Probabilistic decisions

Tomáš Svoboda, Matěj Hoffmann, and Petr Pošík
thanks to, Daniel Novák and Filip Železný

Vision for Robots and Autonomous Systems, Center for Machine Perception
Department of Cybernetics
Faculty of Electrical Engineering, Czech Technical University in Prague

April 23, 2024

(Re-)introduction uncertainty/probability

- Markov Decision Processes (MDP)/RL – uncertainty about outcome of actions
  - *Sequential* decisions (robot/agent goes from s_0 to s_G)
  - $\pi : \mathcal{S} \rightarrow \mathcal{A}$
- Now: uncertainty associated with states
  - Different states may have different prior probabilities.
  - The states $s \in \mathcal{S}$ are not directly observable.
  - They need to be inferred from features $x \in \mathcal{X}$.
  - *Single (repeated)* decision $\delta : \mathcal{X} \rightarrow \mathcal{S}$

(Re-)introduction uncertainty/probability

- Markov Decision Processes (MDP)/RL – uncertainty about outcome of actions
  - *Sequential* decisions (robot/agent goes from s_0 to s_G)
  - $\pi : \mathcal{S} \rightarrow \mathcal{A}$
- Now: uncertainty associated with states
  - Different states may have different prior probabilities.
  - The states $s \in \mathcal{S}$ are not directly observable.
  - They need to be inferred from features $x \in \mathcal{X}$.
  - *Single (repeated)* decision $\delta : \mathcal{X} \rightarrow \mathcal{S}$

Decision example: Insure or not? (from late 1980s) [5]

Known about HIV testing: HIV test falsely positive only in 1 case out of 1000.
A doctor calls: "Your HIV test is positive, 999/1000 you will die in 10 years. I'm sorry ...".
Insurance company does not want to insure a married couple.

- Was the doctor right?
- Was the insurance company rational?

$\mathcal{S} = \{\text{healthy}, \text{infected}\}$, $\mathcal{X} = \{\text{positive\_test}, \text{negative\_test}\}$

Decision example: Insure or not? (from late 1980s) [5]

Known about HIV testing: HIV test falsely positive only in 1 case out of 1000.
A doctor calls: "Your HIV test is positive, 999/1000 you will die in 10 years. I'm sorry . . . ".
Insurance company does not want to insure a married couple.

- Was the doctor right?
- Was the insurance company rational?

$\mathcal{S} = \{\text{healthy}, \text{infected}\}$, $\mathcal{X} = \{\text{positive\_test}, \text{negative\_test}\}$

Decision example: Insure or not? (from late 1980s) [5]

Known about HIV testing: HIV test falsely positive only in 1 case out of 1000.
A doctor calls: "Your HIV test is positive, 999/1000 you will die in 10 years. I'm sorry ...".
Insurance company does not want to insure a married couple.

- Was the doctor right?
- Was the insurance company rational?

$\mathcal{S} = \{\text{healthy}, \text{infected}\}$, $\mathcal{X} = \{\text{positive\_test}, \text{negative\_test}\}$

Decision example: Insure or not? (from late 1980s) [5]

Known about HIV testing: HIV test falsely positive only in 1 case out of 1000.
A doctor calls: "Your HIV test is positive, 999/1000 you will die in 10 years. I'm sorry ...".
Insurance company does not want to insure a married couple.

- Was the doctor right?
- Was the insurance company rational?

$\mathcal{S} = \{\text{healthy}, \text{infected}\}$, $\mathcal{X} = \{\text{positive\_test}, \text{negative\_test}\}$

What is the probability the man is infected?

A: $\frac{1}{1000}$

B: $\frac{999}{1000}$

C: Don't know yet, more info needed, but less than $\frac{1}{2}$

D: Don't know yet, more info needed, but more than $\frac{1}{2}$

Decision: guilty or not? (people of CA vs Collins, 1968) [5]

- Robbery, LA 1964, fuzzy evidence of the offenders:
  - female, around 65 kg
  - wearing something dark
  - hair of light color, between light and dark blond, in a ponytail
- At the same time, additional evidence close to the crime scene:
  - loud scream, yelling, looking at the this direction ...
  - a woman sitting into a yellow car
  - car starts immediately and passes close to the additional witness
  - a black man with beard and moustache was driving
- No more evidence
- Testimony of both the victim and the witness not unambiguous (didn't recognize suspects)
- Still, the suspects were sentenced to jail.

Decision: guilty or not? (people of CA vs Collins, 1968) [5]

$$
\begin{aligned}
P(\text{yellow car}) &= 1/10 \\
P(\text{man with moustache}) &= 1/4 \\
P(\text{black man with beard}) &= 1/10 \\
P(\text{woman with pony tail}) &= 1/10 \\
P(\text{woman blond hair}) &= 1/3 \\
P(\text{mix race pair in a car}) &= 1/1000
\end{aligned}
$$

Assume (wrong!) mutual indepedence:

$$P(?) = \frac{1}{12,000,000}$$

What probability?

A Convicted pair not guilty.

B A randomly selected pair matches characteristics.

C Some other.

Decision: guilty or not? (people of CA vs Collins, 1968) [5]

$$
\begin{aligned}
P(\text{yellow car}) &= 1/10 \\
P(\text{man with moustache}) &= 1/4 \\
P(\text{black man with beard}) &= 1/10 \\
P(\text{woman with pony tail}) &= 1/10 \\
P(\text{woman blond hair}) &= 1/3 \\
P(\text{mix race pair in a car}) &= 1/1000
\end{aligned}
$$

Assume (wrong!) mutual indepedence:

$$P(?) = \frac{1}{12,000,000}$$

What probability?

A Convicted pair not guilty.

B A randomly selected pair matches characteristics.

C Some other.

Decision: guilty or not? (people of CA vs Collins, 1968) [5]

$$
\begin{aligned}
P(\text{yellow car}) &= 1/10 \\
P(\text{man with moustache}) &= 1/4 \\
P(\text{black man with beard}) &= 1/10 \\
P(\text{woman with pony tail}) &= 1/10 \\
P(\text{woman blond hair}) &= 1/3 \\
P(\text{mix race pair in a car}) &= 1/1000
\end{aligned}
$$

Assume (wrong!) mutual indepedence:

$$P(?) = \frac{1}{12,000,000}$$

What probability?

- A Convicted pair not guilty.
- B A randomly selected pair matches characteristics.
- C Some other.

people of CA vs Collins, 1968, [1]

Computed (wrongly):

$$P_r = P(\text{randomly selected pair matches discussed characteristics}) = \frac{1}{12,000,000}$$

Judge needs:

$$P(\text{a pair matching characteristics is guilty}) = ?$$

$P(\text{randomly selected pair does not match}) = 1 - P_r$

possible/existing pairs in California . . . N

$P(\text{pair will never appear in } N) = P(NA) = (1 - P_r)^N$

$P(\text{pair will appear at least once in } N) = P(ALO) = 1 - P(NA) = 1 - (1 - P_r)^N$

$P(\text{pair will appear exactly once in } N) = P(EO) = NP_r(1 - P_r)^{N-1}$

$P(\text{pair will appear more than once in } N) = P(MTO) = P(ALO) - P(EO)$

$P(MTO|ALO) = \frac{P(MTO,ALO)}{P(ALO)} = \frac{P(MTO)}{P(ALO)}$

$$P(MTO|ALO) = \frac{1 - (1 - P_r)^N - NP_r(1 - P_r)^{N-1}}{1 - (1 - P_r)^N}$$

people of CA vs Collins, 1968, [1]

Computed (wrongly):

$$P_r = P(\text{randomly selected pair matches discussed characteristics}) = \frac{1}{12,000,000}$$

Judge needs:

$$P(\text{a pair matching characteristics is guilty}) = ?$$

$P(\text{randomly selected pair does not match}) = 1 - P_r$

possible/existing pairs in California ... N

$P(\text{pair will never appear in } N) = P(NA) = (1 - P_r)^N$

$P(\text{pair will appear at least once in } N) = P(ALO) = 1 - P(NA) = 1 - (1 - P_r)^N$

$P(\text{pair will appear exactly once in } N) = P(EO) = NP_r(1 - P_r)^{N-1}$

$P(\text{pair will appear more than once in } N) = P(MTO) = P(ALO) - P(EO)$

$P(MTO|ALO) = \frac{P(MTO,ALO)}{P(ALO)} = \frac{P(MTO)}{P(ALO)}$

$$P(MTO|ALO) = \frac{1 - (1 - P_r)^N - NP_r(1 - P_r)^{N-1}}{1 - (1 - P_r)^N}$$

people of CA vs Collins, 1968, [1]

Computed (wrongly):

$$P_r = P(\text{randomly selected pair matches discussed characteristics}) = \frac{1}{12,000,000}$$

Judge needs:

$$P(\text{a pair matching characteristics is guilty}) = ?$$

$P(\text{randomly selected pair does not match}) = 1 - P_r$

possible/existing pairs in California ... N

$P(\text{pair will never appear in } N) = P(NA) = (1 - P_r)^N$

$P(\text{pair will appear at least once in } N) = P(ALO) = 1 - P(NA) = 1 - (1 - P_r)^N$

$P(\text{pair will appear exactly once in } N) = P(EO) = NP_r(1 - P_r)^{N-1}$

$P(\text{pair will appear more than once in } N) = P(MTO) = P(ALO) - P(EO)$

$P(MTO|ALO) = \frac{P(MTO,ALO)}{P(ALO)} = \frac{P(MTO)}{P(ALO)}$

$$P(MTO|ALO) = \frac{1 - (1 - P_r)^N - NP_r(1 - P_r)^{N-1}}{1 - (1 - P_r)^N}$$

people of CA vs Collins, 1968, [1]

Computed (wrongly):

$$P_r = P(\text{randomly selected pair matches discussed characteristics}) = \frac{1}{12,000,000}$$

Judge needs:

$$P(\text{a pair matching characteristics is guilty}) = ?$$

$P(\text{randomly selected pair does not match}) = 1 - P_r$
possible/existing pairs in California ... N

$P(\text{pair will never appear in } N) = P(NA) = (1 - P_r)^N$
$P(\text{pair will appear at least once in } N) = P(ALO) = 1 - P(NA) = 1 - (1 - P_r)^N$
$P(\text{pair will appear exactly once in } N) = P(EO) = NP_r(1 - P_r)^{N-1}$
$P(\text{pair will appear more than once in } N) = P(MTO) = P(ALO) - P(EO)$
$P(MTO|ALO) = \frac{P(MTO,ALO)}{P(ALO)} = \frac{P(MTO)}{P(ALO)}$

$$P(MTO|ALO) = \frac{1 - (1 - P_r)^N - NP_r(1 - P_r)^{N-1}}{1 - (1 - P_r)^N}$$

people of CA vs Collins, 1968, [1]

Computed (wrongly):

$$P_r = P(\text{randomly selected pair matches discussed characteristics}) = \frac{1}{12,000,000}$$

Judge needs:

$$P(\text{a pair matching characteristics is guilty}) = ?$$

$P(\text{randomly selected pair does not match}) = 1 - P_r$
possible/existing pairs in California ... N
$P(\text{pair will never appear in } N) = P(NA) = (1 - P_r)^N$
$P(\text{pair will appear at least once in } N) = P(ALO) = 1 - P(NA) = 1 - (1 - P_r)^N$
$P(\text{pair will appear exactly once in } N) = P(EO) = NP_r(1 - P_r)^{N-1}$
$P(\text{pair will appear more than once in } N) = P(MTO) = P(ALO) - P(EO)$
$P(MTO|ALO) = \frac{P(MTO,ALO)}{P(ALO)} = \frac{P(MTO)}{P(ALO)}$

$$P(MTO|ALO) = \frac{1 - (1 - P_r)^N - NP_r(1 - P_r)^{N-1}}{1 - (1 - P_r)^N}$$

people of CA vs Collins, 1968, [1]

Computed (wrongly):

$$P_r = P(\text{randomly selected pair matches discussed characteristics}) = \frac{1}{12,000,000}$$

Judge needs:

$$P(\text{a pair matching characteristics is guilty}) = ?$$

$P(\text{randomly selected pair does not match}) = 1 - P_r$
possible/existing pairs in California ... N
$P(\text{pair will never appear in } N) = P(NA) = (1 - P_r)^N$
$P(\text{pair will appear at least once in } N) = P(ALO) = 1 - P(NA) = 1 - (1 - P_r)^N$
$P(\text{pair will appear exactly once in } N) = P(EO) = NP_r(1 - P_r)^{N-1}$
$P(\text{pair will appear more than once in } N) = P(MTO) = P(ALO) - P(EO)$
$P(MTO|ALO) = \frac{P(MTO,ALO)}{P(ALO)} = \frac{P(MTO)}{P(ALO)}$

$$P(MTO|ALO) = \frac{1 - (1 - P_r)^N - NP_r(1 - P_r)^{N-1}}{1 - (1 - P_r)^N}$$

people of CA vs Collins, 1968, [1]

Computed (wrongly):

$$P_r = P(\text{randomly selected pair matches discussed characteristics}) = \frac{1}{12,000,000}$$

Judge needs:

$$P(\text{a pair matching characteristics is guilty}) = ?$$

$P(\text{randomly selected pair does not match}) = 1 - P_r$
possible/existing pairs in California . . . N
$P(\text{pair will never appear in } N) = P(NA) = (1 - P_r)^N$
$P(\text{pair will appear at least once in } N) = P(ALO) = 1 - P(NA) = 1 - (1 - P_r)^N$
$P(\text{pair will appear exactly once in } N) = P(EO) = NP_r(1 - P_r)^{N-1}$
$P(\text{pair will appear more than once in } N) = P(MTO) = P(ALO) - P(EO)$
$P(MTO|ALO) = \frac{P(MTO,ALO)}{P(ALO)} = \frac{P(MTO)}{P(ALO)}$

$$P(MTO|ALO) = \frac{1 - (1 - P_r)^N - NP_r(1 - P_r)^{N-1}}{1 - (1 - P_r)^N}$$

people of CA vs Collins, 1968, [1]

Computed (wrongly):

$$P_r = P(\text{randomly selected pair matches discussed characteristics}) = \frac{1}{12,000,000}$$

Judge needs:

$$P(\text{a pair matching characteristics is guilty}) = ?$$

$P(\text{randomly selected pair does not match}) = 1 - P_r$
possible/existing pairs in California $\ldots N$
$P(\text{pair will never appear in } N) = P(NA) = (1 - P_r)^N$
$P(\text{pair will appear at least once in } N) = P(ALO) = 1 - P(NA) = 1 - (1 - P_r)^N$
$P(\text{pair will appear exactly once in } N) = P(EO) = NP_r(1 - P_r)^{N-1}$
$P(\text{pair will appear more than once in } N) = P(MTO) = P(ALO) - P(EO)$
$P(MTO|ALO) = \frac{P(MTO,ALO)}{P(ALO)} = \frac{P(MTO)}{P(ALO)}$

$$P(MTO|ALO) = \frac{1 - (1 - P_r)^N - NP_r(1 - P_r)^{N-1}}{1 - (1 - P_r)^N}$$

people of CA vs Collins, 1968, [1]

Computed (wrongly):

$$P_r = P(\text{randomly selected pair matches discussed characteristics}) = \frac{1}{12,000,000}$$

Judge needs:

$$P(\text{a pair matching characteristics is guilty}) = ?$$

$P(\text{randomly selected pair does not match}) = 1 - P_r$

possible/existing pairs in California . . . N

$P(\text{pair will never appear in } N) = P(NA) = (1 - P_r)^N$

$P(\text{pair will appear at least once in } N) = P(ALO) = 1 - P(NA) = 1 - (1 - P_r)^N$

$P(\text{pair will appear exactly once in } N) = P(EO) = NP_r(1 - P_r)^{N-1}$

$P(\text{pair will appear more than once in } N) = P(MTO) = P(ALO) - P(EO)$

$P(MTO|ALO) = \frac{P(MTO,ALO)}{P(ALO)} = \frac{P(MTO)}{P(ALO)}$

$$P(MTO|ALO) = \frac{1 - (1 - P_r)^N - NP_r(1 - P_r)^{N-1}}{1 - (1 - P_r)^N}$$

people of CA vs Collins, 1968, [1]

Computed (wrongly):

$$P_r = P(\text{randomly selected pair matches discussed characteristics}) = \frac{1}{12,000,000}$$

Judge needs:

$$P(\text{a pair matching characteristics is guilty}) = ?$$

$P(\text{randomly selected pair does not match}) = 1 - P_r$

possible/existing pairs in California . . . N

$P(\text{pair will never appear in } N) = P(NA) = (1 - P_r)^N$

$P(\text{pair will appear at least once in } N) = P(ALO) = 1 - P(NA) = 1 - (1 - P_r)^N$

$P(\text{pair will appear exactly once in } N) = P(EO) = NP_r(1 - P_r)^{N-1}$

$P(\text{pair will appear more than once in } N) = P(MTO) = P(ALO) - P(EO)$

$P(MTO|ALO) = \frac{P(MTO,ALO)}{P(ALO)} = \frac{P(MTO)}{P(ALO)}$

$$P(MTO|ALO) = \frac{1 - (1 - P_r)^N - NP_r(1 - P_r)^{N-1}}{1 - (1 - P_r)^N}$$

people of CA vs Collins, 1968, [1]

Computed (wrongly):

$$P_r = P(\text{randomly selected pair matches discussed characteristics}) = \frac{1}{12,000,000}$$

Judge needs:

$$P(\text{a pair matching characteristics is guilty}) = ?$$

$P(\text{randomly selected pair does not match}) = 1 - P_r$
possible/existing pairs in California $\ldots N$
$P(\text{pair will never appear in } N) = P(NA) = (1 - P_r)^N$
$P(\text{pair will appear at least once in } N) = P(ALO) = 1 - P(NA) = 1 - (1 - P_r)^N$
$P(\text{pair will appear exactly once in } N) = P(EO) = NP_r(1 - P_r)^{N-1}$
$P(\text{pair will appear more than once in } N) = P(MTO) = P(ALO) - P(EO)$
$P(MTO|ALO) = \frac{P(MTO,ALO)}{P(ALO)} = \frac{P(MTO)}{P(ALO)}$

$$P(MTO|ALO) = \frac{1 - (1 - P_r)^N - NP_r(1 - P_r)^{N-1}}{1 - (1 - P_r)^N}$$

people of CA vs Collins, 1968, [1]

Computed (wrongly):

$$P_r = P(\text{randomly selected pair matches discussed characteristics}) = \frac{1}{12,000,000}$$

Judge needs:

$$P(\text{a pair matching characteristics is guilty}) = ?$$

$P(\text{randomly selected pair does not match}) = 1 - P_r$
possible/existing pairs in California ... N
$P(\text{pair will never appear in } N) = P(NA) = (1 - P_r)^N$
$P(\text{pair will appear at least once in } N) = P(ALO) = 1 - P(NA) = 1 - (1 - P_r)^N$
$P(\text{pair will appear exactly once in } N) = P(EO) = NP_r(1 - P_r)^{N-1}$
$P(\text{pair will appear more than once in } N) = P(MTO) = P(ALO) - P(EO)$
$P(MTO|ALO) = \frac{P(MTO,ALO)}{P(ALO)} = \frac{P(MTO)}{P(ALO)}$

$$P(MTO|ALO) = \frac{1 - (1 - P_r)^N - NP_r(1 - P_r)^{N-1}}{1 - (1 - P_r)^N}$$

people of CA vs Collins, 1968, [1]

Computed (wrongly):

$$P_r = P(\text{randomly selected pair matches discussed characteristics}) = \frac{1}{12,000,000}$$

Judge needs:

$$P(\text{a pair matching characteristics is guilty}) = ?$$

$P(\text{randomly selected pair does not match}) = 1 - P_r$
possible/existing pairs in California ... N
$P(\text{pair will never appear in } N) = P(NA) = (1 - P_r)^N$
$P(\text{pair will appear at least once in } N) = P(ALO) = 1 - P(NA) = 1 - (1 - P_r)^N$
$P(\text{pair will appear exactly once in } N) = P(EO) = NP_r(1 - P_r)^{N-1}$
$P(\text{pair will appear more than once in } N) = P(MTO) = P(ALO) - P(EO)$
$P(MTO|ALO) = \frac{P(MTO,ALO)}{P(ALO)} = \frac{P(MTO)}{P(ALO)}$

$$P(MTO|ALO) = \frac{1 - (1 - P_r)^N - NP_r(1 - P_r)^{N-1}}{1 - (1 - P_r)^N}$$

people of CA vs Collins, 1968, [1]

Computed (wrongly):

$$P_r = P(\text{randomly selected pair matches discussed characteristics}) = \frac{1}{12,000,000}$$

Judge needs:

$$P(\text{a pair matching characteristics is guilty}) = ?$$

$P(\text{randomly selected pair does not match}) = 1 - P_r$

possible/existing pairs in California ... N

$P(\text{pair will never appear in } N) = P(NA) = (1 - P_r)^N$

$P(\text{pair will appear at least once in } N) = P(ALO) = 1 - P(NA) = 1 - (1 - P_r)^N$

$P(\text{pair will appear exactly once in } N) = P(EO) = NP_r(1 - P_r)^{N-1}$

$P(\text{pair will appear more than once in } N) = P(MTO) = P(ALO) - P(EO)$

$P(MTO|ALO) = \frac{P(MTO,ALO)}{P(ALO)} = \frac{P(MTO)}{P(ALO)}$

$$P(MTO|ALO) = \frac{1 - (1 - P_r)^N - NP_r(1 - P_r)^{N-1}}{1 - (1 - P_r)^N}$$

people of CA vs Collins, 1968, [1]

Computed (wrongly):

$$P_r = P(\text{randomly selected pair matches discussed characteristics}) = \frac{1}{12,000,000}$$

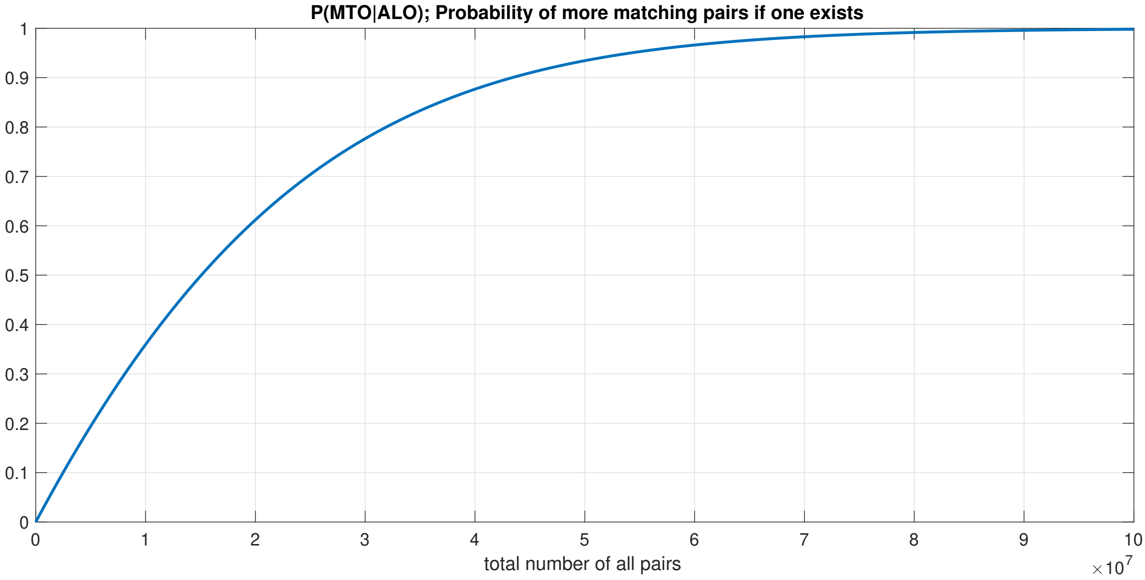
Judge needs:

$$P(\text{a pair matching characteristics is guilty}) = ?$$

$P(\text{randomly selected pair does not match}) = 1 - P_r$
possible/existing pairs in California ... N
$P(\text{pair will never appear in } N) = P(NA) = (1 - P_r)^N$
$P(\text{pair will appear at least once in } N) = P(ALO) = 1 - P(NA) = 1 - (1 - P_r)^N$
$P(\text{pair will appear exactly once in } N) = P(EO) = NP_r(1 - P_r)^{N-1}$
$P(\text{pair will appear more than once in } N) = P(MTO) = P(ALO) - P(EO)$
$P(MTO|ALO) = \frac{P(MTO,ALO)}{P(ALO)} = \frac{P(MTO)}{P(ALO)}$

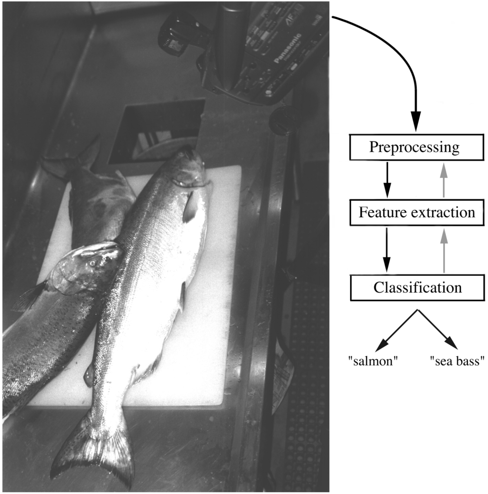
$$P(MTO|ALO) = \frac{1 - (1 - P_r)^N - NP_r(1 - P_r)^{N-1}}{1 - (1 - P_r)^N}$$

$P(MTO|ALO) = f(N)$; people of CA vs Collins, 1968



Probabilistic Classification

Classification example: What's the fish?



- Factory for fish processing
- 2 classes $s_{1,2}$:
  - salmon
  - sea bass
- Features $\vec{x}$: length, width, lightness etc. from a camera

Fish – classification using probability

$$posterior = \frac{likelihood \times prior}{evidence}$$

- Notation for classification problem
  - Classes $s_j \in \mathcal{S}$ (e.g., salmon, sea bass)
  - Features $x_i \in \mathcal{X}$ or feature vectors ($\vec{x_i}$) (also called attributes)
- Optimal classification of $\vec{x}$:

$$\delta^*(\vec{x}) = \arg\max_j P(s_j|\vec{x})$$

- We thus choose the most probable class for a given feature vector .
- Both likelihood and prior are taken into account – recall Bayes rule:

$$P(s_j|\vec{x}) = \frac{P(\vec{x}|s_j)P(s_j)}{P(\vec{x})}$$

- Can we do (classify) better?

Fish – classification using probability

$$posterior = \frac{likelihood \times prior}{evidence}$$

- Notation for classification problem
  - Classes $s_j \in \mathcal{S}$ (e.g., salmon, sea bass)
  - Features $x_i \in \mathcal{X}$ or feature vectors ($\vec{x_i}$) (also called attributes)
- Optimal classification of $\vec{x}$:

$$\delta^*(\vec{x}) = \arg\max_j P(s_j|\vec{x})$$

- We thus choose the most probable class for a given feature vector .
- Both likelihood and prior are taken into account – recall Bayes rule:

$$P(s_j|\vec{x}) = \frac{P(\vec{x}|s_j)P(s_j)}{P(\vec{x})}$$

- Can we do (classify) better?

Decision making under uncertainty

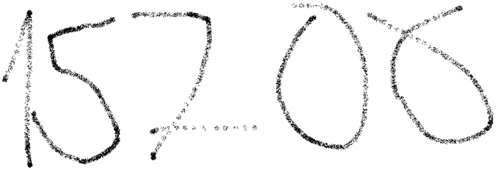
- An important feature of intelligent systems
  - make the best possible decision
  - in uncertain conditions
- **Example**: Take a tram OR subway from A to B?
  - Tram: timetables imply a quicker route, but adherence uncertain.
  - Subway: longer route, but adherence almost certain.
- **Example**: where to route a letter with this ZIP?
  - 15700? 15706? 15200? 15206?
- What is the optimal decision ?
- What is the cost of the decision? What is the associated loss ?
- What is the relation between loss and utility ?

Decision making under uncertainty

- An important feature of intelligent systems
  - make the best possible decision
  - in uncertain conditions
- **Example**: Take a tram OR subway from A to B?
  - Tram: timetables imply a quicker route, but adherence uncertain.
  - Subway: longer route, but adherence almost certain.
- **Example**: where to route a letter with this ZIP?
  - 15700? 15706? 15200? 15206?
- What is the optimal decision ?
- What is the cost of the decision? What is the associated loss ?
- What is the relation between loss and utility ?

Decision making under uncertainty

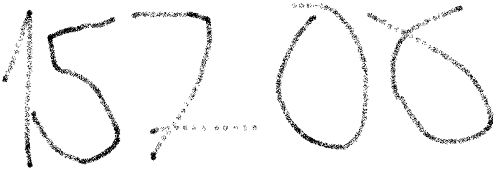
- An important feature of intelligent systems
  - make the best possible decision
  - in uncertain conditions
- **Example**: Take a tram OR subway from A to B?
  - Tram: timetables imply a quicker route, but adherence uncertain.
  - Subway: longer route, but adherence almost certain.
- **Example**: where to route a letter with this ZIP?



  - 15700? 15706? 15200? 15206?
- What is the optimal decision ?
- What is the cost of the decision? What is the associated loss ?
- What is the relation between loss and utility ?

Decision making under uncertainty

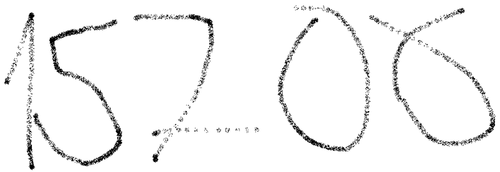
- An important feature of intelligent systems
  - make the best possible decision
  - in uncertain conditions
- **Example**: Take a tram OR subway from A to B?
  - Tram: timetables imply a quicker route, but adherence uncertain.
  - Subway: longer route, but adherence almost certain.
- **Example**: where to route a letter with this ZIP?



  - 15700? 15706? 15200? 15206?
- What is the optimal decision ?
- What is the cost of the decision? What is the associated loss ?
- What is the relation between loss and utility ?

Decision making under uncertainty

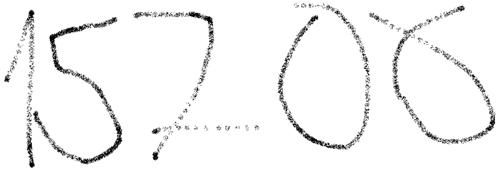
- An important feature of intelligent systems
  - make the best possible decision
  - in uncertain conditions
- **Example**: Take a tram OR subway from *A* to *B*?
  - Tram: timetables imply a quicker route, but adherence uncertain.
  - Subway: longer route, but adherence almost certain.
- **Example**: where to route a letter with this ZIP?



  - 15700? 15706? 15200? 15206?
- What is the optimal decision ?
- What is the cost of the decision? What is the associated loss ?
- What is the relation between loss and utility ?

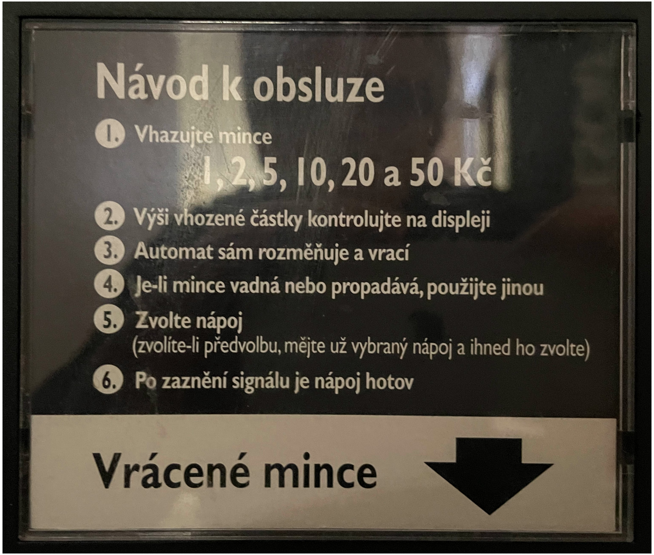
Decision making under uncertainty

- An important feature of intelligent systems
  - make the best possible decision
  - in uncertain conditions
- **Example**: Take a tram OR subway from A to B?
  - Tram: timetables imply a quicker route, but adherence uncertain.
  - Subway: longer route, but adherence almost certain.
- **Example**: where to route a letter with this ZIP?



  - 15700? 15706? 15200? 15206?
- What is the optimal decision ?
- What is the cost of the decision? What is the associated loss ?
- What is the relation between loss and utility ?

Introducing decision loss: Coin recognition



Recognizing/classifying coins: components

- $s \in \{1, 2, 5, 10, 20, 50\}$ – state - the true value
- $x \in \{0.0, 0.1, \cdots, 9.9\}[g]$ – measurement, observation
- $P(s, x)$ joint probability
- $d \in \{1, 2, 5, 10, 20, 50\}$– decision, result of the algorithm

Loss function $\ell(?)$
is a function of:

A s

B s, d

C s, x, d

D d

How many strategies?:

A 100

B 100^6

C 600

D 6^{100}

Strategy $d = \delta(?)$
is a function of:

A x

B s

C s, x

Recognizing/classifying coins: components

- $s \in \{1, 2, 5, 10, 20, 50\}$ – state - the true value
- $x \in \{0.0, 0.1, \cdots, 9.9\}[g]$ – measurement, observation
- $P(s, x)$ joint probability
- $d \in \{1, 2, 5, 10, 20, 50\}$– decision, result of the algorithm

How many strategies?:

A 100

B 100^6

C 600

D 6^{100}

Loss function $\ell(?)$

is a function of:

A s

B s, d

C s, x, d

D d

Strategy $d = \delta(?)$

is a function of:

A x

B s

C s, x

Recognizing/classifying coins: components

- $s \in \{1, 2, 5, 10, 20, 50\}$ – state - the true value
- $x \in \{0.0, 0.1, \cdots, 9.9\}[g]$ – measurement, observation
- $P(s, x)$ joint probability
- $d \in \{1, 2, 5, 10, 20, 50\}$– decision, result of the algorithm

How many strategies?:

A 100

B 100^6

C 600

D 6^{100}

Loss function $\ell(?)$

is a function of:

A s

B s, d

C s, x, d

D d

Strategy $d = \delta(?)$

is a function of:

A x

B s

C s, x

Recognizing/classifying coins: components

- $s \in \{1, 2, 5, 10, 20, 50\}$ – state - the true value
- $x \in \{0.0, 0.1, \cdots, 9.9\}[g]$ – measurement, observation
- $P(s, x)$ joint probability
- $d \in \{1, 2, 5, 10, 20, 50\}$– decision, result of the algorithm

How many strategies?:

A 100

B 100^6

C 600

D 6^{100}

What is the best strategy?

Loss function $\ell(?)$

is a function of:

A s

B s, d

C s, x, d

D d

Strategy $d = \delta(?)$

is a function of:

A x

B s

C s, x

Recognizing/classifying coins: components

- $s \in \{1, 2, 5, 10, 20, 50\}$ – state - the true value
- $x \in \{0.0, 0.1, \cdots, 9.9\}[g]$ – measurement, observation
- $P(s, x)$ joint probability
- $d \in \{1, 2, 5, 10, 20, 50\}$– decision, result of the algorithm

How many strategies?:

A 100

B 100^6

C 600

D 6^{100}

What is the best strategy?

Loss function $\ell(?)$
is a function of:

A s

B s, d

C s, x, d

D d

Strategy $d = \delta(?)$
is a function of:

A x

B s

C s, x

Introducing decision loss: What to cook for dinner [4]

- *Wife is coming back from work. Husband: what to cook for dinner?*
- 3 dishes (decisions) in his repertoire:
  - *nothing* … **don't bother cooking** $\Rightarrow$ no work but makes wife upset
  - *pizza* … **microwave a frozen pizza** $\Rightarrow$ not much work but won't impress
  - *g.T.c.* … **general Tso's chicken** $\Rightarrow$ will make her day, but very laborious
- "Hassle" incurred by the individual options depends on wife's mood.
- For each of the 9 possible situations (3 possible decisions $\times$ 3 possible states), the cost is quantified by a loss function $\ell(d, s)$:

$\ell(s, d)$	$d =$ *nothing*	$d =$ *pizza*	$d =$ *g.T.c.*
$s =$ *good*	0	2	4
$s =$ *average*	5	3	5
$s =$ *bad*	10	9	6

The wife's state of mind is an uncertain state.

Introducing decision loss: What to cook for dinner [4]

- *Wife is coming back from work. Husband: what to cook for dinner?*
- 3 dishes (decisions) in his repertoire:
  - *nothing* . . . **don't bother cooking** $\Rightarrow$ no work but makes wife upset
  - *pizza* . . . **microwave a frozen pizza** $\Rightarrow$ not much work but won't impress
  - *g.T.c.* . . . **general Tso's chicken** $\Rightarrow$ will make her day, but very laborious
- "Hassle" incurred by the individual options depends on wife's mood.
- For each of the 9 possible situations (3 possible decisions $\times$ 3 possible states), the cost is quantified by a loss function $\ell(d, s)$:

$\ell(s, d)$	$d = nothing$	$d = pizza$	$d = g.T.c.$
$s = good$	0	2	4
$s = average$	5	3	5
$s = bad$	10	9	6

The wife's state of mind is an uncertain state.

Introducing decision loss: What to cook for dinner [4]

- *Wife is coming back from work. Husband: what to cook for dinner?*
- 3 dishes (decisions) in his repertoire:
  - *nothing* . . . **don't bother cooking** $\Rightarrow$ no work but makes wife upset
  - *pizza* . . . **microwave a frozen pizza** $\Rightarrow$ not much work but won't impress
  - *g.T.c.* . . . **general Tso's chicken** $\Rightarrow$ will make her day, but very laborious
- "Hassle" incurred by the individual options depends on wife's mood.
- For each of the 9 possible situations (3 possible decisions $\times$ 3 possible states), the cost is quantified by a loss function $\ell(d, s)$:

$\ell(s, d)$	$d = nothing$	$d = pizza$	$d = g.T.c.$
$s = good$	0	2	4
$s = average$	5	3	5
$s = bad$	10	9	6

The wife's state of mind is an uncertain state.

Introducing decision loss: What to cook for dinner [4]

- *Wife is coming back from work. Husband: what to cook for dinner?*
- 3 dishes (decisions) in his repertoire:
  - *nothing* . . . **don't bother cooking** $\Rightarrow$ no work but makes wife upset
  - *pizza* . . . **microwave a frozen pizza** $\Rightarrow$ not much work but won't impress
  - *g.T.c.* . . . **general Tso's chicken** $\Rightarrow$ will make her day, but very laborious
- "Hassle" incurred by the individual options depends on wife's mood.
- For each of the 9 possible situations (3 possible decisions $\times$ 3 possible states), the cost is quantified by a loss function $\ell(d, s)$:

$\ell(s, d)$	$d =$ *nothing*	$d =$ *pizza*	$d =$ *g.T.c.*
$s =$ *good*	0	2	4
$s =$ *average*	5	3	5
$s =$ *bad*	10	9	6

The wife's state of mind is an uncertain state.

Example (cont'd), State uncertain, symptoms, ...

- Husband's experiment. He tells her he accidentally overtaped their wedding video and observes her reaction.
- Anticipates 4 possible reactions:
  - *mild* ... all right, we keep our memories.
  - *irritated* ... how many times do I have to tell you....
  - *upset* ... Why did I marry this guy?
  - *alarming* ... silence
- The reaction is a measurable attribute/symptom ("feature") of the mind state.
- From experience, the husband knows how probable individual reactions are in each state of mind; this is captured by the joint distribution $P(x, s)$.

$P(x, s)$	$x = mild$	$x = irritated$	$x = upset$	$x = alarming$
$s = good$	0.35	0.28	0.07	0.00
$s = average$	0.04	0.10	0.04	0.02
$s = bad$	0.00	0.02	0.05	0.03

Example (cont'd), State uncertain, symptoms, ...

- Husband's experiment. He tells her he accidentally overtaped their wedding video and observes her reaction.
- Anticipates 4 possible reactions:
  - *mild* ... all right, we keep our memories.
  - *irritated* ... how many times do I have to tell you....
  - *upset* ... Why did I marry this guy?
  - *alarming* ... silence
- The reaction is a measurable attribute/symptom ("feature") of the mind state.
- From experience, the husband knows how probable individual reactions are in each state of mind; this is captured by the joint distribution $P(x, s)$.

$P(x, s)$	$x = mild$	$x = irritated$	$x = upset$	$x = alarming$
$s = good$	0.35	0.28	0.07	0.00
$s = average$	0.04	0.10	0.04	0.02
$s = bad$	0.00	0.02	0.05	0.03

Example (cont'd), State uncertain, symptoms, . . .

- Husband's experiment. He tells her he accidentally overtaped their wedding video and observes her reaction.
- Anticipates 4 possible reactions:
  - *mild* . . . all right, we keep our memories.
  - *irritated* . . . how many times do I have to tell you....
  - *upset* . . . Why did I marry this guy?
  - *alarming* . . . silence
- The reaction is a measurable attribute/symptom ("feature") of the mind state.
- From experience, the husband knows how probable individual reactions are in each state of mind; this is captured by the joint distribution $P(x, s)$.

$P(x,s)$	$x = mild$	$x = irritated$	$x = upset$	$x = alarming$
$s = good$	0.35	0.28	0.07	0.00
$s = average$	0.04	0.10	0.04	0.02
$s = bad$	0.00	0.02	0.05	0.03

Example (cont'd), State uncertain, symptoms, ...

- Husband's experiment. He tells her he accidentally overtaped their wedding video and observes her reaction.
- Anticipates 4 possible reactions:
  - *mild* ... all right, we keep our memories.
  - *irritated* ... how many times do I have to tell you....
  - *upset* ... Why did I marry this guy?
  - *alarming* ... silence
- The reaction is a measurable attribute/symptom ("feature") of the mind state.
- From experience, the husband knows how probable individual reactions are in each state of mind; this is captured by the joint distribution $P(x, s)$.

$P(x,s)$	$x =$ *mild*	$x =$*irritated*	$x =$ *upset*	$x =$ *alarming*
$s =$ *good*	0.35	0.28	0.07	0.00
$s =$ *average*	0.04	0.10	0.04	0.02
$s =$ *bad*	0.00	0.02	0.05	0.03

Decision strategy

- Decision strategy : a rule selecting a decision for *any given value* of the measured attribute(s).
- i.e. function $d = \delta(x)$.
- Example of husband's possible strategies:

$\delta(x)$	$x = mild$	$x = irritated$	$x = upset$	$x = alarming$
$\delta_1(x) =$	*nothing*	*nothing*	*pizza*	*g.T.c.*
$\delta_2(x) =$	*nothing*	*pizza*	*g.T.c.*	*g.T.c.*
$\delta_3(x) =$	*g.T.c.*	*g.T.c.*	*g.T.c.*	*g.T.c.*
$\delta_4(x) =$	*nothing*	*nothing*	*nothing*	*nothing*

- How many strategies?
- How to define which strategy is the best? How to sort them by quality?
- Define the risk of a strategy as a mean (expected) loss value.

$$r(\delta) = \sum_x \sum_s \ell(s, \delta(x)) P(x, s)$$

Decision strategy

- Decision strategy : a rule selecting a decision for *any given value* of the measured attribute(s).
- i.e. function $d = \delta(x)$.
- Example of husband's possible strategies:

$\delta(x)$	$x = mild$	$x = irritated$	$x = upset$	$x = alarming$
$\delta_1(x) =$	*nothing*	*nothing*	*pizza*	*g.T.c.*
$\delta_2(x) =$	*nothing*	*pizza*	*g.T.c.*	*g.T.c.*
$\delta_3(x) =$	*g.T.c.*	*g.T.c.*	*g.T.c.*	*g.T.c.*
$\delta_4(x) =$	*nothing*	*nothing*	*nothing*	*nothing*

- How many strategies?
- How to define which strategy is the best? How to sort them by quality?
- Define the risk of a strategy as a mean (expected) loss value .

$$r(\delta) = \sum_x \sum_s \ell(s, \delta(x)) P(x, s)$$

Decision strategy

- Decision strategy : a rule selecting a decision for *any given value* of the measured attribute(s).
- i.e. function $d = \delta(x)$.
- Example of husband's possible strategies:

$\delta(x)$	$x = mild$	$x = irritated$	$x = upset$	$x = alarming$
$\delta_1(x) =$	*nothing*	*nothing*	*pizza*	*g.T.c.*
$\delta_2(x) =$	*nothing*	*pizza*	*g.T.c.*	*g.T.c.*
$\delta_3(x) =$	*g.T.c.*	*g.T.c.*	*g.T.c.*	*g.T.c.*
$\delta_4(x) =$	*nothing*	*nothing*	*nothing*	*nothing*

- How many strategies?
- How to define which strategy is the best? How to sort them by quality?
- Define the risk of a strategy as a mean (expected) loss value .

$$r(\delta) = \sum_x \sum_s \ell(s, \delta(x)) P(x, s)$$

Calculating $r(\delta) = \sum_x \sum_s \ell(s, \delta(x))P(x, s)$

$\ell(s,d)$	$d = nothing$	$d = pizza$	$d = g.T.c.$
$s = good$	0	2	4
$s = average$	5	3	5
$s = bad$	10	9	6

$P(x,s)$	$x = mild$	$x =irritated$	$x = upset$	$x = alarming$
$s = good$	0.35	0.28	0.07	0.00
$s = average$	0.04	0.10	0.04	0.02
$s = bad$	0.00	0.02	0.05	0.03

$\delta(x)$	$x = mild$	$x = irritated$	$x = upset$	$x = alarming$
$\delta_1(x) =$	$nothing$	$nothing$	$pizza$	$g.T.c.$
$\delta_2(x) =$	$nothing$	$pizza$	$g.T.c.$	$g.T.c.$
$\delta_3(x) =$	$g.T.c.$	$g.T.c.$	$g.T.c.$	$g.T.c.$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Do we need to evaluate all possible strategies? $P(x,s) = P(s|x)P(x)$

Calculating $r(\delta) = \sum_x \sum_s \ell(s, \delta(x)) P(x, s)$

$\ell(s,d)$	$d = nothing$	$d = pizza$	$d = g.T.c.$
$s = good$	0	2	4
$s = average$	5	3	5
$s = bad$	10	9	6

$P(x,s)$	$x = mild$	$x = irritated$	$x = upset$	$x = alarming$
$s = good$	0.35	0.28	0.07	0.00
$s = average$	0.04	0.10	0.04	0.02
$s = bad$	0.00	0.02	0.05	0.03

$\delta(x)$	$x = mild$	$x = irritated$	$x = upset$	$x = alarming$
$\delta_1(x) =$	$nothing$	$nothing$	$pizza$	$g.T.c.$
$\delta_2(x) =$	$nothing$	$pizza$	$g.T.c.$	$g.T.c.$
$\delta_3(x) =$	$g.T.c.$	$g.T.c.$	$g.T.c.$	$g.T.c.$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Do we need to evaluate all possible strategies? $P(x,s) = P(s|x)P(x)$

Calculating $r(\delta) = \sum_x \sum_s \ell(s, \delta(x)) P(x, s)$

$\ell(s,d)$	$d = nothing$	$d = pizza$	$d = g.T.c.$
$s = good$	0	2	4
$s = average$	5	3	5
$s = bad$	10	9	6

$P(x,s)$	$x = mild$	$x = irritated$	$x = upset$	$x = alarming$
$s = good$	0.35	0.28	0.07	0.00
$s = average$	0.04	0.10	0.04	0.02
$s = bad$	0.00	0.02	0.05	0.03

$\delta(x)$	$x = mild$	$x = irritated$	$x = upset$	$x = alarming$
$\delta_1(x) =$	$nothing$	$nothing$	$pizza$	$g.T.c.$
$\delta_2(x) =$	$nothing$	$pizza$	$g.T.c.$	$g.T.c.$
$\delta_3(x) =$	$g.T.c.$	$g.T.c.$	$g.T.c.$	$g.T.c.$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Do we need to evaluate all possible strategies? $P(x,s) = P(s|x)P(x)$

Calculating $r(\delta) = \sum_x \sum_s \ell(s, \delta(x)) P(x, s)$

$\ell(s, d)$	$d =$ *nothing*	$d =$ *pizza*	$d =$ *g.T.c.*
$s =$ *good*	0	2	4
$s =$ *average*	5	3	5
$s =$ *bad*	10	9	6

$P(x, s)$	$x =$ *mild*	$x =$ *irritated*	$x =$ *upset*	$x =$ *alarming*
$s =$ *good*	0.35	0.28	0.07	0.00
$s =$ *average*	0.04	0.10	0.04	0.02
$s =$ *bad*	0.00	0.02	0.05	0.03

$\delta(x)$	$x =$ *mild*	$x =$ *irritated*	$x =$ *upset*	$x =$ *alarming*
$\delta_1(x) =$	*nothing*	*nothing*	*pizza*	*g.T.c.*
$\delta_2(x) =$	*nothing*	*pizza*	*g.T.c.*	*g.T.c.*
$\delta_3(x) =$	*g.T.c.*	*g.T.c.*	*g.T.c.*	*g.T.c.*
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Do we need to evaluate all possible strategies? $P(x, s) = P(s|x)P(x)$

Calculating $r(\delta) = \sum_x \sum_s \ell(s, \delta(x)) P(x, s)$

$\ell(s, d)$	$d =$ *nothing*	$d =$ *pizza*	$d =$ *g.T.c.*
$s =$ *good*	0	2	4
$s =$ *average*	5	3	5
$s =$ *bad*	10	9	6

$P(x, s)$	$x =$ *mild*	$x =$ *irritated*	$x =$ *upset*	$x =$ *alarming*
$s =$ *good*	0.35	0.28	0.07	0.00
$s =$ *average*	0.04	0.10	0.04	0.02
$s =$ *bad*	0.00	0.02	0.05	0.03

$\delta(x)$	$x =$ *mild*	$x =$ *irritated*	$x =$ *upset*	$x =$ *alarming*
$\delta_1(x) =$	*nothing*	*nothing*	*pizza*	*g.T.c.*
$\delta_2(x) =$	*nothing*	*pizza*	*g.T.c.*	*g.T.c.*
$\delta_3(x) =$	*g.T.c.*	*g.T.c.*	*g.T.c.*	*g.T.c.*
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Do we need to evaluate all possible strategies? $P(x, s) = P(s|x)P(x)$

Bayes optimal strategy

- The Bayes optimal strategy : one minimizing mean risk.

$$\delta^* = \arg\min_{\delta} r(\delta)$$

- From $P(x, s) = P(s|x)P(x)$ (Bayes rule), we have

$$r(\delta) = \sum_x \sum_s \ell(s, \delta(x))P(x, s) = \sum_s \sum_x \ell(s, \delta(x))P(s|x)P(x)$$

$$= \sum_x P(x) \underbrace{\sum_s \ell(s, \delta(x))P(s|x)}_{\text{Conditional risk}}$$

- The optimal strategy is obtained by minimizing the conditional risk *separately* for each x:

$$\delta^*(x) = \arg\min_{d} \sum_s \ell(s, d)P(s|x)$$

Optimal strategy: $\delta^*(x) = \arg\min_d \sum_s \ell(s,d)P(s|x)$

$\ell(s,d)$	$d =$ *nothing*	$d =$ *pizza*	$d =$ *g.T.c.*
$s =$ *good*	0	2	4
$s =$ *average*	5	3	5
$s =$ *bad*	10	9	6

$P(x,s)$	$x =$ *mild*	$x =$*irritated*	$x =$ *upset*	$x =$ *alarming*
$s =$ *good*	0.35	0.28	0.07	0.00
$s =$ *average*	0.04	0.10	0.04	0.02
$s =$ *bad*	0.00	0.02	0.05	0.03

$\delta(x)$	$x =$ *mild*	$x =$ *irritated*	$x =$ *upset*	$x =$ *alarming*
$\delta^*(x) =$	??	??	??	??

Statistical decision making: wrapping up

- **Given**:
  - A set of possible states : $\mathcal{S}$
  - A set of possible decisions : $\mathcal{D}$
  - A loss function $l : \mathcal{D} \times \mathcal{S} \to \Re$
  - The range $\mathcal{X}$ of the attribute
  - Distribution $P(x, s)$, $x \in \mathcal{X}, s \in \mathcal{S}$.
- **Define**:
  - Strategy : function $\delta : \mathcal{X} \to \mathcal{D}$
  - Risk of strategy δ : $r(\delta) = \sum_x \sum_s \ell(s, \delta(x)) P(x, s)$
- **Bayes problem**:
  - Goal: find the optimal strategy $\delta^* = \arg\min_\delta r(\delta)$
  - Solution: $\delta^*(x) = \arg\min_d \sum_s \ell(s, d) P(s|x)$ (for each x)

A special case - Bayesian *classification*

- Bayesian classification is a special case of statistical decision theory:
  - Attribute vector $\vec{x} = (x_1, x_2, \dots)$: pixels 1, 2, ...
  - **State set $\mathcal{S}$ = decision set $\mathcal{D} = \{0, 1, \dots 9\}$.**
  - State = actual class, Decision = recognized class
  - Loss function:

$$\ell(s,d) = \begin{cases} 0, & d = s \\ 1, & d \neq s \end{cases}$$

$$\delta^*(\vec{x}) = \arg\min_d \sum_s \underbrace{\ell(s,d)}_{0 \text{ if } d=s} P(s|\vec{x}) = \arg\min_d \sum_{s \neq d} P(s|\vec{x})$$

Obviously $\sum_s P(s|\vec{x}) = 1$, then:

$$P(d|\vec{x}) + \sum_{s \neq d} P(s|\vec{x}) = 1$$

Inserting into above:

$$\delta^*(\vec{x}) = \arg\min_d [1 - P(d|\vec{x})] = \arg\max_d P(d|\vec{x})$$

A special case - Bayesian *classification*

- Bayesian classification is a special case of statistical decision theory:
  - Attribute vector $\vec{x} = (x_1, x_2, \dots)$: pixels 1, 2, ...
  - **State set $\mathcal{S}$ = decision set $\mathcal{D} = \{0, 1, \dots 9\}$.**
  - State = actual class, Decision = recognized class
  - Loss function:

$$\ell(s,d) = \begin{cases} 0, & d = s \\ 1, & d \neq s \end{cases}$$

$$\delta^*(\vec{x}) = \arg\min_d \sum_s \underbrace{\ell(s,d)}_{0 \text{ if } d=s} P(s|\vec{x}) = \arg\min_d \sum_{s \neq d} P(s|\vec{x})$$

Obviously $\sum_s P(s|\vec{x}) = 1$, then:

$$P(d|\vec{x}) + \sum_{s \neq d} P(s|\vec{x}) = 1$$

Inserting into above:

$$\delta^*(\vec{x}) = \arg\min_d [1 - P(d|\vec{x})] = \arg\max_d P(d|\vec{x})$$

A special case - Bayesian *classification*

- Bayesian classification is a special case of statistical decision theory:
  - Attribute vector $\vec{x} = (x_1, x_2, \dots)$: pixels 1, 2, ...
  - **State set $\mathcal{S}$ = decision set $\mathcal{D} = \{0, 1, \dots 9\}$.**
  - State = actual class, Decision = recognized class
  - Loss function:

$$\ell(s, d) = \begin{cases} 0, & d = s \\ 1, & d \neq s \end{cases}$$

$$\delta^*(\vec{x}) = \arg\min_d \sum_s \underbrace{\ell(s, d)}_{0 \text{ if } d=s} P(s|\vec{x}) = \arg\min_d \sum_{s \neq d} P(s|\vec{x})$$

Obviously $\sum_s P(s|\vec{x}) = 1$, then:

$$P(d|\vec{x}) + \sum_{s \neq d} P(s|\vec{x}) = 1$$

Inserting into above:

$$\delta^*(\vec{x}) = \arg\min_d [1 - P(d|\vec{x})] = \arg\max_d P(d|\vec{x})$$

A special case - Bayesian *classification*

- Bayesian classification is a special case of statistical decision theory:
  - Attribute vector $\vec{x} = (x_1, x_2, \dots)$: pixels 1, 2, ...
  - **State set $\mathcal{S}$ = decision set $\mathcal{D} = \{0, 1, \dots 9\}$.**
  - State = actual class, Decision = recognized class
  - Loss function:

$$\ell(s,d) = \begin{cases} 0, & d = s \\ 1, & d \neq s \end{cases}$$

$$\delta^*(\vec{x}) = \arg\min_d \sum_s \underbrace{\ell(s,d)}_{0 \text{ if } d=s} P(s|\vec{x}) = \arg\min_d \sum_{s \neq d} P(s|\vec{x})$$

Obviously $\sum_s P(s|\vec{x}) = 1$, then:

$$P(d|\vec{x}) + \sum_{s \neq d} P(s|\vec{x}) = 1$$

Inserting into above:

$$\delta^*(\vec{x}) = \arg\min_d [1 - P(d|\vec{x})] = \arg\max_d P(d|\vec{x})$$

A special case - Bayesian *classification*

- Bayesian classification is a special case of statistical decision theory:
  - Attribute vector $\vec{x} = (x_1, x_2, \dots)$: pixels 1, 2, ...
  - **State set $\mathcal{S}$ = decision set $\mathcal{D} = \{0, 1, \dots 9\}$.**
  - State = actual class, Decision = recognized class
  - Loss function:

$$\ell(s,d) = \begin{cases} 0, & d = s \\ 1, & d \neq s \end{cases}$$

$$\delta^*(\vec{x}) = \arg\min_d \sum_s \underbrace{\ell(s,d)}_{0 \text{ if } d=s} P(s|\vec{x}) = \arg\min_d \sum_{s \neq d} P(s|\vec{x})$$

Obviously $\sum_s P(s|\vec{x}) = 1$, then:

$$P(d|\vec{x}) + \sum_{s \neq d} P(s|\vec{x}) = 1$$

Inserting into above:

$$\delta^*(\vec{x}) = \arg\min_d [1 - P(d|\vec{x})] = \arg\max_d P(d|\vec{x})$$

References I

Further reading: Chapter 13 and 14 of [7] (Chapters 12 and 13 in [8]). Books [2] (for this lecture, read Chapter 1) and [3] are classical textbooks in the field of pattern recognition and machine learning. Interesting insights into how people think and interact with probabilities are presented in [5] (in Czech as [6]).

[1] People v. collins.
https://law.justia.com/cases/california/supreme-court/2d/68/319.html.

[2] Christopher M. Bishop.
Pattern Recognition and Machine Learning.
Springer Science+Bussiness Media, New York, NY, 2006.
https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf.

References II

[3] Richard O. Duda, Peter E. Hart, and David G. Stork.
Pattern Classification.
John Wiley & Sons, 2nd edition, 2001.

[4] Zdeněk Kotek, Petr Vysoký, and Zdeněk Zdráhal.
Kybernetika.
SNTL, 1990.

[5] Leonard Mlodinow.
The Drunkard's Walk. How Randomness Rules Our Lives.
Vintage Books, 2008.

[6] Leonard Mlodinow.
Život je jen náhoda. Jak náhoda ovlivňuje naše životy.
Slovart, 2009.

References III

[7] Stuart Russell and Peter Norvig.
Artificial Intelligence: A Modern Approach.
Prentice Hall, 3rd edition, 2010.
http://aima.cs.berkeley.edu/.

[8] Stuart Russell and Peter Norvig.
Artificial Intelligence: A Modern Approach.
Prentice Hall, 4th edition, 2021.