INTRODUCTION LECTURE TO NAÏVE BAYES CLASSIFIER AND FEATURE DISCRETIZATION

Kristjan Pilt, PhD

PROBABLITY

- **Probability** is a numerical description of how likely an event is to occur.

- The probability of an event A occurrence is calculated:

$$P(\mathrm{A}) = \frac{m}{n}$$

where *n* is the number of all equal elementary events (total number of outcomes) and *m* is the number of elementary events contained in event A (number of ways the event A can happen).

- Probability is a number between 0 and 1, where, roughly speaking, 0 indicates impossibility and 1 indicates certainty.

- Probability can be calculated as well after the experiments and value starts to stabilize around the constant value in case the number of experiments increases.

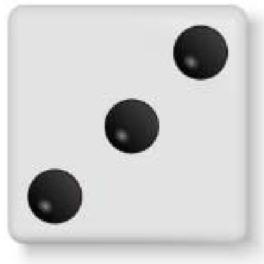
TAL TECH

PROBABILITY EXAMPLE

- **Example: Find the probability that the dice rolls 3.**
  - The number of all equal elementary events (n) is: 6 (as the dice can have equally values between 1 and 6)
  - The number of elementary events contained in event A (m) is: 1 (as the event is: getting number 3 with dice)
  - The probability of getting 3 is: $P(\mathrm{A}) = \frac{m}{n} = \frac{1}{6} = 0.166\ldots$

- This probability can be as well estimated through experiments:

n	m	$P(\mathrm{A})$
10	2	0.2
60	11	0.183
100	16	0.16



TAL TECH

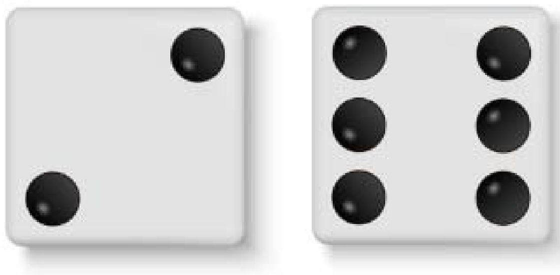
PROBABLITY OF TWO MUTUALLY EXCLUSIVE EVENTS

- The **probability that either first OR second mutually exclusive event occurs** is equal to the sum of the probabilities of these events:

$$P(\mathrm{A} \cup \mathrm{B}) = P(\mathrm{A}) + P(\mathrm{B})$$

- ∪ - means OR. Translation of equation: the probability that occurs event A **OR** occurs event B. The events A and B have to be exclusive! In case event A occurs then it is impossible that B can occur.
- **Example: Find the probability that the dice rolls either 2 OR 6.**

$$P(\mathrm{A} \cup \mathrm{B}) = P(\mathrm{A}) + P(\mathrm{B}) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3} = 0.333 \ldots$$



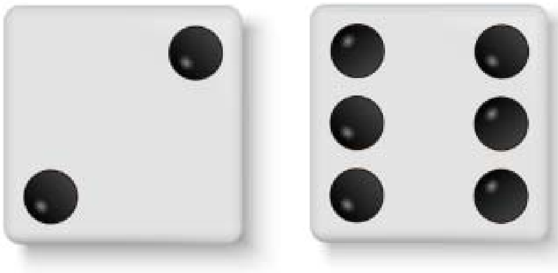
PROBABLITY OF TWO INDEPENDENT EVENTS

- The **probability that first AND second independent event occurs** is equal to the sum of the probabilities of these events:

$$P(\mathrm{A} \cap \mathrm{B}) = P(\mathrm{A}) \cdot P(\mathrm{B})$$

- $\cap$ - means AND. Translation of equation: the probability that occurs event A **AND** occurs event B. The events A and B have to be independent! The probability of event B occurring is independent of whether or not event A occurred.
- **Example: Find the probability that the dice rolls firstly 2 AND thereafter 6.**

$$P(\mathrm{A} \cup \mathrm{B}) = P(\mathrm{A}) \cdot P(\mathrm{B}) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36} = 0.0277 \ldots$$



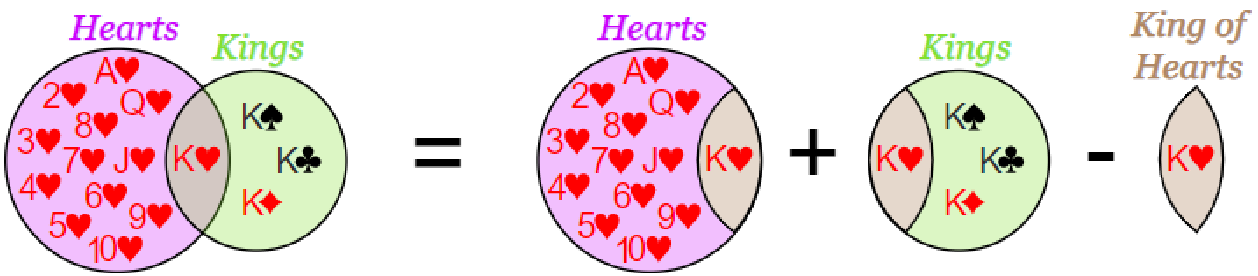
PROBABLITY OF TWO NOT MUTUALLY EXCLUSIVE EVENTS

- The **probability of two not mutually exclusive events** equals the sum of the probabilities of the events minus the probability of the co-occurrence or multiplication of those events:

$$P(\mathrm{A} \cup \mathrm{B}) = P(\mathrm{A}) + P(\mathrm{B}) - P(A \cap B)$$

- Translation of equation: the probability that occurs event A **OR** occurs event B and excluding the probability that they occur together.
- **Example: Find the probability of getting a heart OR king from a 52 card deck?**

$$P(A \cup B) = P(\mathrm{A}) + P(\mathrm{B}) - P(A \cap B) = \frac{1}{4} + \frac{4}{52} - \frac{1}{4} \cdot \frac{4}{52} = \frac{1}{4} + \frac{4}{52} - \frac{1}{52} \approx 0.308$$



TAL TECH

CONDITIONAL PROBABILITY (*TINGLIK TÕENÄOSUS VÕI KA OSASÜNDMUSE TÕENÄOSUS*)

- The **conditional probability** of event A with condition B is the probability of event A assuming event B occurred and is denoted by $P(\mathrm{A}|\mathrm{B})$.

- It is important that the event B has already happened.

- **Example: Two cards are taken from a 52 card deck. What is the probability that the second card is heart, assuming that the first card was heart?**

There are 13 hearts in a 52 card deck. Event B is that the first card was heart.

$$P(\mathrm{A}|\mathrm{B}) = \frac{13-1}{52-1} = \frac{12}{51} \approx 0.24$$

TAL TECH

PROBABLITY OF TWO DEPENDENT EVENTS

- The **probability of two dependent events** A and B occurring together, i.e. the occurrence of one event affects the other event, is equal to the product of the probability of one event and the conditional probability of the other:

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

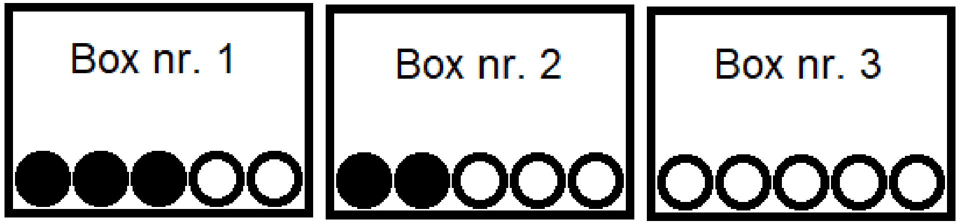
- **Example: Two cards are taken from a 52 card deck. What is the probability that the both cards are hearts?**

$$P(\mathrm{A} \cap \mathrm{B}) = P(\mathrm{A}) \cdot P(\mathrm{B}|\mathrm{A}) = \frac{13}{52} \cdot \frac{12}{51} = \frac{1}{4} \cdot \frac{12}{51} = \frac{12}{204} = \frac{3}{51} \approx 0.059$$

TAL TECH

LAW OF TOTAL PROBABILITY

- Lets consider the case of three boxes. Each of them includes white and black balls:



- Let the $B_1...B_N$ be the coplete system of events (e.g. picking ball from first box (B_1), picking ball from second box (B_2), etc.).
- Let A event (picking white ball) occur only with one of B_n events, with A conditional probability B_n being $P(A|B_n)$.
- Then the probability of event A (picking white ball) occuring is:

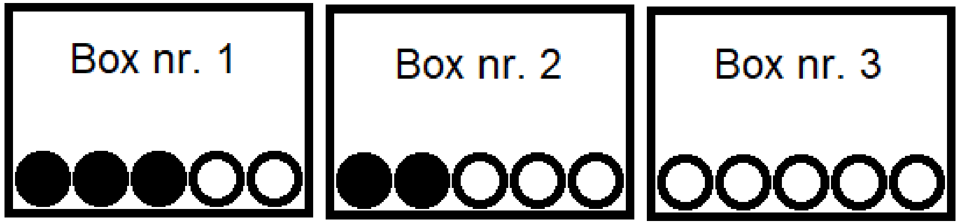
$$P(\mathrm{A}) = \sum_{n=1}^{N} P(\mathrm{B}_n) \cdot P(\mathrm{A}|\mathrm{B}_n)$$ **(Total probability equation)**

(*Täistõenäosuse valem*)

TAL TECH

LAW OF TOTAL PROBABILITY

- Lets consider the case of three boxes. Each of them includes white and black balls. **What is the probability of taking white ball?**



$$P(\mathrm{A}|\mathrm{B}_1) = \frac{2}{5}, \quad P(\mathrm{A}|\mathrm{B}_2) = \frac{3}{5}, \quad P(\mathrm{A}|\mathrm{B}_3) = \frac{5}{5}$$

- Probability that the ball is picked from first box is $P(\mathrm{B}_1) = \frac{1}{3}$. $(P(\mathrm{B}_2) = P(\mathrm{B}_3) = \frac{1}{3})$

$$P(\mathrm{A}) = \sum_{n=1}^{N} P(\mathrm{B}_n) \cdot P(\mathrm{A}|\mathrm{B}_n) = \frac{1}{3} \cdot \frac{2}{5} + \frac{1}{3} \cdot \frac{3}{5} + \frac{1}{3} \cdot \frac{5}{5} = \frac{2}{3}$$

- **Consider current case. White ball is picked. From which box it is taken from most probably? What about picking black ball? What if the probability is higher that the ball is taken from box nr. 2 as the box is colored to green and other boxes are red?**

TAL TECH

BAYES' EQUATION

- Lets call back **probability of two dependent events:**

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

- We can derive following relationships:

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

- Considering the total probability equation, we can write **(Bayes' equation)**:

$$P(\mathrm{B}_i|\mathrm{A}) = \frac{P(\mathrm{B}_i) \cdot P(\mathrm{A}|\mathrm{B}_i)}{\sum_{n=1}^{N} P(\mathrm{B}_n) \cdot P(\mathrm{A}|\mathrm{B}_n)}$$

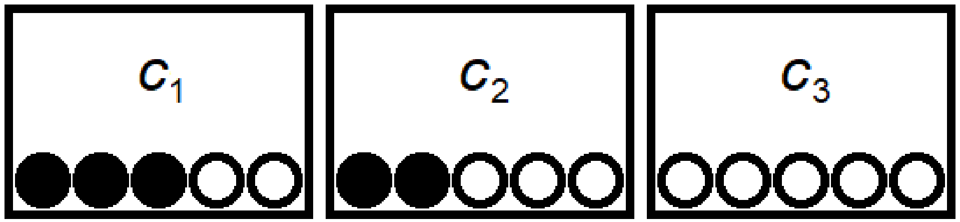
TAL TECH

NOTATION OF VARIABLES

- Let C be the class label. C consists of classes c_1 to c_N. $C \in \{c_1, c_2, \dots c_N\}$
- Let X is the observed features (attributes) $X = \{x_1, x_2, \dots x_M\}$
- Let the **box be class** and **color of ball be feature** in case of the example with boxes and balls.
- The probability distribution between classes and features is defined as follows:

$P(x_1 = \text{white}|c_1) = \frac{2}{5}$, $P(x_1 = \text{white}|c_2) = \frac{3}{5}$, $P(x_1 = \text{white}|c_3) = \frac{5}{5}$, $P(c_1) = \frac{1}{3}$, $P(c_2) = \frac{1}{3}$, $P(c_3) = \frac{1}{3}$,

$P(x_1 = \text{black}|c_1) = \frac{3}{5}$, $P(x_1 = \text{black}|c_2) = \frac{2}{5}$, $P(x_1 = \text{black}|c_3) = 0$

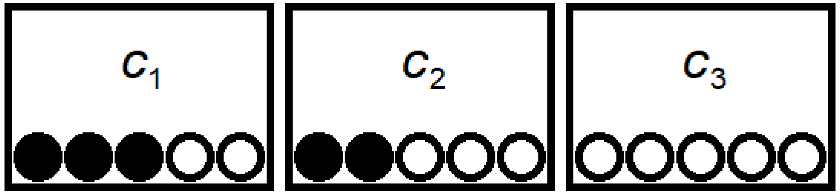


TAL
TECH

EXAMPLE USING BAYES' EQUATION

C	$P(C)$
c_1	0.333...
c_2	0.333...
c_3	0.333...

Colour	$P(X \mid c_1)$	$P(X \mid c_2)$	$P(X \mid c_3)$
white	0.4	0.6	1
black	0.6	0.4	0



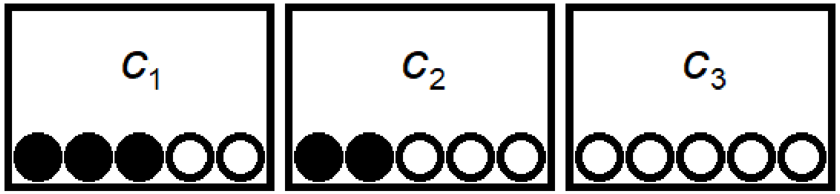
- **From which box the ball is taken if the ball is white?**
- The probability that the ball was taken from first box is calculated using Bayes' equation:

$$P(\mathrm{B}_i|\mathrm{A}) = \frac{P(\mathrm{B}_i) \cdot P(\mathrm{A}|\mathrm{B}_i)}{\sum_{n=1}^{N} P(\mathrm{B}_n) \cdot P(\mathrm{A}|\mathrm{B}_n)}$$

$$P(c_1|x_1 = \text{white}) = \frac{P(c_1) \cdot P(x_1 = \text{white}|c_1)}{\sum_{n=1}^{3} P(c_n) \cdot P(x_1 = \text{white}|c_n)} = \frac{\frac{1}{3} \cdot 0.4}{\frac{1}{3} \cdot 0.4 + \frac{1}{3} \cdot 0.6 + \frac{1}{3} \cdot 1} = 0.2$$

EXAMPLE OF BOX IDENTIFICATION RESULTS

Colour	$P(c_1\|X)$	$P(c_2\|X)$	$P(c_3\|X)$
white	0.2	0.3	**0.5**
black	**0.6**	0.4	0



- Maximal probabilities were found for each class.
- The ball was taken with 50% of probability from box nr. 3 in case the ball was white.
- The ball was taken with 60% of probability from box nr. 1 in case the ball was black.
- As there are only balls with two different colours then the probability that the ball is taken from box nr. 2 is lower than for the other two boxes.

TAL TECH

PROBABILISTIC MODELS AS CLASSIFIERS

- Let X denote the variables we know about, e.g., our instance's feature values; and let C denote the target variables we're interested in, e.g., the instance's class.

$$X \xrightarrow{f} C' \qquad C' \text{ vs } C$$

- The key question in machine learning is how to model the relationship (f) between X and C .

- The **statistician's approach** is to assume that there is some underlying random process that generates the values for these variables, according to a well-defined but unknown probability distribution. We want to use the data to find out more about this distribution.

BAYES EQUATION SIMPLIFICATIONS

- Bayes' equation: $P(c_n|X) = \frac{P(c_n) \cdot P(X|c_n)}{\sum_{k=1}^{K} P(c_k) \cdot P(X|c_k)}$
- Bayes' equation can be extended to become a naïve Bayes classifier with **two simplifications**:

1. To use the conditional **independence assumption**. Each feature (attribute) is conditionally independent of every other feature under given a class label c_n.
   - For example ball mass and its density are **dependent**
   - For example ball colour and size are **independent**
   - Therefore the calculation of $P(X|c_n)$ for number of features M can be carried out as follows:

$$P(x_1, x_2, \dots, x_M|c_n) = P(x_1|c_n) \cdot P(x_2|c_n) \cdot \ldots \cdot P(x_M|c_n) = \prod_{m=1}^{M} P(x_m|c_n)$$

2. To **ignore the denominator** $\sum_{k=1}^{K} P(c_k) \cdot P(X|c_k)$, because it appears in the denominator of $P(c_n|X)$ for all values of n. Removing the denominator will have no impact on the relative probability scores and will simplify calculations.

TAL TECH

NAÏVE BAYES CLASSIFIER

- Naïve Bayes classifier:

$$P(c_n|X) \propto P(c_n) \cdot \prod_{m=1}^{M} P(x_m|c_n)$$

- $\propto$ - means proportional!
- The calculated probability can't be taken as a true probability of $P(c_n|X)$. For all n values the calculated probabilities are proportional with eachother and can be **compared**.

- The Naïve Bayes **classifier is trained** using training set by computing the $P(c_n)$ and $P(x_m|c_n)$ for all possible m and n values (m=1...M and n=1...N).
- In case of **classification**, for each record in the testing set, the Naïve Bayes classifier assigns the classifier label c_n that **maximizes** $P(c_n) \cdot \prod_{m=1}^{M} P(x_m|c_n)$.

TAL TECH