Artificial Intelligence. Decision Tasks.

Petr Pošík

Czech Technical University in Prague
Faculty of Electrical Engineering
Dept. of Cybernetics

Artificial Intelligence

Artificial Intelligence — In a Broad Sense

Studies of *intelligence in general*:

- How do we *perceive* the world?
- How do we *understand* the world?
- How do we *reason* about the world?
- How do we *predict* the consequences of our actions?
- How do we act to *influence* the world?

Artificial Intelligence — In a Broad Sense

Studies of *intelligence in general*:

- How do we *perceive* the world?
- How do we *understand* the world?
- How do we *reason* about the world?
- How do we *predict* the consequences of our actions?
- How do we act to *influence* the world?

Artificial Intelligence (AI) not only wants to understand the "intelligence", but also wants to

- create an intelligent entity (agent, robot)
- imitating or improving
- the human behavior and effects in the outer world, and/or
- the inner human mind processes and reasoning.

Artificial Intelligence — In a Broad Sense

Studies of *intelligence in general*:

- How do we *perceive* the world?
- How do we *understand* the world?
- How do we *reason* about the world?
- How do we *predict* the consequences of our actions?
- How do we act to *influence* the world?

Artificial Intelligence (AI) not only wants to understand the "intelligence", but also wants to

- create an intelligent entity (agent, robot)
- imitating or improving
- the human behavior and effects in the outer world, and/or
- the inner human mind processes and reasoning.

Robot vs. agent:

- very often interchangeable terms describing systems with varying degrees of autonomy able to predict the state of the world and effects of their own actions. Sometimes, however:
- **agent:** the software responsible for the "intelligence"
- **robot:** the hardware, often used as substitute for humans in dangerous situations, in poorly accessible places, or for routine repeating actions

Question: What is AI for you?

In my opinion, the primary goal of AI is to build machines that

A. think like people.

B. act like people.

C. think reasonably, rationally.

D. act reasonably, rationally.

What is AI for us?

The science of making machines

- **think like people?** Not AI anymore, mix of cognitive science and computational neuroscience.

What is AI for us?

The science of making machines

- **think like people?** Not AI anymore, mix of cognitive science and computational neuroscience.
- **act like people?** No matter how they think, actions and behavior must be human-like. Dates back to Turing. But should we mimic even human errors?

What is AI for us?

The science of making machines

- **think like people?** Not AI anymore, mix of cognitive science and computational neuroscience.
- **act like people?** No matter how they think, actions and behavior must be human-like. Dates back to Turing. But should we mimic even human errors?
- **think rationally?** Requires correct thought process. Builds on philosophy and logic: how shall you think in order not to make a mistake? Our limited ability to express the logical deduction.

What is AI for us?

The science of making machines

- **think like people?** Not AI anymore, mix of cognitive science and computational neuroscience.
- **act like people?** No matter how they think, actions and behavior must be human-like. Dates back to Turing. But should we mimic even human errors?
- **think rationally?** Requires correct thought process. Builds on philosophy and logic: how shall you think in order not to make a mistake? Our limited ability to express the logical deduction.
- **act rationally.** Care only about what they do and if they achieve their goals optimally. Goals are described in terms of the utility of the outcomes. *Maximize the expected utility of the outcomes of their decisions.*

What is AI for us?

The science of making machines

- **think like people?** Not AI anymore, mix of cognitive science and computational neuroscience.
- **act like people?** No matter how they think, actions and behavior must be human-like. Dates back to Turing. But should we mimic even human errors?
- **think rationally?** Requires correct thought process. Builds on philosophy and logic: how shall you think in order not to make a mistake? Our limited ability to express the logical deduction.
- **act rationally.** Care only about what they do and if they achieve their goals optimally. Goals are described in terms of the utility of the outcomes. *Maximize the expected utility of the outcomes of their decisions.*

Good decisions:

- Take into account similar situations that happened in the past. **Machine learning.**
- Simulations using a model of the world. Be aware of the consequences of your actions and plan ahead. **Inference, planning.**

Science Disciplines Important for AI

Knowledge representation:

- how to store the model of the world, the relations between the entities in the world, the rules that are valid in the world, ...

Automated reasoning:

- how to infer some conclusions from what is known or answer some questions

Planning:

- how to find an action sequence that puts the world in the desired state

Pattern recognition:

- how to decide about the state of the world based on observations

Machine learning:

- how to create/adapt the model of the world using new observations

Multiagent systems:

- how to coordinate and cooperate in a group of agents to reach the desired goal

Science Disciplines Important for AI

Knowledge representation:

- how to store the model of the world, the relations between the entities in the world, the rules that are valid in the world, . . .

Automated reasoning:

- how to infer some conclusions from what is known or answer some questions

Planning:

- how to find an action sequence that puts the world in the desired state

Pattern recognition:

- how to decide about the state of the world based on observations

Machine learning:

- how to create/adapt the model of the world using new observations

Multiagent systems:

- how to coordinate and cooperate in a group of agents to reach the desired goal

Natural language processing:

- how to understand what people say and how to say something to them

Computer vision:

- how to understand the observed scene, what is going on in a sequence of pictures

Robotics:

- how to move, how to manipulate with objects, how to localize and navigate

. . .

Course outline

1. Bayesian and non-Bayesian decision tasks. Empirical learning.
2. Linear methods for classification and regression.
3. Non-linear model. Overfitting.
4. Nearest neighbors. Kernels, SVM. Decision trees.
5. Bagging. Boosting. Random forests.
6. Neural networks. Error backpropagation.
7. Deep learning. Convolutional and recurrent NNs.
8. Probabilistic graphical models. Bayesian networks.
9. Hidden Markov models.
10. Expectation-Maximization algorithm.
11. Constraint satisfaction problems.
12. Planning. Representations and methods.
13. Scheduling. Local search.

Decision Tasks and Decision Making

Observations and States

An **object (or situation)** of interest is described by two (sets of) parameters:

- $x \in X$ which is *observable*, called **observation**, or evidence, measurement, feature vector, etc.
- $k \in K$ which is *unobservable (hidden)*, called **hidden state**, state of nature, class, etc.

Observations and States

An **object (or situation)** of interest is described by two (sets of) parameters:

- $x \in X$ which is *observable*, called **observation**, or evidence, measurement, feature vector, etc.
- $k \in K$ which is *unobservable (hidden)*, called **hidden state**, state of nature, class, etc.

For a certain observation x (and unknown, but present k), we would like to make a **decision** $d \in D$, where D is the set of possible decisions.

Observations and States

An **object (or situation)** of interest is described by two (sets of) parameters:

- $x \in X$ which is *observable*, called **observation**, or evidence, measurement, feature vector, etc.
- $k \in K$ which is *unobservable (hidden)*, called **hidden state**, state of nature, class, etc.

For a certain observation x (and unknown, but present k), we would like to make a **decision** $d \in D$, where D is the set of possible decisions.

Examples:

- Radar detection of an aircraft:
  - Observation x: a particular observed radar reflection.
  - Hidden state k: the (unknown) truth whether the reflection belongs to an aircraft or not.
  - Decision d: an estimate, guess, or prediction of the true hidden state.
- Patient diagnosis:
  - Observation x: a set of diagnostic measurements – body temperature, blood tests, subjective description of feelings, etc.
  - Hidden state k: the (unknown) disease the patient suffers from.
  - Decision d: the kind of treatment that is to be prescribed to the patient. Ideally, something suitable to her disease.

Observations and States

An **object (or situation)** of interest is described by two (sets of) parameters:

- $x \in X$ which is *observable*, called **observation**, or evidence, measurement, feature vector, etc.
- $k \in K$ which is *unobservable (hidden)*, called **hidden state**, state of nature, class, etc.

For a certain observation x (and unknown, but present k), we would like to make a **decision** $d \in D$, where D is the set of possible decisions.

Examples:

- Radar detection of an aircraft:
  - Observation x: a particular observed radar reflection.
  - Hidden state k: the (unknown) truth whether the reflection belongs to an aircraft or not.
  - Decision d: an estimate, guess, or prediction of the true hidden state.
- Patient diagnosis:
  - Observation x: a set of diagnostic measurements – body temperature, blood tests, subjective description of feelings, etc.
  - Hidden state k: the (unknown) disease the patient suffers from.
  - Decision d: the kind of treatment that is to be prescribed to the patient. Ideally, something suitable to her disease.

The *observation* is almost always *noisy*, *incomplete*, or *corrupted*, i.e. contains various forms of **uncertainty**.

Decision Strategy Design

A general goal:

- Using an **observation** $x \in X$ of an object of interest (with a **hidden state** $k \in K$),
- we should find/design a **decision strategy** $q : X \to D$
- which would be **optimal** with respect to certain criterion,
- taking into account the **uncertainty** of the observation.

Decision Strategy Design

A general goal:

- Using an **observation** $x \in X$ of an object of interest (with a **hidden state** $k \in K$),
- we should find/design a **decision strategy** $q : X \to D$
- which would be **optimal** with respect to certain criterion,
- taking into account the **uncertainty** of the observation.

Bayesian decision theory requires

- complete statistical information about the object of interest in the form of the joint probability distribution $p_{XK}(x, k)$, and
- a suitable penalty/utility function $W : K \times D \to \mathcal{R}$.

Decision Strategy Design

A general goal:

- Using an **observation** $x \in X$ of an object of interest (with a **hidden state** $k \in K$),
- we should find/design a **decision strategy** $q : X \to D$
- which would be **optimal** with respect to certain criterion,
- taking into account the **uncertainty** of the observation.

Bayesian decision theory requires

- complete statistical information about the object of interest in the form of the joint probability distribution $p_{XK}(x, k)$, and
- a suitable penalty/utility function $W : K \times D \to \mathcal{R}$.

Non-Bayesian decision theory studies decision tasks for which some of the above information is not available.

Definitions of concepts

An **object** of interest is characterized by the following parameters:

- **observation** $x \in X$ (vector of numbers, graph, picture, sound, ECG, ...), and
- **hidden state** $k \in K$.
- k is often viewed as the object **class**, but it may be something different, e.g. when we seek for the location k of an object based on the picture x taken by a camera.

Definitions of concepts

An **object** of interest is characterized by the following parameters:

- **observation** $x \in X$ (vector of numbers, graph, picture, sound, ECG, ...), and
- **hidden state** $k \in K$.
- k is often viewed as the object **class**, but it may be something different, e.g. when we seek for the location k of an object based on the picture x taken by a camera.

Joint probability distribution $p_{XK} : X \times K \rightarrow \langle 0, 1 \rangle$

- $p_{XK}(x,k)$ is the joint probability that the object is in the state k and we observe x.
- $p_{XK}(x,k) = p_{X|K}(x|k) \cdot p_K(k)$

Definitions of concepts

An **object** of interest is characterized by the following parameters:

- **observation** $x \in X$ (vector of numbers, graph, picture, sound, ECG, ...), and
- **hidden state** $k \in K$.
- k is often viewed as the object **class**, but it may be something different, e.g. when we seek for the location k of an object based on the picture x taken by a camera.

Joint probability distribution $p_{XK} : X \times K \rightarrow \langle 0, 1 \rangle$

- $p_{XK}(x, k)$ is the joint probability that the object is in the state k and we observe x.
- $p_{XK}(x, k) = p_{X|K}(x|k) \cdot p_K(k)$

Decision strategy (or function or rule) $q : X \rightarrow D$

- D is a set of possible decisions. (Very often $D = K$.)
- q is a function that assigns a decision $d = q(x), d \in D$, to each $x \in X$.
- Q is a set of all possible decision strategies q, $q \in Q$.

Definitions of concepts

An **object** of interest is characterized by the following parameters:

- **observation** $x \in X$ (vector of numbers, graph, picture, sound, ECG, ...), and
- **hidden state** $k \in K$.
- k is often viewed as the object **class**, but it may be something different, e.g. when we seek for the location k of an object based on the picture x taken by a camera.

Joint probability distribution $p_{XK} : X \times K \rightarrow \langle 0, 1 \rangle$

- $p_{XK}(x, k)$ is the joint probability that the object is in the state k and we observe x.
- $p_{XK}(x, k) = p_{X|K}(x|k) \cdot p_K(k)$

Decision strategy (or function or rule) $q : X \rightarrow D$

- D is a set of possible decisions. (Very often $D = K$.)
- q is a function that assigns a decision $d = q(x), d \in D$, to each $x \in X$.
- Q is a set of all possible decision strategies q, $q \in Q$.

Penalty function (or loss function) $W : K \times D \rightarrow \mathcal{R}$ (real numbers)

- $W(k, d)$ is a penalty for decision d if the object is in state k.

Definitions of concepts

An **object** of interest is characterized by the following parameters:

- **observation** $x \in X$ (vector of numbers, graph, picture, sound, ECG, ...), and
- **hidden state** $k \in K$.
- k is often viewed as the object **class**, but it may be something different, e.g. when we seek for the location k of an object based on the picture x taken by a camera.

Joint probability distribution $p_{XK} : X \times K \to \langle 0, 1 \rangle$

- $p_{XK}(x, k)$ is the joint probability that the object is in the state k and we observe x.
- $p_{XK}(x, k) = p_{X|K}(x|k) \cdot p_K(k)$

Decision strategy (or function or rule) $q : X \to D$

- D is a set of possible decisions. (Very often $D = K$.)
- q is a function that assigns a decision $d = q(x), d \in D$, to each $x \in X$.
- Q is a set of all possible decision strategies q, $q \in Q$.

Penalty function (or loss function) $W : K \times D \to \mathcal{R}$ (real numbers)

- $W(k, d)$ is a penalty for decision d if the object is in state k.

Risk $R : Q \to \mathcal{R}$

- the criterion used to evaluate a decision strategy q in Bayesian tasks;
- the mathematical expectation of the penalty which must be paid when using the strategy q.

Question: Decision strategy?

Decision strategy (or function or rule) $q : X \rightarrow D$

- D is a set of possible decisions. (Very often $D = K$.)
- q is a function that assigns a decision $d = q(x), d \in D$, to each $x \in X$.
- Q is a set of all possible decision strategies $q, q \in Q$.

Example: You have a coin which may be biased. By observing 5 coin tosses, you should decide whether the coin is fair, biased, or you may say that you do not know.

- Hidden states: $K = \{\text{biased, fair}\}$
- Observations, number of heads: $X = \{0, \ldots, 5\}$
- Decisions: $D = \{\text{'fair', 'biased', 'I do not know'}\}$

Question: Decision strategy?

Decision strategy (or function or rule) $q : X \to D$

- D is a set of possible decisions. (Very often $D = K$.)
- q is a function that assigns a decision $d = q(x), d \in D$, to each $x \in X$.
- Q is a set of all possible decision strategies q, $q \in Q$.

Example: You have a coin which may be biased. By observing 5 coin tosses, you should decide whether the coin is fair, biased, or you may say that you do not know.

- Hidden states: $K = \{\text{biased, fair}\}$
- Observations, number of heads: $X = \{0, \dots, 5\}$
- Decisions: $D = \{\text{'fair', 'biased', 'I do not know'}\}$

How many different decision strategies are possible?

A. $3 \cdot 6 = 18$

B. $6^2 = 36$

C. $3^6 = 729$

D. $6^3 = 216$

Decision task examples

The description of the concepts is very general—so far we did not specify what the items of the X, K, and D sets actually are, how they are represented.

Application	Observation (measurement)	Decisions
Coin value in a slot machine	$x \in \mathcal{R}^n$	Value
Cancerous tissue detection	Gene-expression profile, $x \in \mathcal{R}^n$	{yes, no}
Medical diagnostics	Results of medical tests, $x \in \mathcal{R}^n$	Diagnosis
Optical character recognition	2D bitmap, intensity image	Words, numbers
License plate recognition	2D bitmap, grey-level image	Characters, numbers
Fingerprint recognition	2D bitmap, grey-level image	Personal identity
Face detection	2D bitmap	{yes, no}
Speech recognition	$x(t)$	Words
Speaker identification	$x(t)$	Personal identity
Speaker verification	$x(t)$	{yes, no}
EEG, ECG analysis	$\boldsymbol{x}(t)$	Diagnosis
Forfeit detection	Various	{yes, no}

Notes on decision tasks

In the following, we consider decision tasks where

- the decisions do not influence the state of nature (unlike *game theory* or *control theory*).
- a single decision is made, time is mostly ignored (unlike *control theory*, where decisions are typically taken continuously in real time).
- the costs of obtaining the observations are not modeled (unlike *sequential decision theory*).

The hidden parameter k (state, class) is considered not observable. Common situations are:

- k can be observed, but at a high cost.
- k is a future state (e.g. price of gold) and will be observed later.

Don't get confused by a different notation!

$X \times K \times D \times W$ used by Schlesinger and Hlaváč [SH12]

- observations X,
- hidden states K,
- decisions D,
- penalty function W.

$X \times \Omega \times A \times W$ used by Duda, Hart, and Stork [DHS01]

- observations X,
- hidden states/classes Ω (Y),
- decisions/actions A,
- penalty function W.

$E \times S \times A \times U$ used by Russel and Norvig [RN10]

- evidence E,
- hidden states S,
- decisions/actions A,
- utility function U.

[DHS01] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. Wiley, New York, 2 edition, 2001.

[RN10] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach (3rd Edition)*. Prentice Hall, 3 edition, 2010.

[SH12] M. I. Schlesinger and Václav Hlaváč. *Ten Lectures on Statistical and Structural Pattern Recognition (Computational Imaging and Vision)*. Springer, 2002 edition, March 2012.

Two types of pattern recognition

1. **Statistical pattern recognition**
   - Objects are represented as points in a vector space.
   - The point (vector) $\boldsymbol{x}$ contains the individual observations (in a numerical form) as its coordinates.
2. **Structural pattern recognition**
   - The object observations contain a structure which is represented and used for recognition.
   - A typical example of the representation of a structure is *a grammar*.

Bayesian Decision Theory

Question: Expected value of penalty?

You should know: Expected value of a discrete random variable V:

- If all values are equally probable: $E(V) = \frac{1}{N} \sum_{v \in V} v$
- In general case: $E(V) = \sum_{v \in V} p(v) \cdot v$

Question: Expected value of penalty?

You should know: Expected value of a discrete random variable V:

- If all values are equally probable: $E(V) = \frac{1}{N} \sum_{v \in V} v$
- In general case: $E(V) = \sum_{v \in V} p(v) \cdot v$

Question: Given

- set of observations X, set of hidden states K, set of decisions D,
- all probability distributions p_{XK}, p_X, p_K, $p_{X|K}$, $p_{K|X}$,
- penalty function $W : K \times D \to \mathcal{R}$, and
- decision strategy $q : X \to D$,

how do you compute the expected value of penalty W for a certain strategy q?

A. $\sum_{x \in X} p_{X|K}(x|k) \cdot W(x, q(k))$

B. $\sum_{x \in X} \sum_{k \in K} p_{XK}(x, k) \cdot W(k, q(x))$

C. $\sum_{x \in X} \sum_{k \in K} p_{X|K}(x|k) \cdot W(x, q(k))$

D. $\sum_{k \in K} p_{K|X}(k|x) \cdot W(k, d)$

Bayesian decision task

Given the sets X, K, and D, and functions $p_{XK} : X \times K \rightarrow \langle 0, 1 \rangle$ and $W : K \times D \rightarrow \mathcal{R}$, find a strategy $q : X \rightarrow D$ which minimizes the **Bayesian risk** of the strategy q

$$R(q) = \sum_{x \in X} \sum_{k \in K} p_{XK}(x, k) \cdot W(k, q(x)).$$

The optimal strategy q, denoted as q^*, is then called **the Bayesian strategy**.

Bayesian decision task

Given the sets X, K, and D, and functions $p_{XK} : X \times K \rightarrow \langle 0, 1 \rangle$ and $W : K \times D \rightarrow \mathcal{R}$, find a strategy $q : X \rightarrow D$ which minimizes the **Bayesian risk** of the strategy q

$$R(q) = \sum_{x \in X} \sum_{k \in K} p_{XK}(x, k) \cdot W(k, q(x)).$$

The optimal strategy q, denoted as q^*, is then called **the Bayesian strategy**. The Bayesian risk can be expressed as

$$\begin{aligned} R(q) &= \sum_{x \in X} \sum_{k \in K} p_{K|X}(k|x) \cdot p_X(x) \cdot W(k, q(x)) = \\ &= \sum_{x \in X} p_X(x) \sum_{k \in K} p_{K|X}(k|x) \cdot W(k, q(x)) = \\ &= \sum_{x \in X} p_X(x) \cdot R(q(x), x), \text{ where} \\ R(d, x) &= \sum_{k \in K} p_{K|X}(k|x) \cdot W(k, d) \end{aligned}$$

is the **partial risk**, i.e. the expected penalty for decision d given the observation x.

Bayesian decision task

Given the sets X, K, and D, and functions $p_{XK} : X \times K \rightarrow \langle 0, 1 \rangle$ and $W : K \times D \rightarrow \mathcal{R}$, find a strategy $q : X \rightarrow D$ which minimizes the **Bayesian risk** of the strategy q

$$R(q) = \sum_{x \in X} \sum_{k \in K} p_{XK}(x, k) \cdot W(k, q(x)).$$

The optimal strategy q, denoted as q^*, is then called **the Bayesian strategy**. The Bayesian risk can be expressed as

$$\begin{aligned} R(q) &= \sum_{x \in X} \sum_{k \in K} p_{K|X}(k|x) \cdot p_X(x) \cdot W(k, q(x)) = \\ &= \sum_{x \in X} p_X(x) \sum_{k \in K} p_{K|X}(k|x) \cdot W(k, q(x)) = \\ &= \sum_{x \in X} p_X(x) \cdot R(q(x), x), \text{ where} \\ R(d, x) &= \sum_{k \in K} p_{K|X}(k|x) \cdot W(k, d) \end{aligned}$$

is the **partial risk**, i.e. the expected penalty for decision d given the observation x. The minimization of the Bayesian risk can be formulated as

$$R(q^*) = \min_{q \in Q} R(q) = \sum_{x \in X} p_X(x) \cdot \min_{d \in D} R(d, x),$$

i.e. the Bayesian strategy can be constructed by choosing the decision d^* that minimizes the partial risk for each observation x.

Bayesian strategy characteristics

Bayesian strategy can be derived for infinite X, K and/or D by replacing summation with integration and probability mass function with probability density function in the formulation of Bayesian decision task.

Bayesian strategy is deterministic.

- q provides the same decision $d = q(x)$ for the same x, although k may be different.
- What if we used a randomized strategy q of the form $q(d|x)$, i.e. if the decision d would be chosen randomly using the probability distribution $q(d|x)$?
- The risk of the randomized strategy $q(d|x)$ is equal or greater than the risk of the deterministic Bayesian strategy $q(x)$.

Bayesian strategy divides the probability space to $|D|$ convex cones $C(d)$.

- *Probability space?* Any observation x is mapped to a point in a $|K|$-dimensional linear space (delimited by the positive coordinates) with the coordinates $(p_{X|1}(x|1), p_{X|2}(x|2), \ldots, p_{X|k}(x|k))$.
- *Cone?* Let S be a linear space. Any subspace $C \subset S$ is a *cone* if for each $x \in C$ also $\alpha x \in C$ for any real number $\alpha > 0$.
- *Convex cone?* For any 2 points $x_1 \in C$ and $x_2 \in C$, and for any point x lying on the line between x_1 and x_2, also $x \in C$.
- The individual $C(d)$ are *linearly separable!!!*

Two special cases of the Bayesian decision task

Probability of error when estimating k

- The task is to decide the object state k, i.e. $D = K$.
- The goal is to minimize $Pr(q(x) \neq k)$.
- $Pr(q(x) \neq k) = R(q)$ if

$$W(k, q(x)) = \begin{cases} 0 & \text{if } q(x) = k, \\ 1 & \text{otherwise.} \end{cases}$$

- In this case:

$$\begin{aligned} q(x) &= \arg\min_{d \in D} \sum_{k \in K} p_{XK}(x, k) W(k, d) = \\ &= \arg\max_{d \in D} p_{K|X}(k|x), \end{aligned} \tag{1}$$

  i.e. compute *posterior probabilities* of all states k given the observation x, and decide for the most probable state.
- **Maximum posterior (MAP) estimation**.

Two special cases of the Bayesian decision task

Probability of error when estimating k

- The task is to decide the object state k, i.e. $D = K$.
- The goal is to minimize $Pr(q(x) \neq k)$.
- $Pr(q(x) \neq k) = R(q)$ if

$$W(k, q(x)) = \begin{cases} 0 & \text{if } q(x) = k, \\ 1 & \text{otherwise.} \end{cases}$$

- In this case:

$$\begin{aligned} q(x) &= \arg\min_{d \in D} \sum_{k \in K} p_{XK}(x,k) W(k,d) = \\ &= \arg\max_{d \in D} p_{K|X}(k|x), \end{aligned} \tag{1}$$

  i.e. compute *posterior probabilities* of all states k given the observation x, and decide for the most probable state.
- **Maximum posterior (MAP) estimation**.

Bayesian strategy with the `dontknow` decision

- Using the partial risk $R(d,x) = \sum_{k \in K} p_{K|X}(k|x) \cdot W(k,d)$, for each observation x, we shall provide the decision d minimizing $R(d,x)$.
- But even this optimal $R(d,x)$ may not be sufficiently low, i.e. x does not convey sufficient information for a low-risk decision.
- Let's use $D = K \cup \{\texttt{dontknow}\}$ and define

$$W(k,d) = \begin{cases} 0 & \text{if } d = k, \\ 1 & \text{if } d \neq k \text{ and } d \neq \texttt{dontnow} \\ \epsilon & \text{if } d = \texttt{dontknow}. \end{cases}$$

- In this case:

$$q(x) = \begin{cases} \arg\max_{k \in K} p_{K|X}(k|x) \\ \qquad \text{if } \max_{k \in K} p_{K|X}(k|x) > 1 - \epsilon, \\ \texttt{dontknow} \\ \qquad \text{if } \max_{k \in K} p_{K|X}(k|x) \leq 1 - \epsilon. \end{cases}$$

Limitations of the Bayesian approach

To use the Bayesian approach we need to know:

1. The penalty function W.
2. The *a priori* probabilities of states $p_K(k)$.
3. The conditional probabilities of observations $p_{X|K}(x|k)$.

Limitations of the Bayesian approach

To use the Bayesian approach we need to know:

1. The penalty function W.
2. The *a priori* probabilities of states $p_K(k)$.
3. The conditional probabilities of observations $p_{X|K}(x|k)$.

Penalty function:

- Important: $W(k, d) \in R$
- We cannot use the Bayesian formulation for tasks where identifying the penalties with R substantially deforms the task, i.e. *when the penalties cannot be measured in (or easily transformed to) the same units.*
- How do you compare the following penalties:
  - games, fairy tales:
    loose your horse vs. loose your sword vs. loose your fiancee

Limitations of the Bayesian approach

To use the Bayesian approach we need to know:

1. The penalty function W.
2. The *a priori* probabilities of states $p_K(k)$.
3. The conditional probabilities of observations $p_{X|K}(x|k)$.

Penalty function:

- Important: $W(k, d) \in R$
- We cannot use the Bayesian formulation for tasks where identifying the penalties with R substantially deforms the task, i.e. *when the penalties cannot be measured in (or easily transformed to) the same units.*
- How do you compare the following penalties:
  - games, fairy tales:
    loose your horse vs. loose your sword vs. loose your fiancee
  - system diagnostics, health diagnosis:
    false alarm (costs you some money) vs. overlooked danger (may cost you a human life)

Limitations of the Bayesian approach

To use the Bayesian approach we need to know:

1. The penalty function W.
2. The *a priori* probabilities of states $p_K(k)$.
3. The conditional probabilities of observations $p_{X|K}(x|k)$.

Penalty function:

- Important: $W(k, d) \in R$
- We cannot use the Bayesian formulation for tasks where identifying the penalties with R substantially deforms the task, i.e. *when the penalties cannot be measured in (or easily transformed to) the same units.*
- How do you compare the following penalties:
  - games, fairy tales:
    loose your horse vs. loose your sword vs. loose your fiancee
  - system diagnostics, health diagnosis:
    false alarm (costs you some money) vs. overlooked danger (may cost you a human life)
  - judicial error:
    to convict an innocent (huge harm for 1 innocent person) vs. to free a killer (potential harm to many innocent persons)

Artificial Intelligence
Decision Making
Bayesian DT
- Question
- Bayesian dec. task
- Two special cases
- Limitations

Non-Bayesian DT
Summary

Limitations of the Bayesian approach (cont.)

Prior probabilities of states:

- Probabilities $p_K(k)$
  - may be unknown (then we can determine them by further study), or
  - may not exist at all (if the state k is not random).
- E.g. we observe a plane x and we want to decide if it is an enemy aircraft or not.
  - $p_{X|K}(x|k)$ may be quite complex, but known (it at least exists).
  - $p_K(k)$, however, do not exist—the frequency of enemy plane observation does not converge to any number.

Limitations of the Bayesian approach (cont.)

Prior probabilities of states:

- Probabilities $p_K(k)$
  - may be unknown (then we can determine them by further study), or
  - may not exist at all (if the state k is not random).
- E.g. we observe a plane x and we want to decide if it is an enemy aircraft or not.
  - $p_{X|K}(x|k)$ may be quite complex, but known (it at least exists).
  - $p_K(k)$, however, do not exist—the frequency of enemy plane observation does not converge to any number.

Conditional probabilities of observations:

- Again, probabilities $p_{X|K}(x|k)$ may not be known or may not exist.
- E.g. if we want to decide what characters are on paper cards written by several persons, the observation x of the state k is influenced by an unobservable non-random intervention—by the writer z.
  - We can only talk about $p_{X|K,Z}(x|k,z)$, not about $p_{X|K}(x|k)$.
  - If Z was random and if we knew $p_Z(z)$, than we could compute also $p_{X|K}(x|k)$.

Non-Bayesian Decision Theory

Non-Bayesian decision tasks

When?

- Tasks where W, p_K, or $p_{X|K}$ are not known.
- Even if all the events are random and all probabilities are known, it is sometimes helpful to approach the problem as a non-Bayesian task.
- In practical tasks, it can be more intuitive for the customer to express the desired strategy properties as allowed rates of false positives (false alarm) and false negatives (overlooked danger).

Non-Bayesian decision tasks

When?

- Tasks where W, p_K, or $p_{X|K}$ are not known.
- Even if all the events are random and all probabilities are known, it is sometimes helpful to approach the problem as a non-Bayesian task.
- In practical tasks, it can be more intuitive for the customer to express the desired strategy properties as allowed rates of false positives (false alarm) and false negatives (overlooked danger).

There are several special cases of practically useful non-Bayesian formulations for which the solution is known:

- The strategies that solve these non-Bayesian tasks are of the same form as Bayesian strategies—**they divide the probability space to a set of convex cones**.
- These non-Bayesian tasks can be formulated as linear programs and **solved by linear programming methods**.

Non-Bayesian decision tasks

When?

- Tasks where W, p_K, or $p_{X|K}$ are not known.
- Even if all the events are random and all probabilities are known, it is sometimes helpful to approach the problem as a non-Bayesian task.
- In practical tasks, it can be more intuitive for the customer to express the desired strategy properties as allowed rates of false positives (false alarm) and false negatives (overlooked danger).

There are several special cases of practically useful non-Bayesian formulations for which the solution is known:

- The strategies that solve these non-Bayesian tasks are of the same form as Bayesian strategies—**they divide the probability space to a set of convex cones**.
- These non-Bayesian tasks can be formulated as linear programs and **solved by linear programming methods**.

There are many other non-Bayesian tasks for which the solution is not known yet.

Neyman-Pearson task

Situation:

- Observation $x \in X$, states $k = 1$ (normal), $k = 2$ (dangerous), $K = \{1, 2\}$.
- The probability distribution $p_{X|K}(x|k)$ exists and is known.
- Given the observation x, the task is to decide k, i.e. if the object is in normal or dangerous state.
- In this formulation, $p_K(k)$ and $W(k, d)$ is not needed.

Neyman-Pearson task

Situation:

- Observation $x \in X$, states $k = 1$ (normal), $k = 2$ (dangerous), $K = \{1, 2\}$.
- The probability distribution $p_{X|K}(x|k)$ exists and is known.
- Given the observation x, the task is to decide k, i.e. if the object is in normal or dangerous state.
- In this formulation, $p_K(k)$ and $W(k, d)$ is not needed.

Each strategy q is characterized by 2 numbers:

- Probability of false positive (false alarm):

$$\omega(1) = \sum_{\{x \in X : q(x) = 2\}} p_{X|K}(x|1)$$

- Probability of false negative (overlooked danger):

$$\omega(2) = \sum_{\{x \in X : q(x) = 1\}} p_{X|K}(x|2)$$

Neyman-Pearson task

Situation:

- Observation $x \in X$, states $k = 1$ (normal), $k = 2$ (dangerous), $K = \{1, 2\}$.
- The probability distribution $p_{X|K}(x|k)$ exists and is known.
- Given the observation x, the task is to decide k, i.e. if the object is in normal or dangerous state.
- In this formulation, $p_K(k)$ and $W(k, d)$ is not needed.

Each strategy q is characterized by 2 numbers:

- Probability of false positive (false alarm):

$$\omega(1) = \sum_{\{x \in X: q(x)=2\}} p_{X|K}(x|1)$$

- Probability of false negative (overlooked danger):

$$\omega(2) = \sum_{\{x \in X: q(x)=1\}} p_{X|K}(x|2)$$

Neyman-Pearson task formulation:
Find a strategy q such that

- the probability of overlooked danger (FN) is not larger than a predefined value ϵ, i.e.

$$\omega(2) = \sum_{\{x \in X: q(x)=1\}} p_{X|K}(x|2) \le \epsilon,$$

- and the probability of false alarm (FP) is minimal, i.e.

$$\text{minimize } \omega(1) = \sum_{\{x \in X: q(x)=2\}} p_{X|K}(x|1).$$

Solution: The optimal strategy q^* decides according to the *likelihood ratio*:

$$q^*(x) = \begin{cases} 1 & \text{iff } \frac{p_{X|K}(x|1)}{p_{X|K}(x|2)} > \theta, \\ 2 & \text{iff } \frac{p_{X|K}(x|1)}{p_{X|K}(x|2)} < \theta. \end{cases}$$

Minimax task

Situation:

- Observation $x \in X$, states $k \in K$, $|K| \geq 2$.
- $q : X \rightarrow K$ — given the observation x, the strategy decides the object state k.
- Again, $p_K(k)$ and $W(k, d)$ are not required.

Minimax task

Situation:

- Observation $x \in X$, states $k \in K$, $|K| \geq 2$.
- $q : X \rightarrow K$ — given the observation x, the strategy decides the object state k.
- Again, $p_K(k)$ and $W(k, d)$ are not required.

Each strategy is described by $|K|$ numbers

$$\omega(k) = \sum_{\{x \in X : q(x) \neq k\}} p_{X|K}(x|k),$$

i.e. by the conditional probabilities of a wrong decision under the condition that the true hidden state is k.

Minimax task

Situation:

- Observation $x \in X$, states $k \in K$, $|K| \geq 2$.
- $q : X \rightarrow K$ — given the observation x, the strategy decides the object state k.
- Again, $p_K(k)$ and $W(k,d)$ are not required.

Each strategy is described by $|K|$ numbers

$$\omega(k) = \sum_{\{x \in X : q(x) \neq k\}} p_{X|K}(x|k),$$

i.e. by the conditional probabilities of a wrong decision under the condition that the true hidden state is k.

Minimax task formulation:
Find the optimal strategy q^* as

$$q^* = \underset{q \in Q}{\operatorname{argmin}} \max_{k \in K} \omega(k)$$

Minimax task

Situation:

- Observation $x \in X$, states $k \in K$, $|K| \geq 2$.
- $q : X \to K$ — given the observation x, the strategy decides the object state k.
- Again, $p_K(k)$ and $W(k, d)$ are not required.

Each strategy is described by $|K|$ numbers

$$\omega(k) = \sum_{\{x \in X : q(x) \neq k\}} p_{X|K}(x|k),$$

i.e. by the conditional probabilities of a wrong decision under the condition that the true hidden state is k.

Minimax task formulation:
Find the optimal strategy q^* as

$$q^* = \operatorname*{argmin}_{q \in Q} \max_{k \in K} \omega(k)$$

Solution:

- The solution is of the same form as the Bayesian strategies.
- The solution for the $|K| = 2$ case is similar to the Neyman-Pearson task, with the exception that in minimax task the probability of FN cannot be controlled explicitly.

Wald task

Motivation:

- The Neyman-Pearson task is asymmetric: the prob. of FN is controlled explicitly, while the probability of FP is minimized (but can be quite high).
- Can we find a strategy for which *both* the error probabilities would not exceed a predefined ϵ? No, the demands often cannot be accomplished in the same time.

Wald's relaxation:

- The task is to guess the hidden state k, the strategy can say „I don't know“, i.e., $D = K \cup \{\texttt{dontknow}\}$.
- Strategy of this form is characterized by 4 numbers:
  - the conditional prob. of a wrong decision about the state k,

$$\omega(1) = \sum_{\{x \in X : q(x) = 2\}} p_{X|K}(x|1) \quad \text{and} \quad \omega(2) = \sum_{\{x \in X : q(x) = 1\}} p_{X|K}(x|2),$$

  - the conditional prob. of the `dontknow` decision when the object state is k,

$$\chi(1) = \sum_{\{x \in X : q(x) = \texttt{dontknow}\}} p_{X|K}(x|1) \quad \text{and} \quad \chi(2) = \sum_{\{x \in X : q(x) = \texttt{dontknow}\}} p_{X|K}(x|2).$$

- The requirements $\omega(1) \leq \epsilon$ and $\omega(2) \leq \epsilon$ are no longer contradictory for an arbitrarily small $\epsilon > 0$, since the strategy $\forall x \in X : q(x) = \texttt{dontknow}$ is plausible.
- Each strategy fulfilling $\omega(1) \leq \epsilon$ and $\omega(2) \leq \epsilon$ is then characterized by how often the strategy refuses to decide, i.e. by the number $\max(\chi(1), \chi(2))$.

Wald task (cont.)

Wald task formulation:
Find a strategy q^* which minimizes

$$\max(\chi(1), \chi(2))$$

subject to conditions $\omega(1) \leq \epsilon$ and $\omega(2) \leq \epsilon$.

Solution: The optimal decision is based on the likelihood ratio and 2 thresholds $\theta_1 > \theta_2$:

$$q^*(x) = \begin{cases} 1 & \text{iff } \frac{p_{X|K}(x|1)}{p_{X|K}(x|2)} > \theta_1, \\ 2 & \text{iff } \frac{p_{X|K}(x|1)}{p_{X|K}(x|2)} < \theta_2, \\ \texttt{dontknow} & \text{otherwise.} \end{cases}$$

In [SH12], also the generalization for $|K| > 2$ is given.

[SH12] M. I. Schlesinger and Václav Hlaváč. *Ten Lectures on Statistical and Structural Pattern Recognition (Computational Imaging and Vision)*. Springer, 2002 edition, March 2012.

Linnik tasks

a.k.a. statistical decisions with non-random interventions
a.k.a. evaluations of complex hypotheses.

Previous non-Bayesian tasks did not require

- the a priori probabilities of the states $p_K(k)$, and
- the penalty function $W(k, d)$ to be known.

In Linnik tasks,

- the conditional probabilities $p_{X|K}(x|k)$ do not exist,
- the a priori probabilities $p_K(k)$ may exist (it depends on the fact if the state k is a random variable or not),
- but the conditional probabilities $p_{X|K,Z}(x|k, z)$ do exist, i.e. the random observation x depends not only on the (random or non-random) object state k, but also on a non-random intervention z.

Goal:

- find a strategy that minimizes the probability of incorrect decision in case of the worst intervention z.

See examples in [SH12].

[SH12] M. I. Schlesinger and Václav Hlaváč. *Ten Lectures on Statistical and Structural Pattern Recognition (Computational Imaging and Vision)*. Springer, 2002 edition, March 2012.

Summary of PR

- The aim of PR is to design decision strategies (classifiers) which—given an observation x of an object with a hidden state k—provide a decision d such that this decision making process is optimal with respect to a certain criterion.
- If the statistical properties of (x, k) are completely known, and if we are able to design a suitable penalty function $W(k, d)$, we should solve the task in the *Bayesian framework* and search for the *Bayesian strategy* which optimizes the *Bayesian risk* of the strategy.
  - The minimization of the probability of an error is a special case, the resulting Bayesian strategy decides for the state with the *maximum a posteriori probability*.
- If the statistical properties are known only partially, or are not known at all, or if a reasonable penalty function cannot be constructed, we face a *non-Bayesian task*.
  - Several practically important special cases of non-Bayesian tasks are well-analyzed and solved (Neyman-Pearson, minimax, Wald, . . .).
  - There are plenty of non-Bayesian tasks we can say nothing about.

Summary

Competencies

After this lecture, a student shall be able to . . .

1. explain various views on AI and describe the differences of their personal view of AI;
2. list the fields of science most related to AI;
3. define Bayesian decision task and all its components (decision strategy, risk, penalty function, observation, hidden state, joint probability distribution);
4. solve simple instances of Bayesian decision task by hand, write a computer program solving Bayesian decision tasks;
5. explain features of Bayesian strategy;
6. recognize special cases of Bayesian decision task (minimization of error probability when estimating hidden state, strategy with ”dontknow”decision);
7. describe reasons and examplify situations when the Bayesian approach cannot be used;
8. define and describe examples of non-Bayesian tasks which can be solved to some extent without learning (Neyman-Pearson, minimax, Wald);
9. solve simple instances of the above non-Bayesian decision tasks by hand, write a computer program solving them;
10. define the decision strategy design as a learning from data;
11. describe the differences between Bayesian decision tasks, non-Bayesian decision tasks and decision tasks solved by learning;

References

[DHS01] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. Wiley, New York, 2 edition, 2001.

[RN10] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach (3rd Edition)*. Prentice Hall, 3 edition, 2010.

[SH12] M. I. Schlesinger and Václav Hlaváč. *Ten Lectures on Statistical and Structural Pattern Recognition (Computational Imaging and Vision)*. Springer, 2002 edition, March 2012.