Probabilistic classification

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

Department of Cybernetics, Vision for Robotics and Autonomous Systems,
Center for Machine Perception (CMP)

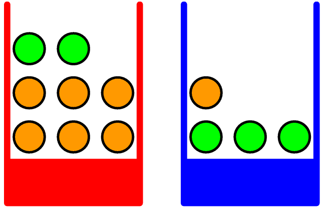
May 25, 2018

(Re-)introduction uncertainty/probability

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

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 the box we chose was the blue one?

Example from Chapter 1.2 [1]

Example serves for probability recap (sum, product rules, conditional probabilities, Bayes)
Random variables:

- Identity of the box B, two possible values r, b
- Identity of the fruit F, two possible values a, o

Info about picking a box $P(B = r) = 0.4$ and $P(B = r) = 0.6$.
Conditional probabilities, given box selected:
$P(o|r) = 3/4, P(a|r) = 1/4, P(o|b) = 1/4, P(a|b) = 3/4$.
Answering questions:

- $P(F = a) = P(a|r)P(r) + P(a|b)P(b) = 11/20$
- $P(B = b|F = o) = P(b|o)$

$$P(b|o) = \frac{P(o|b)P(b)}{P(o)} = \frac{P(o|b)P(b)}{P(o|b)P(b) + P(o|r)P(r)} = 1/3$$

$P(B)$ prior probability - *before* we observe the fruit; $P(B|F)$ - aposteriori probability - *after* we observe the fruit.

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: $P(X) = \sum_{Y} P(X, Y)$

This and the following slides are just to formally recap what we learned when discussion boxes and fruits

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

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

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

Boxes and fruits: prior (before observation) - $P(B)$, likelihood (of observation) - $P(F|B)$, evidence (total observations) $P(F)$, posterior (after observation) $P(B|F)$.

Think about these terms, it helps to understand and remember.

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

A doctor calls: "999/1000 you die in 10 years, I'm sorry ...". Insurance company does not want to insure 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, have family, no drugs, no risk behavior.

Equations/formulas are simple but not easy to (fully) understand. doctor: $P(\text{positive test}|\text{healthy})$ but this the likelihood which we learn before the pacient diagnosis (classification). More interesting and important is to know: $P(\text{healthy}|\text{positive test})$. Think about 10000 samples of heterosexual males, family, ... Statistically 1 HIV positive inbetween. Assume $P(\text{negative test}|\text{healthy}) \to 0$. 1 person HIV positive will be tested positive, but also 10 other healthy persons will be tested positive. Hence $P(\text{healthy}|textpositivetest) = 10/11$!. The fact that a disease is rare matters a lot!

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

A doctor calls: “999/1000 you die in 10 years, I'm sorry …”. Insurance company does not want to insure 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, have family, no drugs, no risk behavior.

Equations/formulas are simple but not easy to (fully) understand. doctor: $P(\text{positive test}|\text{healthy})$ but this the likelihood which we learn before the pacient diagnosis (classification). More interesting and important is to know: $P(\text{healthy}|\text{positive test})$. Think about 10000 samples of heterosexual males, family, ... Statistically 1 HIV positive inbetween. Assume $P(\text{negative test}|\text{healthy}) \rightarrow 0$. 1 person HIV positive will be tested positive, but also 10 other healthy persons will be tested positive. Hence $P(\text{healthy}|textpositivetest) = 10/11!$. The fact that a disease is rare matters a lot!

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

A doctor calls: "999/1000 you die in 10 years, I'm sorry ...". Insurance company does not want to insure 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, have family, no drugs, no risk behavior.

Equations/formulas are simple but not easy to (fully) understand. doctor: $P(\text{positive test}|\text{healthy})$ but this the likelihood which we learn before the pacient diagnosis (classification). More interesting and important is to know: $P(\text{healthy}|\text{positive test})$. Think about 10000 samples of heterosexual males, family, ... Statistically 1 HIV positive inbetween. Assume $P(\text{negative test}|\text{healthy}) \rightarrow 0$. 1 person HIV positive will be tested positive, but also 10 other healthy persons will be tested positive. Hence $P(\text{healthy}|textpositivetest) = 10/11!$. The fact that a disease is rare matters a lot!

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

A doctor calls: "999/1000 you die in 10 years, I'm sorry ...". Insurance company does not want to insure 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, have family, no drugs, no risk behavior.

Equations/formulas are simple but not easy to (fully) understand. doctor: $P(\text{positive test}|\text{healthy})$ but this the likelihood which we learn before the pacient diagnosis (classification). More interesting and important is to know: $P(\text{healthy}|\text{positive test})$. Think about 10000 samples of heterosexual males, family, ... Statistically 1 HIV positive inbetween. Assume $P(\text{negative test}|\text{healthy}) \to 0$. 1 person HIV positive will be tested positive, but also 10 other healthy persons will be tested positive. Hence $P(\text{healthy}|textpositivetest) = 10/11!$. The fact that a disease is rare matters a lot!

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

A doctor calls: "999/1000 you die in 10 years, I'm sorry . . . ". Insurance company does not want to insure married couple.

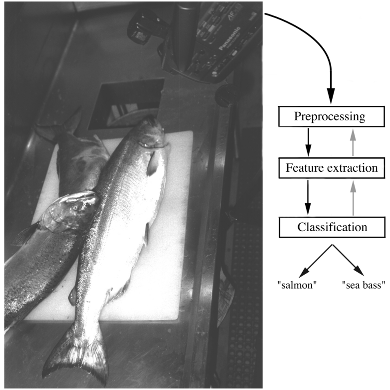
- Was the doctor right?
- Was the insurance company rational?

What the doctor (and the company) knew:

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

Equations/formulas are simple but not easy to (fully) understand. doctor: $P(\text{positive test}|\text{healthy})$ but this the likelihood which we learn before the pacient diagnosis (classification). More interesting and important is to know: $P(\text{healthy}|\text{positive test})$. Think about 10000 samples of heterosexual males, family, ... Statistically 1 HIV positive inbetween. Assume $P(\text{negative test}|\text{healthy}) \to 0$. 1 person HIV positive will be tested positive, but also 10 other healthy persons will be tested positive. Hence $P(\text{healthy}|textpositivetest) = 10/11!$. The fact that a disease is rare matters a lot!

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 $s_j \in S$ (e.g., salmon, sea bass)
  - Features $x_i \in X$ 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:

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

Fish – classification using probability

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

- Notation for classification problem
  - Classes $s_j \in S$ (e.g., salmon, sea bass)
  - Features $x_i \in X$ 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:

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

Bayes classification in practice

- Usually we are not given $P(s|\vec{x})$
- It has to be estimated from already classified examples – training data
- For discrete $\vec{x}$, training examples $(\vec{x}_1, s_1), (\vec{x}_2, s_2), \ldots (\vec{x}_l, s_l)$
  - so-called i.i.d (independent, identically distributed) multiset
  - every $(\vec{x}_i, s)$ is drawn independently from $P(\vec{x}, s)$
- Without knowing anything about the distribution, a non-parametric estimate:

$$P(s|\vec{x}) \approx \frac{\#\text{ examples where } \vec{x}_i = \vec{x} \text{ and } s_i = s}{\#\text{ examples where } \vec{x}_i = \vec{x}}$$

- Hard in practice:
  - To reliably estimate $P(s|\vec{x})$, the number of examples grows exponentially with the number of elements of $\vec{x}$.
    - e.g. with the number of pixels in images
    - curse of dimensionality
    - denominator often 0

Why hard? Way too many various $\vec{x}$. Think about simple binary 10×10 image - $\vec{x}$ contains $0, 1$, position matters. What is the total number of unique images? Think binary, 1×8 binary image?

Bayes classification in practice

- Usually we are not given $P(s|\vec{x})$
- It has to be estimated from already classified examples – training data
- For discrete $\vec{x}$, training examples $(\vec{x}_1, s_1), (\vec{x}_2, s_2), \ldots (\vec{x}_l, s_l)$
  - so-called i.i.d (independent, identically distributed) multiset
  - every $(\vec{x}_i, s)$ is drawn independently from $P(\vec{x}, s)$
- Without knowing anything about the distribution, a non-parametric estimate:

$$P(s|\vec{x}) \approx \frac{\#\text{ examples where } \vec{x}_i = \vec{x} \text{ and } s_i = s}{\#\text{ examples where } \vec{x}_i = \vec{x}}$$

- Hard in practice:
  - To reliably estimate $P(s|\vec{x})$, the number of examples grows exponentially with the number of elements of $\vec{x}$.
    - e.g. with the number of pixels in images
    - curse of dimensionality
    - denominator often 0

Why hard? Way too many various $\vec{x}$. Think about simple binary 10×10 image - $\vec{x}$ contains $0, 1$, position matters. What is the total number of unique images? Think binary, 1×8 binary image?

Bayes classification in practice

- Usually we are not given $P(s|\vec{x})$
- It has to be estimated from already classified examples – training data
- For discrete $\vec{x}$, training examples $(\vec{x}_1, s_1), (\vec{x}_2, s_2), \ldots (\vec{x}_l, s_l)$
  - so-called i.i.d (independent, identically distributed) multiset
  - every $(\vec{x}_i, s)$ is drawn independently from $P(\vec{x}, s)$
- Without knowing anything about the distribution, a non-parametric estimate:

$$P(s|\vec{x}) \approx \frac{\#\text{ examples where } \vec{x}_i = \vec{x} \text{ and } s_i = s}{\#\text{ examples where } \vec{x}_i = \vec{x}}$$

- Hard in practice:
  - To reliably estimate $P(s|\vec{x})$, the number of examples grows exponentially with the number of elements of $\vec{x}$.
    - e.g. with the number of pixels in images
    - curse of dimensionality
    - denominator often 0

Why hard? Way too many various $\vec{x}$. Think about simple binary 10×10 image - $\vec{x}$ contains $0, 1$, position matters. What is the total number of unique images? Think binary, 1×8 binary image?

Bayes classification in practice

- Usually we are not given $P(s|\vec{x})$
- It has to be estimated from already classified examples – training data
- For discrete $\vec{x}$, training examples $(\vec{x}_1, s_1), (\vec{x}_2, s_2), \ldots (\vec{x}_l, s_l)$
  - so-called i.i.d (independent, identically distributed) multiset
  - every $(\vec{x}_i, s)$ is drawn independently from $P(\vec{x}, s)$
- Without knowing anything about the distribution, a non-parametric estimate:

$$P(s|\vec{x}) \approx \frac{\#\text{ examples where } \vec{x}_i = \vec{x} \text{ and } s_i = s}{\#\text{ examples where } \vec{x}_i = \vec{x}}$$

- Hard in practice:
  - To reliably estimate $P(s|\vec{x})$, the number of examples grows exponentially with the number of elements of $\vec{x}$.
    - e.g. with the number of pixels in images
    - curse of dimensionality
    - denominator often 0

Why hard? Way too many various $\vec{x}$. Think about simple binary 10×10 image - $\vec{x}$ contains $0, 1$, position matters. What is the total number of unique images? Think binary, 1×8 binary image?

Naïve Bayes classification

- For efficient classification we must thus rely on additional assumptions.
- In the exceptional case of statistical independence between $\vec{x}$ components for each class s it holds

$$P(\vec{x}|s) = P(x[1]|s) \cdot P(x[2]|s) \cdot \ldots$$

- Use simple Bayes law and maximize:

$$P(s|\vec{x}) = \frac{P(\vec{x}|s)P(s)}{P(\vec{x})} = \frac{P(s)}{P(\vec{x})} P(x[1]|s) \cdot P(x[2]|s) \cdot \ldots =$$

- No combinatorial curse in estimating $P(s)$ and $P(x[i]|s)$ separately for each i and s.
- No need to estimate $P(\vec{x})$. (Why?)
- $P(s)$ may be provided apriori.
- naïve $=$ when used despite statistical dependence

Why naïve at all? Consider $N-$ dimensional space, $8-bit$ values. Instead of problem 8^N we have $8 \times N$ problem.

Think about statistical independence. Example1: person's weight and height. Are they independent? Example2: pixel values in images.

Decision making under uncertainty

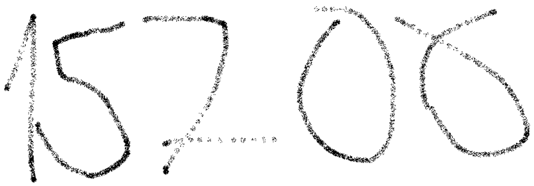
- An important feature of intelligent systems
  - make the best possible decision
  - in uncertain conditions.
- **Example**: Take a tram OR subway from A to B?
  - Tram: timetables imply a quicker route, but adherence uncertain.
  - Subway: longer route, but adherence almost certain.
- **Example**: where to route a letter with this ZIP?
  - 15700? 15706? 15200? 15206?
- What is the "optimal decision"?
- Both examples fall into the same framework.

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 ?
- Both examples fall into the same framework.

Decision making under uncertainty

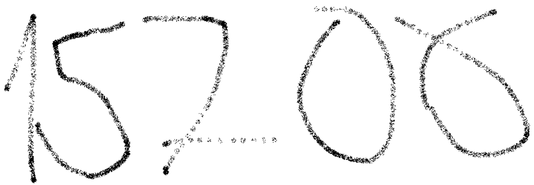
- An important feature of intelligent systems
  - make the best possible decision
  - in uncertain conditions.
- **Example**: Take a tram OR subway from A to B?
  - Tram: timetables imply a quicker route, but adherence uncertain.
  - Subway: longer route, but adherence almost certain.
- **Example**: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



  - 15700? 15706? 15200? 15206?
- What is the optimal decision ?
- Both examples fall into the same framework.

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 ?
- Both examples fall into the same framework.

Example: What to cook for a dinner [3]

- *Wife 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 wife's feeling
- For each of the 9 possible situation (3 possible decisions $\times$ 3 possible states) the hassle 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

Her state of mind is an uncertain state.

Was the state known the decision would be simple.

Example: What to cook for a dinner [3]

- *Wife 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 wife's feeling
- For each of the 9 possible situation (3 possible decisions $\times$ 3 possible states) the hassle 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

Her state of mind is an uncertain state.

Was the state known the decision would be simple.

Example: What to cook for a dinner [3]

- *Wife 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 wife's feeling
- For each of the 9 possible situation (3 possible decisions $\times$ 3 possible states) the hassle 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

Her state of mind is an uncertain state.

Was the state known the decision would be simple.

Example: What to cook for a dinner [3]

- *Wife 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 wife's feeling
- For each of the 9 possible situation (3 possible decisions $\times$ 3 possible states) the hassle 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

Her state of mind is an uncertain state.

Was the state known the decision would be simple.

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

- 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 ("feature") of the mind state.
- From experience, the husband knows how individual reactions are probable 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

Joint distibution. Husband tried similar experiment multiple times, gathered some evidence ...

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

- 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 ("feature") of the mind state.
- From experience, the husband knows how individual reactions are probable 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

Joint distibution. Husband tried similar experiment multiple times, gathered some evidence ...

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

- 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 ("feature") of the mind state.
- From experience, the husband knows how individual reactions are probable 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

Joint distibution. Husband tried similar experiment multiple times, gathered some evidence ...

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

- 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 ("feature") of the mind state.
- From experience, the husband knows how individual reactions are probable 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

Joint distibution. Husband tried similar experiment multiple times, gathered some evidence ...

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

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

Overall, $3^4 = 81$ possible strategies (3 possible decisions for each of the 4 possible attribute values).

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

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

Overall, $3^4 = 81$ possible strategies (3 possible decisions for each of the 4 possible attribute values).

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

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

Overall, $3^4 = 81$ possible strategies (3 possible decisions for each of the 4 possible attribute values).

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

$P(x,s)$	$x = mild$	$x = irritated$	$x = upset$	$x = alarming$
$s = good$	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

$\delta(x)$	$x = mild$	$x = irritated$	$x = upset$	$x = alarming$
$\delta_1(x) =$	*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*

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

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

Risk depend on strategy(decisions). Strategy(decisions) depends on obeservation. Loss comines decision and state. The total weighted average is weighted by joint probability of observation and state. Calculate $r(\delta_1)$ and $r(\delta_2)$, what is better strategy?

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

$P(x,s)$	$x = mild$	$x = irritated$	$x = upset$	$x = alarming$
$s = good$	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

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

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

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

Risk depend on strategy(decisions). Strategy(decisions) depends on obeservation. Loss comines decision and state. The total weighted average is weighted by joint probability of observation and state. Calculate $r(\delta_1)$ and $r(\delta_2)$, what is better strategy?

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

$P(x,s)$	$x = mild$	$x = irritated$	$x = upset$	$x = alarming$
$s = good$	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

$\delta(x)$	$x = mild$	$x = irritated$	$x = upset$	$x = alarming$
$\delta_1(x) =$	*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*

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

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

Risk depend on strategy(decisions). Strategy(decisions) depends on obeservation. Loss comines decision and state. The total weighted average is weighted by joint probability of observation and state. Calculate $r(\delta_1)$ and $r(\delta_2)$, what is better strategy?

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*
$\delta^*(x) =$	??	??	??	??

We need to recompute the table of joint probability $P(s,x)$ into table of conditional probabilies $P(s|x)$. Having the table of all $P(s|x)$ we just mechanically instert into equation in the slide title.

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 $\delta^* = \arg\min_{\delta \in \Delta} r(\delta)$
  - Solution: $\delta^*(x) = \arg\min_d \sum_s l(s, d)P(s|x)$

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:

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

- Classification as opposed to Decision
- Loss function simply counts errors (misclassifications)
- We consider alle errors equally painful!
- More example during the lab ...
- The final result is not that surprising, is it?

A special case - Bayesian *classification*

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

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

- Classification as opposed to Decision
- Loss function simply counts errors (misclassifications)
- We consider alle errors equally painful!
- More example during the lab ...
- The final result is not that surprising, is it?

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:

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

- Classification as opposed to Decision
- Loss function simply counts errors (misclassifications)
- We consider alle errors equally painful!
- More example during the lab ...
- The final result is not that surprising, is it?

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:

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

- Classification as opposed to Decision
- Loss function simply counts errors (misclassifications)
- We consider alle errors equally painful!
- More example during the lab ...
- The final result is not that surprising, is it?

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:

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

- Classification as opposed to Decision
- Loss function simply counts errors (misclassifications)
- We consider alle errors equally painful!
- More example during the lab ...
- The final result is not that surprising, is it?

References I

Further reading: Chapter 13 and 14 of [6]. Books [1] and [2] are classical textbooks in the field of pattern recognition and machine learning. An 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.

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

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

References II

[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/.