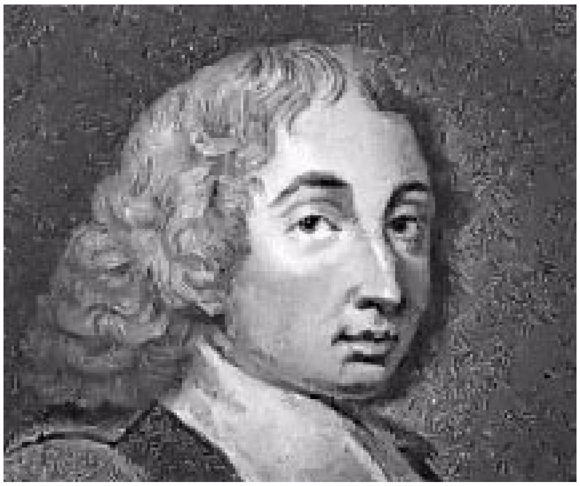
Introduction to Probability and Reasoning Under Uncertainty



Blaise Pascal
1623-1662

Pondering the probabilities of winning and losing in games of chance.

History of Probability Theory

- Modern theory of probability
  - *Pascal* and *Pierre de Fermat* (1601-1665)
  - Study of permutations and combinations
  - Foundation for the study of statistics
  - Began by trying understand games of chance
    - *E.g.,* how many times must a pair of dice must be thrown before the chance that a six will appear is 50-50?
    - *E.g.,* if two players of equal ability competing in a match to be won by the first to win ten games, are obliged to suspend play when one player has won five games, and the other seven games, how should the stakes be divided?

- Applications of probability theory
  - Widely used in the physical, biological, and social sciences to interpret the outcome of experiments
  - Provides the foundations for the theory of quantum mechanics and molecular genetics.
  - Used to analyze the performance of computer algorithms that employ randomization.
  - Provides the basis in actuarial science for computing insurance risks and premiums.
  - More generally it provides the basis for rational decision making under uncertainty.

Reasoning Under Uncertainty

- Uncertainty about
  - the truth of a proposition
    - Is the elevator working?
    - Is Microsoft stock overvalued?
  - the occurrence of an event
    - Will the elevator stop on the third floor?
    - Will Netscape rebound in the second quarter?

- Sources of uncertainty
  - natural sources of stochasticity
    - radioactive decay
  - ignorance, incomplete information
    - causes of cancer and many other diseases
  - complexity, feasibility, and utility
    - summary statistics suffice for many applications

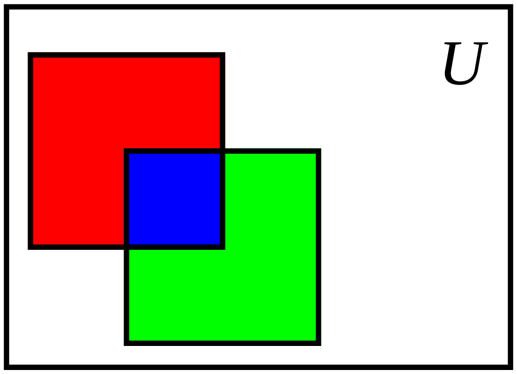
Probabilities

- Examples of probabilities
  - There’s a 70% chance of rain this evening.
  - With probability 0.6 Apple will close up for the day.
  - The degree of my belief in passing the exam is 0.9.

- Interpretations of probabilities
  - *Frequency interpretation*
    - A probability is a property of a set of events.
  - *Subjective interpretation*
    - A probability is an expression of a person’s degree of belief regarding the truth of a proposition or the occurrence of an event.

Introduction to Probability Theory

- Interpretation in terms of finite sets



$$A = \color{red}{\blacksquare} + \color{blue}{\blacksquare}$$

$$B = \color{green}{\blacksquare} + \color{blue}{\blacksquare}$$

$$A - B = \color{red}{\blacksquare}$$

$$A \cap B = \color{blue}{\blacksquare}$$

$$A \cup B = \color{red}{\blacksquare} + \color{blue}{\blacksquare} + \color{green}{\blacksquare}$$

$$\Pr(A) = \frac{|A|}{|U|}$$

$$\Pr(A \wedge B) = \frac{|A \cap B|}{|U|}$$

$$\Pr(A \vee B) = \frac{|A \cup B|}{|U|}$$

$$\Pr(\neg A) = \frac{|U - A|}{|U|}$$

$$\Pr(A \supset B) = \Pr(\neg A \vee B) = \frac{|(U - A) \cup B|}{|U|}$$

- where in this case A denotes both the set A and the event of selecting an item from the set A.

Probability Distributions

- Now let X represent a *random variable* and Ω_X represent the set of possible values for X.
  - Ω_X is called the *sample space* for the variable X.
  - Assuming $\Omega_X = \{0, 1\}$ then $\Pr(X = 0)$ denotes the probability that the variable X takes on the value 0.

- We say $\Pr(X)$ is a *probability distribution* if

$$\forall v \in \Omega_X, 0 \le \Pr(X = v) \le 1 \quad \text{and}$$

$$1 = \sum_{v \in \Omega_X} \Pr(X = v)$$

- If X and Y are random variables then X, Y is a random variable with *joint space* $\Omega_X \times \Omega_Y$.
  - If $\Pr(X, Y)$ is a distribution, then it is called the *joint distribution* for $\{X, Y\}$.
  - By standard convention

$$\Pr(X = u \wedge Y = v) = \Pr(X = u, Y = v)$$

  - With a little bit of thought you should be able to reinterpret the formulas on the previous slide, *e.g.,* $\Pr(A)$ and $\Pr(A \wedge B)$, in terms of random variables, sample spaces, and probability distributions.

Conditional Probabilities

- In practice every probability is conditioned on some context-setting or *conditioning* event.
  - Instead of writing $\Pr(A)$ we should have written $\Pr(A|U)$ and interpreted it as "the probability that we select from A given that we select from U."
  - $\Pr(A|U)$ is called a *conditional probability distribution.*
  - Some times we'll write $\Pr(A)$ for $\Pr(A|U)$ to avoid stating the obvious, but other times explicitly indicating the conditioning event is essential to understanding:

$$\Pr(A|B) = \frac{|A \cap B|}{|B|}$$

  - We can define conditional probabilities in terms of other (unconditional) probabilities:

$$\Pr(X|Y) = \frac{\Pr(X, Y)}{\Pr(Y)}$$

- Conditional probabilities are very useful!
  - What's the probability that she will get the flu given that she get's vacinated in the fall? (hypothetical)
  - What's the probability that he has jaundice given that he has yellow skin and he just ate four pounds of carrots? (evidence and observation)

Conditional Probabilities (continued)

- Given the equalities

$$\Pr(X|Y) = \frac{\Pr(X, Y)}{\Pr(Y)}$$

$$\Pr(Y|X) = \frac{\Pr(Y, X)}{\Pr(X)}$$

$$\Pr(X, Y) = \Pr(Y, X)$$

- we can easily derive *Bayes rule*

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

- which turns out to be enormously useful for a wide range of problems, including medical diagnosis.

$$\Pr(\text{disease}|\text{symptoms})$$

$$= \frac{\Pr(\text{symptoms}|\text{disease}) \times \Pr(\text{disease})}{\Pr(\text{symptoms})}$$

$$\propto \Pr(\text{symptoms}|\text{disease}) \times \Pr(\text{disease})$$

Manipulating Probability Formulas

- Given the joint distribution $\Pr(X, Y, Z)$, we can obtain $\Pr(X, Y)$ by marginalizing the joint distribution to $\{X, Y\}$ as follows

$$\Pr(X, Y) = \sum_{z \in \Omega_Z} \Pr(X, Y, Z = z)$$

  - We could continue to marginalize out variables as in

$$\Pr(X) = \sum_{y \in \Omega_Y} \Pr(X, Y = y)$$

- The addition rule

$$\Pr(X) = \sum_{y \in \Omega_Y} \Pr(X | Y = y) \times \Pr(Y = y)$$

  - follows from marginalization and the definition of conditional probability.

- The chain rule

$$\Pr(X, Y, Z) = \Pr(X | Y, Z)\Pr(Y | Z)\Pr(Z)$$

  - shown for the case of three variables enables us to factor a joint distribution into a product of conditional probabilities in much the same way that you can factor an integer into a product of smaller integers.

Conditional Independence

- The variable A is conditionally independent of B given C if

$$\Pr(A|B, C) = \Pr(A|C)$$

- Note that if A is conditionally independent of B given C then

$$\Pr(A, B|C) = \Pr(A|C) \times \Pr(B|C)$$

  - You should be able to prove this from conditional independence and the definition of conditional probability.

- Note that if A is conditionally independent of B given C then B is conditionally independent of A given C.

- Finally, if A is not conditionally independent of B given C then we say that A is conditionally dependent on B given C.
  - Conditional independence assumptions are crucial in enabling practical calculations involving probabilities.

Bayesian Model Selection

- Find a model (hypothesis) that best explains the data by finding the hypothesis that maximizes:

$$\Pr(\text{hypothesis} \mid \text{data})$$

- Again, Bayes rule allows us to make use of distributions that we are likely to have:

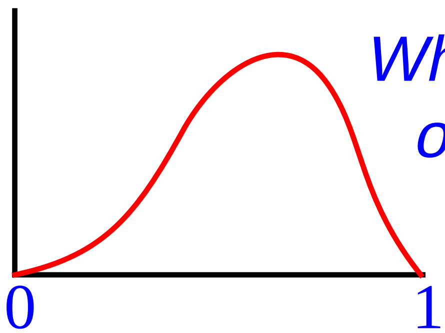
$$\Pr(\text{hypothesis} \mid \text{data})$$

$$\propto \Pr(\text{data} \mid \text{hypothesis}) \times \Pr(\text{hypothesis})$$

- As simple as this formula may seem, it is at the heart of a great many techniques for drawing appropriate conclusions from data.

Bayesian Model Selection (continued)

- What about $\Pr(\text{data} \mid \text{hypothesis})$?
  - Consider the problem of guessing the distribution governing a biased coin?



What is the probability of observing h heads out of m tosses of the coin given that the hypothesis is correct?

  - In general, it seems that the probability of the data given the hypothesis should be related to the error that we might incur if we were to use the hypothesis.

$$\Pr(\text{data} \mid \text{hypothesis}) \propto -\sum_{(x,\, y) \in \text{ data}} (\text{hypothesis}(x) - y)^2$$

- What about $\Pr(\text{hypothesis})$?
  - What if the distribution is uniform over all hypotheses?
  - Why might it be something other than uniform?

Conclusions

- Probability theory allows us to build on the foundations of propositional logic to reason about uncertainty.
- We draw conclusions in much the same way as we made inferences in propositional logic, by manipulating formulas.
- In the lectures on concept learning, we'll see that learning and statistical inference are closely related.
- In the lectures on reinforcement learning, we'll use probability theory to generate optimal plans in uncertain environments.
- In the lectures on Bayesian networks, we'll introduce a general method for drawing conclusions from uncertain knowledge.