Probabilistic classification

Tomáš Svoboda and Matěj Hoffmann
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

May 2, 2022

(Re-)introduction uncertainty/probability

- Markov Decision Processes (MDP) – uncertainty about outcome of actions
- Now: uncertainty may be also associated with states
  - Different states may have different prior probabilities.
  - The states $s \in \mathcal{S}$ may not be directly observable.
  - They need to be inferred from features $x \in \mathcal{X}$.
- This is addressed by the rules of probability *(such as Bayes theorem)* and leads on to
  - Bayesian classification
  - Bayesian decision making

red and **b**lue boxes; **a**pples and **o**ranges

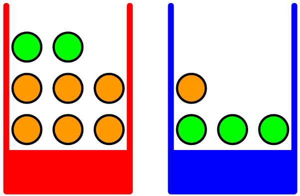
Dark warehouse, color not directly observable. Getting stats:

- Pick box at random, pull it out to the light.
- Pick a fruit from the box, at random.

Random variables: B box color, F fruit kind

Probability example: Picking fruits

- red box: 2 apples, 6 oranges
- blue box: 3 apples, 1 orange

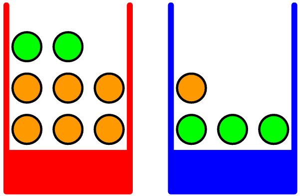


- Scenario: Pick a box—say red box in 40% cases. *Then* pick a fruit at random.
- (Frequent) questions:
  - What is the overall probability that the selection procedure will pick an apple?
  - Given that we have chosen an orange, what is the probability that it was from the blue box?

Example from Chapter 1.2 [1]

Picking fruits. What is the probability that . . . ?

- red box: 2 apples, 6 oranges
- blue box: 3 apples, 1 orange



Procedure: Pick a box (say red box in 40% cases), then pick a fruit at random.

Quiz 1: What is the probability that the selection procedure will pick an apple?

A: 11/20

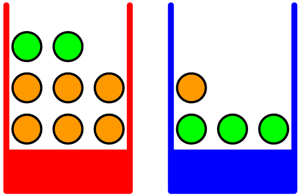
B: 6/8

C: 1/2

D: Different value.

Picking fruits. What is the probability that ...?

- red box: 2 apples, 6 oranges
- blue box: 3 apples, 1 orange



Procedure: Pick a box (say red box in 40% cases), then pick a fruit at random.

Quiz 2: Given that we have chosen an orange, what is the probability that it was from the blue box?

A: $1/4$

B: $3/5$

C: $1/3$

D: Different value.

Rules of probability and notation I

- random variables X, Y
- x_i where $i = 1, ..., M$ – values taken by variable X
- y_j where $j = 1, ..., L$ – values taken by variable Y
- $P(X = x_i, Y = y_i)$ – probability that X takes the value x_i *and* Y takes y_i – joint probability
- $P(X = x_i)$ – probability that X takes the value x_i
- Sum rule of probability :
  - $P(X = x_i) = \sum_{j=1}^{L} P(X = x_i, Y = y_j)$
  - $P(X = x_i)$ is sometimes called marginal probability – obtained by marginalizing / summing out the other variables
  - general rule, compact notation: $P(X) = \sum_Y P(X, Y)$

Rules of probability and notation II

- Conditional probability : $P(Y = y_j | X = x_i)$
- Product rule of probability :
  - $P(X = x_i, Y = y_i) = P(Y = y_j | X = x_i) P(X = x_i)$
  - general rule, compact notation: $P(X, Y) = P(Y|X)P(X)$
- Bayes theorem :
  - from $P(X, Y) = P(Y, X)$ and product rule

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

$$\begin{aligned} P(disease|symptoms) &= \frac{P(symptoms|disease) \times P(disease)}{P(symptoms)} \\ posterior &= \frac{likelihood \times prior}{evidence} \end{aligned}$$

- Independence : $P(X, Y) = P(X)P(Y)$

Rules of probability and notation II

- Conditional probability : $P(Y = y_j | X = x_i)$
- Product rule of probability :
  - $P(X = x_i, Y = y_i) = P(Y = y_j | X = x_i) P(X = x_i)$
  - general rule, compact notation: $P(X, Y) = P(Y|X)P(X)$
- Bayes theorem :
  - from $P(X, Y) = P(Y, X)$ and product rule

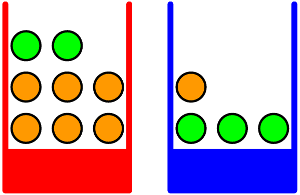
$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

$$\begin{aligned} P(disease|symptoms) &= \frac{P(symptoms|disease) \times P(disease)}{P(symptoms)} \\ posterior &= \frac{likelihood \times prior}{evidence} \end{aligned}$$

- Independence : $P(X, Y) = P(X)P(Y)$

Boxes and Fruits: posterior? likelihood? prior? evidence?

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



Connect with lines:

- posterior
  after observation
- likelihood
  of an observation
- prior
  before observation
- evidence
  total observations

- $P(B)$
- $P(F)$
- $P(F \mid B)$
- $P(B \mid F)$

Decision example: Insure or not? (from late 1980s) [4]

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?

What the doctor (and the company) knew:

- HIV test falsely positive only in 1 case out of 1000.

Decision example: Insure or not? (from late 1980s) [4]

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?

What the doctor (and the company) knew:

- HIV test falsely positive only in 1 case out of 1000.

Decision example: Insure or not? (from late 1980s) [4]

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?

What the doctor (and the company) knew:

- HIV test falsely positive only in 1 case out of 1000.

Decision example: Insure or not? (from late 1980s) [4]

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?

What the doctor (and the company) knew:

- HIV test falsely positive only in 1 case out of 1000.

Decision example: Insure or not? (from late 1980s) [4]

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?

What the doctor (and the company) knew:

- HIV test falsely positive only in 1 case out of 1000.

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 example: Insure or not? (from late 1980s) [4]

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?

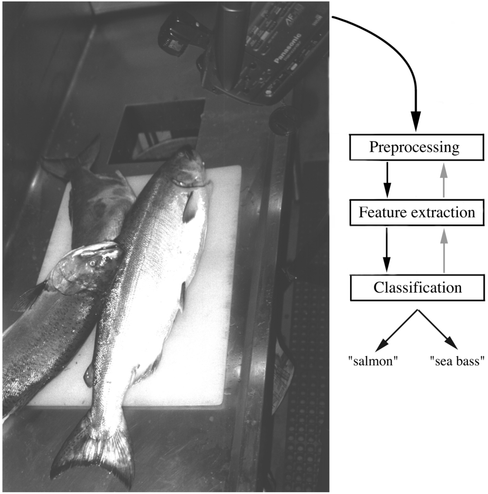
What the doctor (and the company) knew:

- HIV test falsely positive only in 1 case out of 1000.
- Heterosexual male, has family, no drugs, no risk behavior.

Decision: guilty or not? (people of CA vs Collins, 1968) [4]

- 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.

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

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

- 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

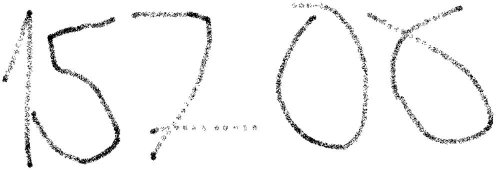
- An important feature of intelligent systems
  - make the best possible decision
  - in uncertain conditions
- **Example**: Take a tram OR subway from A to B?
  - Tram: timetables imply a quicker route, but adherence uncertain.
  - Subway: longer route, but adherence almost certain.
- **Example**: where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

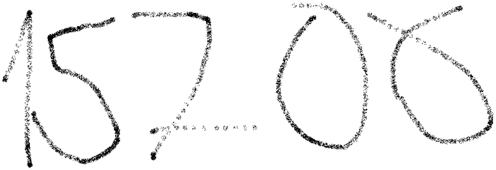
- An important feature of intelligent systems
  - make the best possible decision
  - in uncertain conditions
- **Example**: Take a tram OR subway from A to B?
  - Tram: timetables imply a quicker route, but adherence uncertain.
  - Subway: longer route, but adherence almost certain.
- **Example**: where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

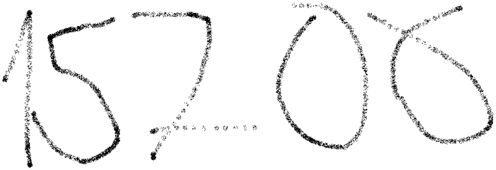
- An important feature of intelligent systems
  - make the best possible decision
  - in uncertain conditions
- **Example**: Take a tram OR subway from A to B?
  - Tram: timetables imply a quicker route, but adherence uncertain.
  - Subway: longer route, but adherence almost certain.
- **Example**: where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

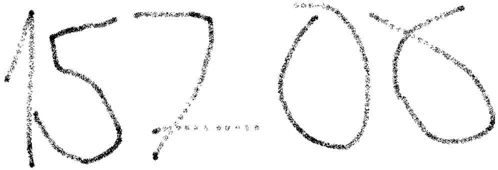
- An important feature of intelligent systems
  - make the best possible decision
  - in uncertain conditions
- **Example**: Take a tram OR subway from A to B?
  - Tram: timetables imply a quicker route, but adherence uncertain.
  - Subway: longer route, but adherence almost certain.
- **Example**: where 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: What to cook for dinner [3]

- *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 $l(d, s)$:

$l(s, d)$	$d = nothing$	$d = pizza$	$d = g.T.c.$
$s = good$	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 [3]

- *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 $l(d, s)$:

$l(s, d)$	$d = nothing$	$d = pizza$	$d = g.T.c.$
$s = good$	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 [3]

- *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 $l(d, s)$:

$l(s, d)$	$d = nothing$	$d = pizza$	$d = g.T.c.$
$s = good$	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 [3]

- *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 $l(d,s)$:

$l(s,d)$	$d = nothing$	$d = pizza$	$d = g.T.c.$
$s = good$	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.

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

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 .

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

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 .

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

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

$l(s,d)$	$d = nothing$	$d = pizza$	$d = g.T.c.$
$s = good$	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 $r(\delta) = \sum_x \sum_s l(s, \delta(x)) P(x, s)$

$l(s, d)$	$d =$ *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 $r(\delta) = \sum_x \sum_s l(s, \delta(x)) P(x, s)$

$l(s,d)$	$d = nothing$	$d = pizza$	$d = g.T.c.$
$s = good$	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 $r(\delta) = \sum_x \sum_s l(s, \delta(x)) P(x, s)$

$l(s, d)$	$d =$ *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 $r(\delta) = \sum_x \sum_s l(s, \delta(x)) P(x, s)$

$l(s,d)$	$d =$ *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

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

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

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

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

$l(s,d)$	$d =$ *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 : $r(\delta) = \sum_x \sum_s l(s, \delta(x)) P(x, s)$
- **Bayes problem**:
  - Goal: find the optimal strategy $\delta^* = \arg\min_\delta r(\delta)$
  - Solution: $\delta^*(x) = \arg\min_d \sum_s l(s, d) P(s|x)$ (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:

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

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

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:

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

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

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:

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

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

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:

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

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

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:

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

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

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 [6] (Chapters 12 and 13 in [7]). Books [1] (for this lecture, read Chapter 1) and [2] 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 [4] (in Czech as [5]).

[1] 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.

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

References II

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

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

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

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

References III

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