any textbook on probability (binomial, geometric, Poisson, . . .).

Basic notions related to probability spaces

We suppose $\langle \Omega, \mathcal{B}, P\rangle$ is a probability space and $A, B, \ldots \in \mathcal{B}$. Furthermore, we require all quantities to be defined.

- **Conditional probability.** $P(A) > 0$ we call $P(B|A)$ defined by

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

  the *conditional probability* that B occurs given that A occurs. The thus induced function $P(\cdot|A)$ is again a probability measure (verify!), called a conditional probability measure.
  Example:

- **Independence.** Events $A, B \subseteq \Omega$ are called *independent* if $P(A \cap B) = P(A) \cdot P(B)$. Clearly, if A and B are independent then $P(A|B) = P(A)$ (provided $P(B) > 0$) and $P(B|A) = P(B)$ (provided $P(B) > 0$).
  Example:

- **Law of complete probabilities.** Let $B_1, \ldots, B_n$ be pairwise disjoint events such that $P(B_i) > 0$ and $P(\bigcup_{i=1}^{n} B_i) = 1$. Then

  $$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i).$$

Example:

- **Bayes theorem.** We have

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

Indeed, observe that $P(A \cap B) = P(B)P(A|B) = P(A)P(B|A)$. If $B_1, \dots, B_n$ be pairwise disjoint events such that $P(B_i) > 0$ and $P(\bigcup_{i=1}^{n} B_i) = 1$, then

$$P(B_k|A) = \frac{P(A|B_k)P(B_k)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}.$$

Indeed, use the above form of Bayes theorem and the law of complete probabilities. Bayes rule is crucial in inductive reasoning. $B_1, \dots, B_n$ are called hypotheses, $P(B_k|A)$ is called posterior probability of B_k after occurrence of A, $P(B_k)$ is called prior probability of B_k.
Example: We have two coins, one fair (comes up heads and tails with probabilities $\frac{1}{2}$ and $\frac{1}{2}$) and one biased (comes up heads). Experiment: We chose a coin at random and flip the coin twice. What is the probability that the chosen coin is biased if it comes up heads twice? Solution: We can put

$$\Omega = \{\langle c, a, b\rangle \mid c \in \{f, b\}, a, b \in \{h, t\}\}.$$

Let A denote the event "biased coin was chosen", B denote the event "the coin comes up heads both times". We are interested in $P(A|B)$. By Bayes theorem,

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} = \frac{P(A)P(B|A)}{P(B|A)P(A) + P(B|\overline{A})P(\overline{A})} =$$
$$= \frac{1/2 \cdot 1}{1 \cdot 1/2 + 1/4 \cdot 1/2} = \frac{4}{5}.$$

- **Joint and marginal probability.** If $\Omega = X \times Y$ then a probability measure P on Ω is also called a *joint probability* (joint are $X = \{x_1, \dots, x_m\}$ and $Y = \{y_1, \dots, y_n\}$). Therefore, P is given by $p_{ij} = P(x_i, y_j)$. The functions $P_1(x) = \sum_{y \in Y} P(x, y)$ for $x \in X$ (or simply $P(x)$) and $P_2(y)$ for $y \in Y$ (or $P(y)$) defined analogously, are called *marginal probability distributions*. Furthermore, $P(x_i|y_j) = \frac{P(\{x_i\} \times \{y_j\})}{P(y_j)}$ is called the *conditional probability* of x_i given y_j (note that this is a special case of the above definition of a conditional probability $P(B|A)$: just take $A = \{\langle x, y_j \rangle \mid x \in X\}$ and $B = \{\langle x_i, y \rangle \mid y \in Y\}$). We say that x_i and y_j are independent if $P(x_i, y_j) = P(x_i)P(y_j)$ (again, this is a special case of the independence defined above).

Random variables, discrete and continuous

Definition A **random variable** on a probability space $\langle \Omega, \mathcal{B}, P\rangle$ is a function $X : \Omega \to \mathbf{R}$ (assigning to any elementary event ω its numeric characteristic $X(\omega) \in \mathbf{R}$ we are interested in) such that for each $x \in \mathbf{R}$ we have $\{\omega \mid X(\omega) \leq x\} \in \mathcal{B}$, i.e. $\{\omega \mid X(\omega) \leq x\}$ is an event of the corresponding measurable space. A tuple $\langle X_1, \dots, X_n\rangle$ of random variables on $\langle \Omega, \mathcal{B}, P\rangle$ is called a **random vector**.

Example (1) dice, elementary events are $x_1, \dots x_6$, $X(x_i) = i$;
(2) experiment with outputs x_i each of them characterized by a real value a_i (e.g. a_i is the weight of a randomly chosen ball x_i).

Notation $P(X \leq x)$ denotes $P(\{\omega \in \Omega \mid X(\omega) \leq x\})$, etc.

Basic notions and facts Let X be a random variable on $\langle \Omega, \mathcal{B}, P\rangle$.

- X induces a (cumulative) **distribution function** $F : \mathbf{R} \to [0, 1]$ of X by
$$F(x) = P(\{\omega \in \Omega \mid X(\omega) \leq x\}).$$
- Properties of F (see course on probability).
- X induces a probability space $\langle \mathbf{R}, \mathcal{B}_X, P_X\rangle$ where $\mathcal{B}_X$ is a Borel field on $\mathbf{R}$ and
$$P_X(A) = P(\{\omega \mid X(\omega) \in A\})$$

for each $A \in \mathcal{B}_X$.

- Two basic types of random variables:
  - **discrete random variable** if X takes at most countably many values in $\mathbf{R}$, i.e. $\{X(\omega) \mid \omega \in \Omega\}$ is finite or countably infinite. That is, there are at most countably many values $x_1, x_2, \ldots$ such that $P(X = x_i) > 0$ and we have $\sum_{i=1}^{\infty} P(X = x_i) = 1$. Function $p_i = p(x_i) = P(X = x_i)$ is called a probability distribution (density) of X
  - **continuous random variable** if $F(x) = \int_{-\infty}^{t} f(f)\mathrm{d}t$ for some non-negative real function f. Then f is called the probability density of X.

- **Expected value** of X.
  For X discrete: $E(X) = \sum_{x_i} x_i \cdot p_i$;
  Example: $n = 3$, $p_1 = 0.1$, $x_1 = 30$, $p_2 = 0.4$, $x_2 = 10$, $p_3 = 0.5$, $x_3 = 100$; $E(X) = 57$.
  For X continuous: $E(X) = \int_{-\infty}^{\infty} x f(x)\mathrm{d}x$.

- More generally: If $Y = h(X)$, i.e. $Y(\omega) = h(X(\omega))$ for some (Borel measurable) function $h : \mathbf{R} \to \mathbf{R}$, then Y is again a random variable on $\langle \Omega, \mathcal{B}, P \rangle$. Then for the expected value of $h(X)$ we have

$E(h(X)) = \sum_{x_i} h(x_i) \cdot p_i$ for discrete X, and
$E(h(X)) = \int_{-\infty}^{\infty} h(x) f(x) \mathrm{d}x$ for continuous X.
Check this for discrete case. (The point is that by definition, $E(h(X)) = \sum_{x_i} h(x_i) \cdot q_i$ where $q_i = P_{h(X)}(X = x_i)$ is the distribution of $h(X)$.)

- **Random vector, joint distribution, marginal distributions**
  - **Random vector** is a vector $\mathbf{X} = \langle X_1, \dots, X_n \rangle$ of random variables on $\langle \Omega, \mathcal{B}, P \rangle$.
  - The distribution function of $\mathbf{X}$ is a function $F : \mathbf{R}^n \to [0, 1]$ defined by

    $$F(x_1, \dots, x_n) = P(\{\omega \in \Omega \mid X_1(\omega) \le x_1, \dots, X_n(\omega) \le x_n\})$$

    is called a **joint distribution function** of $\mathbf{X}$.
  - $\mathbf{X}$ is discrete if there is a finite or countably infinite series of n-tuples $\langle x_1, \dots, x_n \rangle$ such that $P(X_1 = x_1, \dots X_n = x_n) > 0$ and $\sum_{\langle x_1, \dots, x_n \rangle} P(X_1 = x_1, \dots X_n = x_n) = 1$. For $n = 2$ we also write $p_{ij} = p(x_i, x_j) = p_{X_1 X_2}(x_i, x_j) = P(X_1 = x_i, X_2 = x_j)$.
  - **Marginal distribution** ($n = 2$, discrete $\mathbf{X} = \langle X_1, X_2 \rangle$) Marginal distribution for X_1 is a function

    $$p_{X_1}(x_i) = \sum_{y_j} p_{X_1 X_2}(x_i, y_j);$$

marginal distribution for X_2 is a function

$$p_{X_2}(y_j) = \sum_{x_i} p_{X_1 X_2}(x_i, y_j).$$

- **Independence of random variables** Random variables X and Y on $\langle \Omega, \mathcal{B}, P \rangle$ are called **independent** if for any Borel sets $S, T \subseteq \mathbf{R}$ we have that events $A = X^{-1}(S) = \{\omega \mid X(\omega) \in S\} \in \mathcal{B}$ and $B = Y^{-1}(T) = \{\omega \mid Y(\omega) \in T\} \in \mathcal{B}$ are independent events. One can prove that X and Y are independent iff for their distribution functions F_X and F_Y, and the joint distribution function F we have

  $$F(x_1, x_2) = F_X(x_1) \cdot F_Y(x_2)$$

  for any $x_1, x_2 \in \mathbf{R}$. This is true iff:
  $p_{XY}(x_i, y_j) = p_X(x_1) \cdot p_Y(y_j)$ for any x_i, y_j for discrete case, and
  $f(x, y) = f(x) \cdot f(y)$ for any x, y for continuous case.

- **Conditional distribution** Let $\langle X, Y \rangle$ be a two-dimensional discrete random vector with joint probability distribution $p_{ij} = P(X = x_i, Y = y_j)$. Then conditional distribution of X under condition $Y = y_j$ is denoted by $P(X = x_i | Y = y_j)$ (also by $p(x_i | y_j)$) and is defined by

  $$P(X = x_i | Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)}$$

if $P(Y = y_j)$ where $P(Y = y_j)$ is defined by $P(Y = y_j) = \sum_{x_i} P(X = x_i, Y = y_j)$.

- **Conditional expected value** is the expected value of a conditional distribution of X under condition $Y = y_j$. That is,

$$E(X|Y = y_j) = \sum_{x_i} x_i \cdot P(X = x_i|Y = y_j).$$

- Note that if X and Y are independent, then $P(X = x_i|Y = y_j) = P(X = x_i)$.

part I

INFORMATION THEORY
BASIC CONCEPTS AND RESULTS

Entropy

We suppose p is a probability distribution of a discrete random variable X (on some probability space $\langle \Omega, \mathcal{B}, P \rangle$) such that X takes only a finite number of values, namely $\{x_1, \ldots, x_n\}$. Furthermore, put and $p_i = p(x_i) = P(X = x_i)$.

Definition Uncertainty (or **Shannon's entropy**) $E(p)$ of a probability distribution p is defined to be

$$E(X) = E(p) = E(p_1, \ldots, p_n) = C \sum_{i=1}^{n} -p_i \log p_i.$$

log denotes logarithm to the base 2, i.e. $\log p = \log_2 p$.

We assume $p_i > 0$ for all i (but we may have $p_i = 0$ and then put $0 \log 0 := 0$.

C is a constant (it is only of a technical importance; we will assume $C = 1$ in which case the units of E are **bits**).

$E(p)$ is to be understood as follows. Let p describe all what we know about the actual outcome of some experiment (but the notion of an experiment is to be understood in a broad sense, like experiment=any process ending with a result, the set of all results is $\{x_1, \ldots, x_n\}$). Then $E(p)$ expresses the amount of uncertainty associated with p.

Example (1) If $X = \{x_1, x_2\}$ and $p_1 = p_2 = 1/2$ then $E(p) = 1$ (bit). So there is an uncertainty of 1 (bit) associated with a random process with two equally likely outputs.

(2) Tossing a fair die vs. tossing a biased die.

(3) Graph of $E(p, 1-p)$ for $p \in [0, 1]$ shows intuitive meaning of E.

The following theorem shows that E is not arbitrarily chosen. It is uniquely determined by a system of natural axioms (requirements).

Theorem (uniqueness) E is the only function satisfying

1. $f(n) = E(1/n, \ldots, 1/n)$ is a monotonically increasing function of n (here, $E(1/n, \ldots, 1/n)$ is $E(p)$ for $p_1 = 1/n$, dots, $p_n = 1/n$);

2. $f(mn) = f(m) + f(n)$;

3. (branching)

$$
\begin{aligned}
&E(p_1, \ldots, p_n) = E(p_1 + \cdots + p_r, p_{r+1} + \cdots + p_n) + \\
&+(p_1 + \cdots + p_r) E(p_1 / \sum_{i=1}^{r} p_i, \ldots, p_r / \sum_{i=1}^{r} p_i) + \\
&+(p_{r+1} + \cdots + p_n) E(p_{r+1} / \sum_{i=r+1}^{n} p_i, \ldots, p_n / \sum_{i=r+1}^{n} p_i);
\end{aligned}
$$

4. $E(p, 1-p)$ is a continuous function of p (note that for any $p \in [0,1]$, $p_1 = p$, $p_2 = 1 - p$ describes a probability distribution on a two-element set).

Two interpretations of E

First. E is (by the very definition) the average value of a random variable $X(x) = -\log p(x)$.

Second. Imagine a test consisting of “yes/no” questions whose aim is to determine the actual value x_i (and that the probability that the actual value is x_i is $p_i = p(x_i)$). For example, one possible test (but not optimal) would be a series of n questions “is the value x_1?”, . . . , “is the value x_n?” It can be shown that $E(p)$ is the minimum average number of questions of a test necessary to determine the value x_i (that is: we run a series of experiments with the optimal test; we choose x_i randomly according to p; then we run the test and record the number of questions asked in the test (for different x_i’s the numbers may be different because each such a test is in fact a binary tree of questions); then the assertion says that there is a test which ends in average after $E(p)$ questions and there is no better test!).

Basic properties

Lemma For positive numbers $p_1, \ldots, p_n$ and $q_1, \ldots, q_n$ with $\sum_{i=1}^n p_i = \sum_{i=1}^n q_i = 1$:
$-\sum_{i=1}^n p_i \log p_i \leq -\sum_{i=1}^n p_i \log q_i$ with equality iff $p_i = q_i$ for all i.

Proof For convenience, we use natural logarithms (which is OK since $\log_2 x = \log_2 e \cdot \log_e x$).

- $\log_e$ is a convex function, thus
- tangent at $x = 1$: point on $\log_e$ lie below the tangent which is a function $x - 1$ (note that $(\log_e x)' = 1/x$), i.e. $\log_e x \leq x - 1$ with equality iff $x = 1$, thus
- $\log(q_i/p_i) \leq q_i/p_i - 1$ with equality iff $p_i = q_i$, thus multiplying by p_i a summing over i we get
- $\sum_i p_i \log_e q_i/p_i \leq \sum_i (q_i - p_i) = 1 - 1 = 0$ with equality iff $p_i = q_i$, thus
- $\sum_i p_i \log_e q_i - \sum_i p_i \log_e p_i \leq 0$ with equality iff $p_i = q_i$. □

Theorem $E(p) \leq \log n$, with equality if and only if $p_i = 1/n$.

Proof By above Lemma with $q_i = 1/n$:
$-\sum_i p_i \log_e p_i \leq -\sum_i p_i \log 1/n = \log n \sum_i p_i = \log n$. □

Therefore, E takes the maximal value if all results x_i have probability $1/n$—this is surely the situation intuitively considered as most uncertain. On the other hand, if for some i we have $p_i = 1$ while $p_j = 0$ for $j \neq i$, then $E(p) = 0$, i.e. there is no uncertainty.

Joint uncertainty

Suppose we have two random variables X and Y on $\langle \Omega, \mathcal{B}, P \rangle$ (i.e. the same experiment).

Suppose $p_{ij} = p(x_i, y_j)$ $(i = 1, \dots n, j = 1, \dots, m)$ is a joint probability distribution of random vector $\langle X, Y \rangle$.

Denote the marginal probability distributions by p_i $(i = 1, \dots, n,$ distribution for $X)$ and p_j $(j = 1, \dots, m,$ distribution for $Y)$.

The **joint uncertainty** (joint entropy) is defined by

$$E(X,Y) = \sum_{i=1}^{n} \sum_{j=1}^{m} -p_{ij} \log p_{ij}.$$

Theorem We have $E(X,Y) \leq E(X) + E(Y)$ with equality holding if and only if X and Y (i.e. $p_{ij} = p_i \cdot p_j$) are independent.

Proof Since $p(x_i) = \sum_j p(x_i, y_j)$, $p(y_j) = \sum_i p(x_i, y_j)$, we have

$$E(X) = -\sum_i p(x_i) \log p(x_i) = -\sum_i \sum_j p(x_i, y_j) \log p(x_i)$$

and

$$E(Y) = -\sum_j p(y_j) \log p(y_j) = -\sum_i \sum_j p(x_i, y_j) \log p(y_j).$$

It follows

$$\begin{aligned} E(X) + E(Y) &= -\sum_i \sum_j p(x_i, y_j)[\log p(x_i) + \log p(y_j)] = \\ &= -\sum_i \sum_j p(x_i, y_j) \log(p(x_i)p(y_j)) = \\ &= -\sum_i \sum_j p(x_i, y_j) \log q_{ij} \end{aligned}$$

with $q_{ij} = p(x_i)p(y_j)$.
By definition, $E(X, Y) = \sum_{i=1}^{n} \sum_{j=1}^{m} -p_{ij} \log p_{ij}$.
Applying Lemma, we get

$$-\sum_i \sum_j p_{ij} \log p_{ij} \leq -\sum_i \sum_j p_{ij} \log q_{ij}$$

with equality iff $p_{ij} = q_{ij}$.
Note that Lemma can be applied since

$$\sum_i \sum_j q_{ij} = \sum_i p(x_i) \sum_j p(y_j) = 1 \cdot 1 = 1.$$

Therefore, $E(X,Y) \leq E(X) + E(Y)$ with equality iff $p_{ij} = q_{ij} = p(x_i)p(y_j)$, i.e. iff X and Y are independent. □

Conditional uncertainty

Recall that conditional distribution of Y under condition $X = x_i$ is given by $p(y_j|x_i) = p(x_i, y_j)/p(x_i)$ (x_i is fixed).

Under the above notation, the **conditional uncertainty of Y given $X = x_i$** is defined by

$$E(Y|X = x_i) = -\sum_{j=1}^{m} p(y_j|x_i) \log p(y_j|x_i).$$

The **conditional uncertainty of Y given X**, denoted $E(Y|X)$, is defined as the weighted average of $E(Y|X = x_i)$, i.e. by

$$E(Y|X) = \sum_i p(x_i) E(Y|X = x_i) = -\sum_i p(x_i) \sum_j p(y_j|x_i) \log p(y_j|x_i).$$

Using $p(x_i, y_j) = p(x_i)p(y_j|x_i)$, we get

$$E(Y|X) = -\sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \log p(y_j|x_i).$$

Theorem $E(X, Y) = E(X) + E(Y|X) = E(Y) + E(X|Y)$.

Proof Directly by definition:

$$\begin{aligned} E(X,Y) = -\sum_i\sum_j p(x_i,y_j)\,\mathsf{log}\,p(x_i,y_j) = \\ = & -\sum_i\sum_j p(x_i,y_j)\,\mathsf{log}\,p(x_i)p(y_j|x_i) = \\ = & -\sum_i\sum_j p(x_i,y_j)\,\mathsf{log}\,p(x_i) - \sum_i\sum_j p(x_i,y_j)\,\mathsf{log}\,p(y_j|x_i) = \\ = & -\sum_i p(x_i)\,\mathsf{log}\,p(x_i) + E(Y|X) = \\ = & \; E(X) + E(Y|X), \end{aligned}$$

and similarly for $E(X,Y) = E(Y) + E(X|Y)$. □

Theorem $E(Y|X) \leq E(Y)$ with equality iff X and Y are independent.

Proof By the above Theorems, $E(X,Y) = E(X) + E(Y|X)$ and $E(X,Y) \leq E(X)+E(Y)$ with equality iff X and Y are independent. The assertion directly follows. □

Note: As $E(X)$, both $E(X,Y)$ and $E(Y|X)$ can be seen as expected values of random variables (of $W(X,Y) = -\mathsf{log}\,p(x_i,y_j)$ and $W(Y|X) = -\mathsf{log}\,p(y_j,x_i)$).

Information conveyed about X by Y

(or mutual information between X and Y) is defined by

$$I(X|Y) = E(X) - E(X|Y).$$

Natural interpretation: the difference the uncertainty about X minus uncertainty about X given Y.

Theorem We have

- $I(X|Y) \geq 0$,
- $I(X|Y) = 0$ iff X and Y are independent,
- $I(X|Y) = I(Y|X)$.

Proof By above Theorem, $E(X|Y) \leq E(X)$ and $E(X|Y) = E(X)$ iff X and Y are independent. Therefore, $I(X|Y) = E(X) - E(X|Y) \geq 0$, and $I(X|Y) = 0$ iff X and Y are independent.

Furthermore, by above Theorem, $E(X|Y) = E(X,Y) - E(Y)$, thus

$$I(X|Y) = E(X) + E(Y) - E(X,Y).$$

But since $E(X,Y) = E(Y,X)$, the required identity $I(X|Y) = I(Y|X)$ readily follows. □

(!) Sometimes it is said that information theory (and statistics in general) is bad if we want to explain causality (it is the same both ways).

Example (information conveyed about X by Y) Two coins, one is unbiased (probability distribution $p_H = p_T = \frac{1}{2}$), one is two-headed (probability distribution $p_H = 1$, $p_T = 0$), H . . . head, T . . . tail.

Experiment: A coin is tossed twice. The number of heads is recorded. How much information is conveyed about which coin has been tossed by the number of heads obtained (identity of the coin tossed)?

Initial observations: If the number of heads is < 2 then the unbiased coin have been used; if the number of heads is $= 2$ then it is more likely that the two-headed coin has been used (intuition).

Approach via information-theoretic concepts. Consider two random variables: X ($X = 0$ . . . unbiased coin, $X = 1$ . . . two-headed coin), Y (number of heads, i.e. we may have $Y = 0, 1, 2$). Consider a random vector $\langle X, Y\rangle$.

Then we have (verify):
$P(X = 0) = \frac{1}{2}$, $P(X = 0) = \frac{1}{2}$,
$P(Y = 0) = \frac{1}{8}$, $P(Y = 1) = \frac{1}{4}$, $P(Y = 2) = \frac{5}{8}$,
$P(X = 0|Y = 0) = 1$, $P(X = 0|Y = 1) = 1$, $P(X = 0|Y = 2) = \frac{1}{5}$.

Before knowing the number of heads, the uncertainty of the identity of the coin is

$$E(X) = \log 2 = 1.$$

The uncertainty of X given Y is

$$\begin{aligned}
E(X|Y) &= P(Y=0)H(X|Y=0) + P(Y=1)H(X|Y=1) + P(Y=2)H(X|Y=\\
&= \frac{1}{8}0 + \frac{1}{4}0 - \frac{5}{8}(\frac{1}{5}\log\frac{1}{5} + \frac{4}{5}\log\frac{4}{5}) =\\
&= 0.45.
\end{aligned}$$

Therefore, the answer is

$$I(X|Y) = E(X) - E(X|Y) = 0.55\text{bits}.$$

Gain: Quantitative answer, well-founded. □

Exercise(Ash, Information Theory) In a school, $\frac{3}{4}$ of the students pass and $\frac{1}{4}$ fail.
Of those who pass, 10 percent own cars, of those who fail, 50 percent own cars.
All of the car-owning students belong to fraternities,
40 percent of those who do not own cars but pass belong to fraternities,
40 percent of those who do not own cars but fail belong to fraternities.

Questions:
(a) How much information is conveyed about a student's academic standing by specifying whether or not he owns a car?
(b) How much information is conveyed about a student's academic standing by specifying whether or not he belongs to fraternity?
(c) If a student's academic standing, car-owning status, and fraternity status are transmitted by three successive binary digits, how much information is conveyed by each digit?

part II

INFORMATION THEORY SELECTED APPLICATIONS

DECISION TREES

basic features

- a data mining technique
- we present a method of construction of decision trees developed in Quinlan J. R.: Induction of decision trees. *Machine Learning* **11**(1)(1986), 81–106.
- more advanced method(s) can be found in Quinlan J. R.: *C4.5: Programs for Machine Learning.* Morgan Kaufman, San Francisco, 1993.
- any textbook on machine learning/data mining (e.g. Berka P.: Dobývání znalostí z databází, Academia, Praha, 2003).

basic concepts

- input: data table with input attributes $y_1, \dots, y_m$ and output attribute y with val$(y) = \{c_1, \dots, c_k\}$ (val(y) . . . values of attribute y)

- c_i . . . classes (labels of classes) to which objects are to be classified

- example: (S . . . small, M . . . medium, T . . . tall)

name	gender	height	class
A.	male	185cm	T
B.	female	150cm	S
C.	female	179cm	T
D.	male	163cm	S
E.	male	170cm	M
F.	male	180cm	M
G.	female	169cm	M

  here:
  y_1 . . . gender (val(y_1) = {male, female}),
  y_2 . . . height (val$(y_2) = [50, 250]$),
  y . . . class (val(y) = {S, M, T}).

- objects: A.–G.

- row (A., male, 185cm, T) says: object with values of input variables being "male" and "185" is to be classified as "T".

- goal: given a data table, develop a *suitable* decision tree

What is a **decision tree**:

- in short: a tree used for classification

- rooted tree where

- each leaf is labeled by a class label (from $c_1, \dots, c_k$)

- each inner node (including root) is assigned an input variable y_i (one of $y_1, \dots, y_m$),

- out of each inner node there goes a finite (usually small) number of edges to its successor nodes;

- each edge going from a node labeled by y_i is assigned a binary condition C for values of y_i (this means: each value $a \in \mathrm{val}(y_i)$ either satisfies C or not) such that different edges going from a node are assigned mutually exclusive conditions (that is: each value $a \in \mathrm{val}(y_i)$ satisfies at most one condition)

- examples of conditions: “gender is male”, “gender is female”, “height is $\geq$ 185cm”, “height is = 180cm”, etc.

given a decision tree, a classification of an object given by values $(a_1, \ldots, a_m)$ of input variables $(y_1, \ldots, y_m)$ proceeds as follows:

- current node := root node;

- while current node is not a leaf do
  current node := successor S of current node N (N is labeled by y_i) for which the condition assigned to the edge from N to S is satisfied for a_i;

- output value is the class label assigned to the leaf at which we arrived

A decision tree which correctly classifies all the objects from the input data table is not unique. So:

What makes a decision tree a **good decision tree**?

- a decision tree should be as small as possible (height, width): if presented to a user, it should be comprehensible

- it should classify well the data in the input table but also further data which may come for classification in future (this is referred to as generalization capability of decision trees)

In the following, we present algorithm ID3 (proposed by Quinlan).

First, some simplifying assumptions and features of ID3:

- For each input variable y_i, the set val(y_i) is finite. This excludes e.g. real-valued variables like "height". For these variables, one can divide val(y_i) into a finite number of intervals ($[50, 55), [55, 60), \ldots, [245, 250)$). Division into intervals can even be found automatically to be optimal (details omitted).

- We require all input data be classified correctly. However, it might be useful to allow for some small number of errors in classification of input data. Namely, input data may contain noise (errors). Insisting on correct classification of all input data would lead to poor generalization capability.

- Variable y_i assigned to an inner node is called a **splitting variable**. Each condition assigned an edge going from a node with splitting variable has the form $y_i = a$ where $a \in$ val(y_i).

More advanced versions of Quinlan's basic ID3 algorithm can be found in: Quinlan J. R.: *C4.5: Programs for Machine Learning.* Morgan Kaufman, San Francisco, 1993.

ID3 algorithm for inducing decision trees

INPUT: data table $\mathcal{T} = \langle X, Y, T \rangle$,
with $Y = \{y_1, \dots, y_m, y\}$, y_i ...input variable, y ...output variable, val(y_i) finite, val(y) $= \{c_1, \dots, c_k\}$, $T(x, y_i)$ ...value of y_i on x

OUTPUT: decision tree $\mathcal{DT}$

ID3 ALGORITHM

```
buildDT($\mathcal{T}$)

1. if all objects from $\mathcal{T}$ have the same value $c_j$ of output variable $y$ then
   $\mathcal{DT}$ := single node labeled by $c_j$ and STOP;
   else
2. determine the best splitting variable $y_i$ for $\mathcal{T}$;
3. $\mathcal{DT}$ := new node labeled by $y_i$;
4. for each value $a \in$ val($y_i$) of $y_i$:
   (a) create an edge going from the node labeled by $y_i$ and label it by $y_i = a$;
   (b) $\mathcal{T}'$ := new table which results from $\mathcal{T}$ by restricting on objects satisfying
       $y_i = a$ (i.e. objects $x$ with $T(x, y_i) = a$, all other objects get deleted)
   (c) add to the new edge a decision tree which results as buildDT($\mathcal{T}'$)
```

Remarks

- buildDT() is a recursive procedure;

- for a given $a \in$ val(y_i), $\mathcal{T}' = \langle X', Y, T' \rangle$ results from $\mathcal{T} = \langle X', Y, T \rangle$ by $X' = \{x \in X \mid T(x, y_i) = a\}$ and $T'(x, y_j) = T(x, y_j)$ for $x \in X'$ and $y_j \in Y$ (T' is a restriction of T)

- what remains to be specified is the choice of a best splitting variable for a given table $\mathcal{T}$.

Choice of the **best splitting variable** for $\mathcal{T}$:

1. intuition:
   pick input variable $y_i \in Y$ such that, if possible, for each $a \in$ val(y_i), all the objects from X which satisfy $y_i = a$, have the same value of the output variable c (why: then each node to which there leads an edge labeled $y_i = a$ becomes a leaf and the tree is low)

2. more precisely (and in general):
   if this is not possible, pick y_i such that "on an average choice of $a \in$ val(y_i)", the distribution of the values of the output variable y to the objects from X which satisfy $y_i = a$ has "small entropy" (entropy=0 means all have the same value of y).

NOW:
2. can be formalized using information theory, namely:

best splitting attribute = attribute y_i with minimal conditional entropy (uncertainty) $E(y|y_i)$

BUT: How can we see attributes $y_1, \dots, y_n, y$ as random variables? As follows:

- consider a probability distribution p on X, $|X| = n$, with $p(\{x\}) = 1/n$ (equiprobable elementary events, rovnomerne rozdeleni);

- then each $z \in Y = \{y_1, \dots, y_m, y\}$ can be seen as a random variable on X assigning to an object $x \in X$ a value $T(x, z) \in \text{val}(z)$;

- and so we have

$$p(z = a) = p(\{x \in X \mid T(x, z) = a\}) = \frac{|\{x \in X \mid T(x, z) = a\}|}{n}$$

and we can speak of entropies $E(y_i)$ of input variables, entropy $E(y)$ of the output variable, conditional entropies $E(y|y_i)$, etc.

Recall basic facts (see part Information Theory):

- $p(y = a, y_i = b) = p(\{x \in X \mid T(x, y) = a, T(x, y_i) = b\})$, $p(y = a|y_i = b) = p(y = a, y_i = b)/p(y_i = b)$,

- $E(y|y_i) = \sum_{b \in \mathsf{val}(y_i)} p(y_i = b) \cdot E(y|y_i = b)$
- $E(y|y_i = b) = -\sum_{a \in \mathsf{val}(y)} p(y = a|y_i = b) \log p(y = a|y_i = b) = -\sum_{a \in \mathsf{val}(y)} p(y = a, y_i = b)/p(y_i = b) \log p(y = a, y_i = b)/p(y_i = b)$

Note that $E(y|y_i) = 0$ means that for each $b \in \mathsf{val}(y_i)$ we have: either $E(y|y_i = b) = 0$, i.e. for each $b \in \mathsf{val}(y_i)$, all objects satisfying $y_i = b$ have the same value of y (belong to the same class)
or $p(y_i = b) = 0$, i.e. value b does not appear in column labeled y_i

Remark The probability distribution p (and all the other variates including conditional probability $E(y|y_i)$) are relative to the data table $\mathcal{T}$. $\mathcal{T}$ changes during ID3!

Decision trees: example

input taken from Berka P.: Dobývání znalostí z databází, Academia, Praha, 2003.

client	income	account	gender	unemployed	credit
1	H	H	F	No	Yes
2	H	H	M	No	Yes
3	L	L	M	No	No
4	L	H	F	Yes	Yes
5	L	H	M	Yes	Yes
6	L	L	F	Yes	No
7	H	L	M	No	Yes
8	H	L	F	Yes	Yes
9	L	M	M	Yes	No
10	H	M	F	No	Yes
11	L	M	F	Yes	No
12	L	M	M	No	Yes

income (i): H . . . high, L . . . low,

account (a, amount of money on account): H . . . high, M . . . medium, L . . . low,

gender (g): M . . . male, F . . . female credit (c): whether a credit is to be approved for a client

case 1. At the beginning, $\mathcal{T}$ is the whole table. We need to compute $E(c|i)$, $E(c|a)$, $E(c|g)$, $E(c|u)$:

$$\begin{aligned}
E(c|i) &= p(i=H)\cdot E(c|i=H) + p(i=L)\cdot E(c|i=L) = \\
&= 5/12 \cdot E(c|i=H) + 7/12 \cdot E(c|i=L) \approx 5/12 \cdot 0 + 7/12 \cdot 0.985 = \\
&\approx 0.575,
\end{aligned}$$

since

$$\begin{aligned}
E(c|i=H) &= -p(c=Yes, i=H)/p(i=H) \cdot \log[p(c=Yes, i=H)/p(i=H)] - \\
&\quad -p(c=No, i=H)/p(i=H) \cdot \log[p(c=No, i=H)/p(i=H)] = \\
&= 1 \log 1 - 0 \cdot \log 0 = 0 - 0 = 0
\end{aligned}$$

and

$$\begin{aligned}
E(c|i=L) &= -p(c=Yes, i=L)/p(i=L) \cdot \log[p(c=Yes, i=L)/p(i=L)] - \\
&\quad -p(c=No, i=L)/p(i=L) \cdot \log[p(c=No, i=L)/p(i=L)] = \\
&= -3/7 \log 3/7 - 4/7 \cdot \log 4/7 =\approx 0.985.
\end{aligned}$$

And similarly:

$$E(c|a) \approx 0.667, E(c|g) \approx 0.918, E(c|u) \approx 0.825.$$

⇒ the best splitting attribute is i, which leads to creation of a new node $N1$ labeled by i from which there are two edges, one labeled by $i = H$, the other one labeled $i = L$.

case 1.1. for $i = H$, the corresponding $\mathcal{T}'$ results from $\mathcal{T}$ by deleting all objects with values of i different from H. In this table, all the objects have a value of c equal to Yes. This leads to creation of a new node $N2$ labeled by Yes (leaf).

case 1.2. for $i = L$, the corresponding $\mathcal{T}'$ results from $\mathcal{T}$ by deleting all objects with values of i different from L. In this table, not all of the objects have the same value of c. The algorithm continues by computing buildDT($\mathcal{T}'$). To get the best splitting attribute, we do not need to consider i since it was used already (it is easily seen that $E(c|i) = E(c) \geq E(c|z)$ for any $z = a, g, u$; and this holds true in general).

Proceeding this way (but note that the table changed and we need to compute the entropies for the new table $\mathcal{T}'$), we get

$$E(c|a) \approx 0.396, E(c|g) \approx 0.965, E(c|u) \approx 0.979.$$

⇒ the best splitting atribute is a and we create a new node $N3$ labeled a with three edges labeled by $a = H$, $a = M$, $a = L$.

case 1.2.1. For $a = H$, all objects (i.e. satisfying $i = L$ and $a = H$) have value of c equal to Yes. This leads to creation of a new node $N4$ labeled by Yes (leaf).

case 1.2.2. for $a = M$, the corresponding $\mathcal{T}''$ results from $\mathcal{T}'$ by deleting all objects with values of a different from M. In this table, not all of the objects have the same value of c. The algorithm continues by computing buildDT($\mathcal{T}''$). We need to compute $E(c|g)$ and $E(c|u)$ from $\mathcal{T}''$. We get

$$E(c|g) \approx 0.667, E(c|u) = 0.$$

$\Rightarrow$ the best splitting atribute is u and we create a new node $N5$ labeled u with three edges labeled by $u = Yes$, $u = No$.

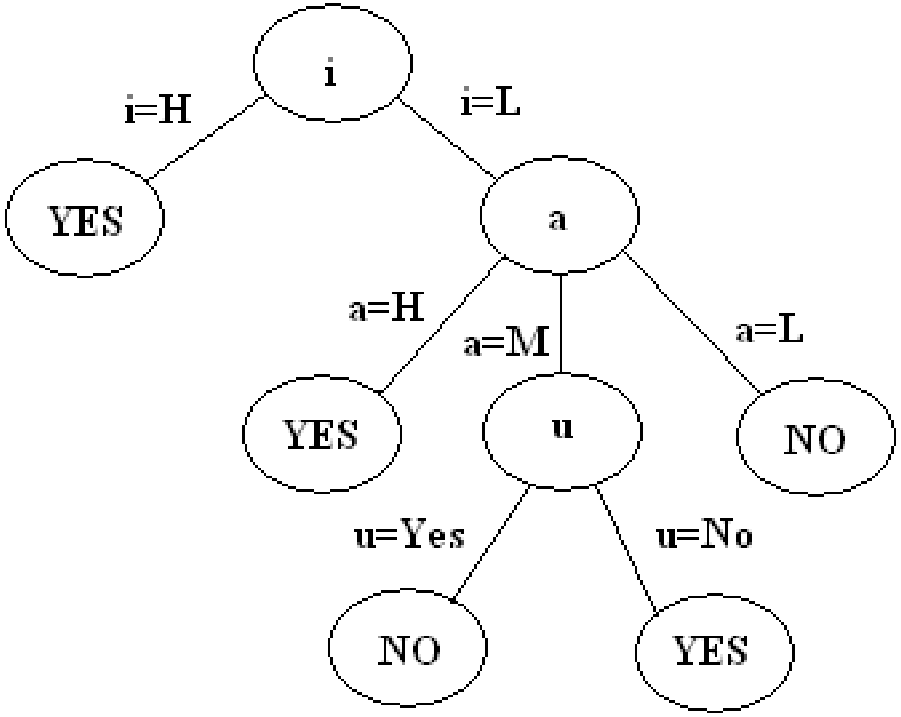
case 1.2.2.1 For $u = Yes$, all objects (i.e. satisfying $i = L$, $a = L$, $u = Yes$) have value of c equal to No. This leads to creation of a new node $N7$ labeled by No (leaf).

case 1.2.2.2 For $u = No$, all objects (i.e. satisfying $i = L$, $a = L$, $u = No$) have value of c equal to Yes. This leads to creation of a new node $N8$ labeled by Yes (leaf).

case 1.2.3. For $a = L$, all objects (i.e. satisfying $i = L$ and $a = L$) have value of c equal to No. This leads to creation of a new node $N6$ labeled by No (leaf).

OUTPUT decision tree:
nodes (node, label) are $(N1, i)$, $(N2, YES)$, $(N3, a)$, $(N4, Yes)$, $(N5, u)$, $(N6, No)$, $(N7, No)$, $(N8, Yes)$,

edges (sourceNode, targetNode, label) are $(N1, N2, i = H)$, $(N1, N3, i = L)$, $(N3, N4, a = H)$, $(N3, N5, a = M)$, $(N3, N6, a = L)$, $(N5, N7, u = Yes)$, $(N5, N8, u = No)$.



Further topics and extensions of ID3

Ovetraining: Occurs when decision tree is constructed to classify correctly all records in the input (training) table. This might not be desirable because input data may contain noise which is then learned.

Way out: Stop stop splitting nodes earlier. For example, stop splitting a node if p% (e.g., $p = 90$) of records corresponding to this node has the same value for the output attribute. Note: Stopping criterion for exact classification uses $p = 100$ (employed in the above ID3).

Algorithm C4.5: extension consists in

– Handles missing data (very simple, ignores items needed for computation which contain missing data).

– Continuous data. Divides data into ranges based on the values of continuous attributes.

– Pruning (simplification of decision trees). Subtree replacement (a subtree is replaced by a node if the classification error after such replacement is close to the one before replacement) is one strategy.

– Rules. Deals explicitly with classification rules corresponding to decision trees. Techniques for simplification of rules.

– Other strategies for the choice of splitting attributes.

Algorithm C5.0: commercial extension of C4.5

– Targeted for large datasets.

– Fast and memory efficient.

– Better accuracy (achieved due to boosting strategies).

– Algorithms not available.

Other approaches to classification

- Regression. Particularly for numeric data (linear/non-linear regression).

- Bayes classification. Highly developed approach. Based on Bayes theorem.
  - Assume the tuples in the dataset are $t_i = \langle a_i, c_i \rangle$ (just one input attribute for simplicity).
  - From the data, we can compute $P(a_i)$, $P(a_i|c_j)$, $P(c_j)$.
  - Using Bayes theorem, we get

$$P(c_j|a_i) = \frac{P(a_i|c_j)P(c_j)}{\Sigma_{j=1}^{k} P(a_i|c_j)P(c_j)}$$

    k is the number of values of the output attribute ($c_1, \dots, c_k$). Given a_i, we select the class c_j for which $P(c_j|a_i)$ is the largest.

  - This can be extended to $t_i = \langle a_{i1}, \dots, a_{in}, c_i \rangle$ under the assumption of independence of input variables. This way we come to so-called naive Bayes classifiers ("naive" because we assume independence).

- Distance-based classification, e.g., kNN (k nearest neighbors). For a new record (tuple) t to classify, we select k nearest records from the training data. Then t belongs to the class which contains the majority of the k nearest records.

- Neural networks. Backpropagation networks (multilayered feedforward networks), RBF (radial basis function networks), SVM (support vector machines).
- Recent approaches based on discrete mathematical structures. LAD (logical analysis of data), approaches based on FCA.