CS786: Lecture 1

- May 1st
- Basics: review of probability theory

Theories to deal with uncertainty

- Dempster-Shafer theory
- Fuzzy set theory
- Possibility theory

- **Probability theory**
  - Well established
    - Axioms of probability theory rediscovered by many scientists over time
  - Theory used by most scientists today

Probabilities

- Objectivist/Frequentist viewpoint:
  - *Pr(q)* denotes the relative frequency that q was observed to be true
- Subjectivist/Bayesian viewpoint:
  - We'll *quantify* our beliefs using *probabilities*
  - *Pr(q)* denotes probability that you believe *q* is true
  - Note: statistics/data *influence* degrees of belief
- Let's formalize things…

Random Variables

- Assume set ***V*** of *random variables*: *X, Y*, etc.
  - Each RV *X* has a *domain* of values *Dom(X)*
  - *X* can take on any value from *Dom(X)*
  - Assume **V** and *Dom(X)* finite
- Examples
  - *Dom(X)* = $\{x_1, x_2, x_3\}$
  - *Dom(Weather) = {sunny, cloudy, rainy}*
  - *Dom(StudentInPascalsOffice) =*
    {bob, georgios, veronica, tianhan…}
  - *Dom(CraigHasCoffee) = {T,F}* (boolean var)

Random Variables/Possible Worlds

- A *formula* is a logical combination of variable assignments:
  - $X = x_1$; $(X = x_2 \vee X = x_3) \wedge Y = y_2$; $(x_2 \vee x_3) \wedge y_2$
  - chc ∧ ~cm, etc…
  - let $\mathcal{L}$ denote the set of formulae (our language)
- A *possible world* is an assignment of values to each RV
  - these are analogous to truth assts (interpretations)
  - Let W be the set of worlds

Probability Distributions

- A probability distribution Pr: $\mathcal{L} \rightarrow [0,1]$ s.t.
  - $0 \leq Pr(\alpha) \leq 1$
  - $Pr(\alpha) = Pr(\beta)$ if α is logically equivalent to β
  - $Pr(\alpha) = 1$ if α is a tautology (always true)
  - $Pr(\alpha) = 0$ if α is impossible (always false)
  - $Pr(\alpha \vee \beta) = Pr(\alpha) + Pr(\beta) - Pr(\alpha \wedge \beta)$
- For continuous random variables, we use probability densities.

Example Distribution

T – mail truck outside
M – mail waiting
C – craig wants coffee
A – craig is angry

$t\ c\ m\ \underline{a}$	0.162	$\overline{\underline{t}}\ c\ m\ \underline{a}$	0.0
$t\ c\ \underline{m}\ a$	0.018	$\underline{t}\ c\ \underline{m}\ a$	0.0
$t\ c\ \underline{m}\ \underline{a}$	0.016	$\underline{t}\ c\ \underline{m}\ \underline{a}$	0.0
$t\ \underline{c}\ m\ a$	0.004	$\underline{t}\ \underline{c}\ m\ a$	0.0
$t\ \underline{c}\ m\ \underline{a}$	0.432	$\underline{t}\ \underline{c}\ m\ \underline{a}$	0.0
$t\ \underline{c}\ \underline{m}\ a$	0.288	$\underline{t}\ \underline{c}\ \underline{m}\ a$	0.0
$t\ \underline{c}\ \underline{m}\ \underline{a}$	0.008	$\underline{t}\ \underline{c}\ \underline{m}\ \underline{a}$	0.0
$t\ c\ m\ a$	0.072	$t\ c\ m\ a$	0.0

Pr(t) =1
Pr(-t) = 0
Pr(c) = .2
Pr(-c) = .8
Pr(m) = .9
Pr(a) = .618
Pr(c & m) = .18
Pr(c v m) = .92
Pr(a -> m)
= Pr(-a v m)
= 1 – Pr(a & -m)
= .976

Conditional Probability

- Conditional probability critical in inference

$$\Pr(b \mid a) = \frac{\Pr(b \wedge a)}{\Pr(a)}$$

- if Pr(a) = 0, we often treat Pr(b|a)=1 by convention

Intuitive Meaning of Cond. Prob.

- Intuitively, if you learned a, you would change your degree of belief in b from Pr(b) to Pr(b|a)
- In our example:
  - Pr(m|c) = 0.9
  - Pr(m| ~c) = 0.9
  - Pr(a) = 0.618
  - Pr(a|~m) = 0.27
  - Pr(a|~m & c) = 0.8
- Notice the *nonmonotonicity* in the last three cases when additional evidence is added
  - contrast this with logical inference

Some Important Properties

- **Product Rule:** Pr(ab) = Pr(a|b)Pr(b)
- **Summing Out Rule:**

$$\Pr(a) = \sum_{b \in Dom(B)} \Pr(a \mid b) \Pr(b)$$

- **Chain Rule:**

  Pr(abcd) = Pr(a|bcd)Pr(b|cd)Pr(c|d)Pr(d)

  - holds for any number of variables

Bayes Rule

- **Bayes Rule:**

$$\Pr(a \mid b) = \frac{\Pr(b \mid a)\Pr(a)}{\Pr(b)}$$

- Bayes rule follows by simple algebraic manipulation of the defn of condition probability
  - why is it so important? why significant?
  - usually, one "direction" easier to assess than other

Example of Use of Bayes Rule

- Disease ∊ {malaria, cold, flu}; Symptom = fever
  - Must compute Pr(D | fever) to prescribe treatment
- Why not assess this quantity directly?
  - Pr(mal | fever) is not natural to assess; Pr(fever | mal) reflects the underlying "causal" mechanism
  - Pr(mal | fever) is not "stable": a malaria epidemy changes this quantity (for example)
- So we use Bayes rule:
  - Pr(mal | fever) = Pr(fever | mal) Pr(mal) / Pr(fever)
  - note that Pr(fev) = Pr(m&fev) + Pr(c&fev) + Pr(fl&fev)
  - so if we compute Pr of each disease given fever using Bayes rule, normalizing constant is "free"

Probabilistic Inference

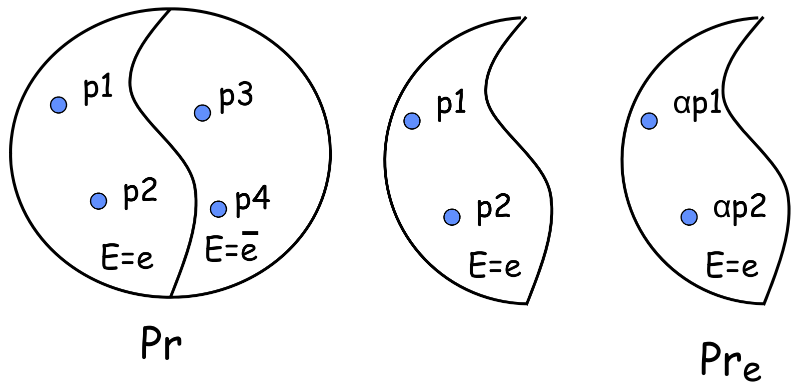
- By probabilistic inference, we mean
  - given a *prior* distribution *Pr* over variables of interest, representing degrees of belief
  - and given new evidence $E=e$ for some var E
  - Revise your degrees of belief: *posterior* Pr_e

- How do your degrees of belief change as a result of learning $E=e$ (or more generally $\mathbf{E}=\mathbf{e}$, for set $\mathbf{E}$)

Conditioning

- We define $Pr_e(\alpha) = Pr(\alpha \mid e)$
- That is, we produce Pr_e by *conditioning* the prior distribution on the observed evidence e
- Intuitively,
  - we set Pr(w) = 0 for any world falsifying e
  - we set Pr(w) = Pr(w) / Pr(e) for any world consistent with e
  - last step known as normalization (ensures that the new measure sums to 1)

Semantics of Conditioning



$\alpha = 1/(p1+p2)$
normalizing constant

Inference: Computational Bottleneck

- Semantically/conceptually, picture is clear; but several issues must be addressed
- Issue 1: How do we specify the full joint distribution over $X_1, X_2, \ldots, X_n$?
  - exponential number of possible worlds
  - e.g., if the X_i are boolean, then 2^n numbers (or 2^n -1 parameters/degrees of freedom, since they sum to 1)
  - these numbers are not robust/stable
  - these numbers are not natural to assess (what is probability that “Pascal wants coffee; it’s raining in Toronto; robot charge level is low; …”?)

Inference: Computational Bottleneck

- Issue 2: Inference in this rep'n frightfully slow
  - Must sum over exponential number of worlds to answer query *Pr(α)* or to condition on evidence e to determine $Pr_e(\alpha)$
- How do we avoid these two problems?
  - no solution in general
  - but in practice there is structure we can exploit
- We'll use conditional independence

Independence

- Recall that x and y are *independent* iff:
  - Pr(x) = Pr(x|y) iff Pr(y) = Pr(y|x) iff Pr(xy) = Pr(x)Pr(y)
  - intuitively, learning y doesn't influence beliefs about x
- x and y are *conditionally independent given* z iff:
  - Pr(x|z) = Pr(x|yz) iff Pr(y|z) = Pr(y|xz) iff
    Pr(xy|z) = Pr(x|z)Pr(y|z) iff …
  - intuitively, learning y doesn't influence your beliefs about x *if you already know z*
  - e.g., learning someone's mark on 886 project can influence the probability you assign to a specific GPA; but if you already knew 886 **final grade**, learning the project mark would *not* influence GPA assessment

What does independence buy us?

- Suppose (say, boolean) variables $X_1, X_2, \ldots, X_n$ are mutually independent
  - we can specify full joint distribution using only n parameters (linear) instead of 2^n -1 (exponential)
- How?
  - Simply specify $Pr(x_1), \ldots Pr(x_n)$
  - from this I can recover probability of any world or any (conjunctive) query easily
  - e.g. $Pr(x_1 \sim x_2 x_3 x_4) = Pr(x_1)\ (1-Pr(x_2))\ Pr(x_3)\ Pr(x_4)$
  - we can condition on observed value $X_k = x_k$ trivially by changing Pr(x_k) to 1, leaving Pr(x_i) untouched for $i \neq k$

The Value of Independence

- Complete independence reduces both *representation of joint* and *inference* from O(2^n) to O(n): pretty significant!
- Unfortunately, such complete mutual independence is very rare. Most realistic domains do not exhibit this property.
- Fortunately, most domains do exhibit a fair amount of conditional independence. And we can exploit conditional independence for representation and inference as well.
- **Bayesian networks** do just this