Basics of Probability Theory

Fabrizio Riguzzi

•Acknowledgments: some slides from
- Andrew Moore's tutorials http://www.autonlab.org/tutorials/
- Irina Rish and Moninder Singh's tutorial http://www.research.ibm.com/people/r/rish/

Summary

- Definition
- Joint Probability
- Conditional probability
- Random Variables
- Continuous Random Variables

Uncertainty

- Reasoning requires simplifications:
  - Birds fly
  - Smoke suggests fire
- Treatment of exceptions
- How to reason from uncertain knowledge?

How to Perform Inference?

- Use non-numerical techniques
  - Logicist: non monotonic logic
- Assign to each proposition a numerical measure of uncertainty
  - Neo-probabilist: use probability theory
  - Neo-calculist: use other theories:
    - fuzzy logic
    - certainty factors
    - Dempster-Shafer

Probability Theory

- A: Proposition,
  - Ex: A=The coin will land heads
- P(A): probability of A
- Frequentist approach: probability as relative frequency
  - Repeated random experiments (possible worlds)
  - P(A) is the fraction of experiments in which A is true
- Bayesian approach: probability as a degree of belief
- Example: B=burglary tonight

Frequentist Approach

- A=The coin will land heads
- 100 throws, for each throw we record whether A is tre
- Results:

A	¬A
61	49

$P(A)=\frac{61}{100}=0.61=61\%$ $\qquad P(\neg A)=\frac{49}{100}=0.49=49\%$

Frequentist Approach

- H="having a headache"
- 400 patients

H	¬H
40	360

$$P(A)=\frac{40}{400}=0.1=10\%$$

Frequentist Approach

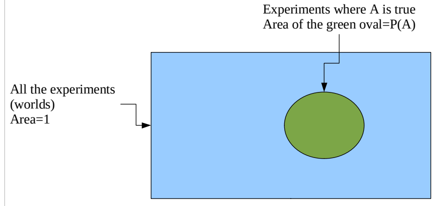
- F="having the flu"
- 400 patients

F	¬F
10	390

$$P(A)=\frac{10}{400}=0.025=2.5\%$$

Visualizing the Frequentist Approach

- P(A)



Axioms of Probability Theory

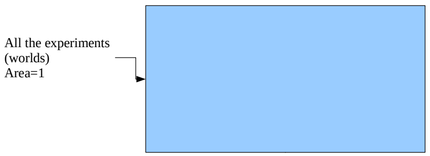
$$0 \le P(A) \le 1$$

$$P(\text{Sure Proposition})=1$$

$$P(A \vee B)=P(A)+P(B)$$
if A and B are mutually exclusive

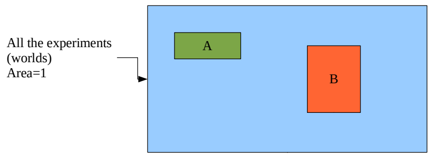
Visualizing the Axioms

- 0<=P(A)<=1: the area cannot get smaller than 0 and larger than 1



Visualizing the Axioms

- P(A v B)=P(A)+P(B) if they are mutually exclusive
- Mutually exclusive=> no world in common=> non overlapping=> the area is the sum



Joint Probability

- Consider the events
  - H="having a headache"
  - F="having the flu"
- **Joint event**: H∧F="having a headache and the flu"
- Also written as H,F
- **Joint probability**: P(H∧F)=P(H,F)
- Frequentist interpretation:
  - P(H∧F)=P(H,F) is the fraction of experiments (in this case patients) where both H and F holds

Joint Probability

- Example: 400 patients

	H	¬H	
F	5	5	10
¬F	35	355	390
	40	360	400

$$P(H,F)=\frac{5}{400}=0.0125=1.25\% \qquad P(H,\neg F)=\frac{35}{400}=0.0875=8.75\%$$

$$P(\neg H,F)=\frac{5}{400}=0.0125=1.25\% \qquad P(\neg H,\neg F)=\frac{355}{400}=0.8875=88.75\%$$

Probability Rules

- Any event A can be written as the or of two disjoint events (A ∧ B) and (A ∧ ¬B)

$$P(A)=P(A,B)+P(A,\neg B)$$ **marginalization/ sum rule**

- In general, if B_i i=1,2,...,n is a set of exhaustive and mutually exclusive propositions

$$P(A)=\sum_i P(A,B_i)$$

- Moreover, picking A=true:

$$P(B)+P(\neg B)=1$$

Conditional Probabilities

- P(A|B)= belief of A given that I know B
- Definition according to the frequentist approach:

$$P(A|B)=\frac{P(A,B)}{P(B)}$$

- Interpretation: fraction of the worlds where B is true in which also A is true
- If P(B)=0 than p(A|B) is not defined

Example

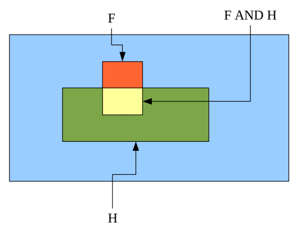
- H="having a headache", F="having the flu"
- P(H|F)="having a headache given that I have the flu"

	H	¬H	
F	5	5	10
¬F	35	355	390
	40	360	400

$$P(H|F)=\frac{P(H,F)}{P(F)}=\frac{\frac{5}{400}}{\frac{10}{400}}=\frac{5}{10}=0.5=50\%$$

- P(H|F)=0.5: H and F are rare but if I have the flu, it is probable that I have a headache

Example



Product Rule

- From

$$P(A|B)=\frac{P(A,B)}{P(B)}$$

- We can derive

$$P(A,B)=P(A|B)P(B) \qquad \textbf{product rule}$$

- In the Bayesian approach, the conditional probability is fundamental and the joint probability is derived with the product rule.

Bayes Theorem

- Relationship between P(A|B) and P(B|A):
- P(A,B)=P(A|B)P(B), P(A,B)=P(B|A)P(A) =>

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

- P(A): **prior probability**
- P(A|B): **posterior probability** (after learning B)

Example

- H=”having a headache”
- F=”having the flu”
- P(H)=0.1 P(F)=0.025
- P(H|F)=0.5

$$P(F|H)=\frac{P(H|F)P(F)}{P(H)}=\frac{0.5*0.025}{0.1}=0.125$$

- Knowing that I have a headache, the probability of having the flu raises to 1/8

Chain Rule

- n events $E_1,\ldots,E_n$
- Joint event $(E_1,\ldots,E_n)$

$$P(E_n,\ldots,E_1)=P(E_n|E_{n-1}\ldots,E_1)P(E_{n-1},\ldots,E_1)$$
$$P(E_{n-1},\ldots,E_1)=P(E_{n-1}|E_{n-2}\ldots,E_1)P(E_{n-2},\ldots,E_1)$$
$$\ldots$$

- Chain rule:

$$P(E_n,\ldots,E_1)=P(E_n|E_{n-1}\ldots,E_1)\ldots P(E_2|E_1)P(E_1)=\prod_{i=1}^{n}P(E_i|E_{i-1},\ldots E_1)$$

Multivalued Hypothesis

- Propositions can be seen as binary variables, i.e. variables taking values true or false
  - Burglary B: true or false
- Generalization: multivalued variables
  - Semaphore S, values: green, yellow, red
  - Propositions are a special case with two values

Discrete Random Variables

- Variable V, values v_i i=1,...,n
- V is also called a **discrete random variable**
- V=v_i is a proposition
- Propositions V=v_i i=1,...,n exhaustive and mutually exclusive
- P(v_i) stands for P(V=v_i)
- V is described by the set {P(v_i)|i=1,...,n}, the **probability distribution** of V, indicated with P(V)

Notation

- We indicate with v a generic value of V
- Set or vector of variables **V**, values **v**

Marginalization

- Multivalued variables A and B
- b_i i=1,...,n values of B

$$P(a)=\sum_i P(a,b_i)$$

- Or

$$P(a)=\sum_b P(a,b)$$

- In general

$$P(\mathbf{x})=\sum_{\mathbf{y}} P(\mathbf{x},\mathbf{y})$$ **sum rule or marginalization**

Conjunctions

- A conjunction of two Boolean variables can be considered as a single random variable that takes 4 values
- Example:
  - H and F, values {true, false}
  - (H,F), values {(true,true),(true,false),(false,true),(false,false)}

Conditional Probabilities

- P(a|b)= belief of A=a given that I know B=b
- Relation to P(a,b)

$$P(a,b)=P(a|b)P(b)$$ **product rule**

$$P(a|b)=\frac{P(a,b)}{P(b)}$$

- Bayes theorem

$$P(a|b)=\frac{P(b|a)\,p(a)}{P(b)}$$

Continuous Random Variables

- A multivalued variable V that takes values from a real interval [a,b] is called a **continuous random variable**
- P(V=v)=0, we want to compute $P(c \leq V \leq d)$
- V is described by a **probability density function** ρ: [a,b]→[0,1]
- ρ(v) is such that

$$P(c \leq V \leq d)=\int_c^d \rho(v)\,dv$$

Properties of Continuous Random Variables

- The same as those of discrete random variables where summation is replaced by integration:
- Marginalization (sum rule)

$$\rho(\mathbf{x})=\int \rho(\mathbf{x},\mathbf{y})\,d\,\mathbf{y}$$

- Conditional probability (product rule)

$$\rho(\mathbf{x},\mathbf{y})=\rho(\mathbf{x}|\mathbf{y})\rho(\mathbf{y})$$

....

Mixed Distribution

- We can have a conjunction of discrete and continuous variables
- Joint: if one of the variables is continuous, the joint is a density:
  - X discrete, Y continuous: $\rho(x,y)$
- Conditional joint:
  - X discrete, Y continuous: $P(x|y)$
  - X discrete, Y continuous, Z discrete: $\rho(x,y|z)$