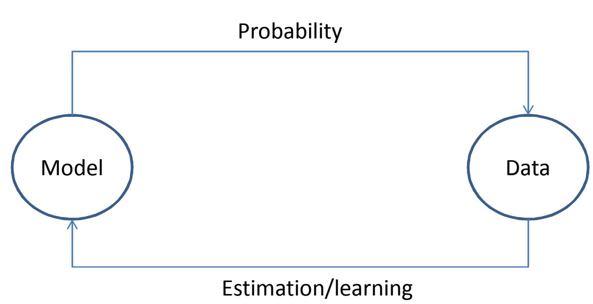
Probability and Statistics Review

Thursday Sep 11

The Big Picture



But how to specify a model?

Graphical Models

- How to specify the model?
  - What are the variables of interest?
  - What are their ranges?
  - How likely their combinations are?
- You need to specify a joint probability distribution
  - But in a compact way
    - Exploit local structure in the domain
- **Today: we will cover some concepts that formalize the above statements**

Probability Review

- Events and Event spaces
- Random variables
- Joint probability distributions
  - Marginalization, conditioning, chain rule, Bayes Rule, law of total probability, etc.
- Structural properties
  - Independence, conditional independence
- Examples
- Moments

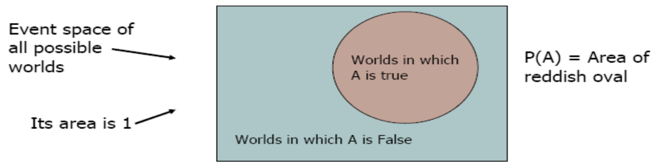
Sample space and Events

- Ω : Sample Space, result of an experiment
  - If you toss a coin twice $\Omega = \{HH, HT, TH, TT\}$
- Event: a subset of Ω
  - First toss is head = {HH,HT}
- S: event space, a set of events:
  - Closed under finite union and complements
    - Entails other binary operation: union, diff, etc.
  - Contains the empty event and Ω

Probability Measure

- Defined over (Ω, S) s.t.
  - $P(\alpha) >= 0$ for all α in S
  - $P(\Omega) = 1$
  - If α, β are disjoint, then
    - $P(\alpha \text{ U } \beta) = p(\alpha) + p(\beta)$
- We can deduce other axioms from the above ones
  - Ex: $P(\alpha \text{ U } \beta)$ for non-disjoint event

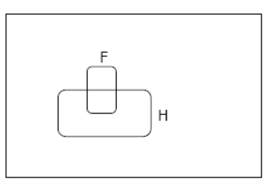
Visualization



- We can go on and define conditional probability, using the above visualization

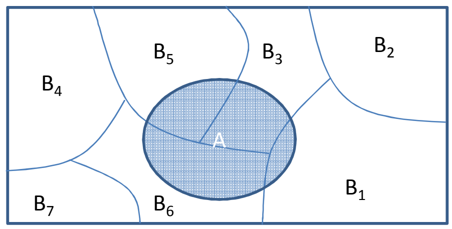
Conditional Probability

-P(F|H) = Fraction of worlds in which H is true that also have F true



$$p(f \mid h) = \frac{p(F \cap H)}{p(H)}$$

Rule of total probability

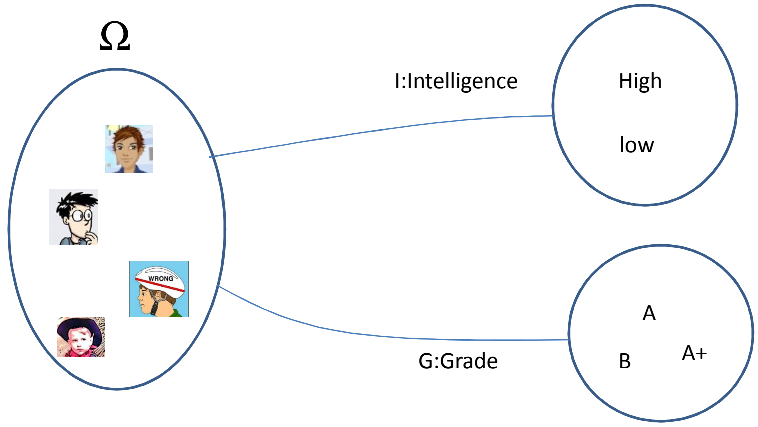


$$p(A) = \sum P(B_i)P(A \mid B_i)$$

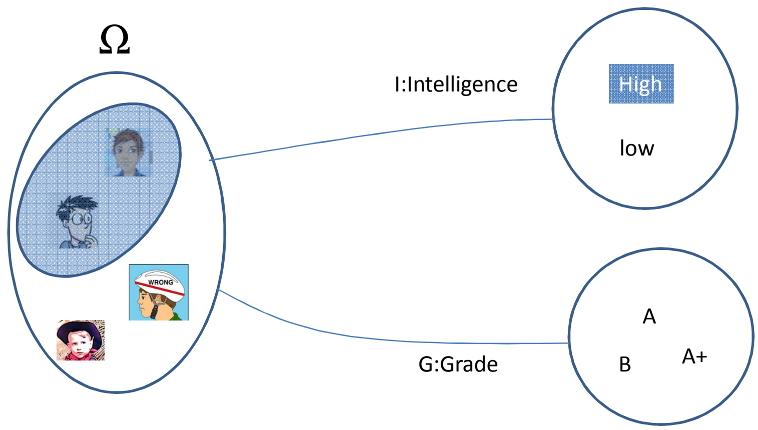
From Events to Random Variable

- Almost all the semester we will be dealing with RV
- Concise way of specifying attributes of outcomes
- Modeling students (Grade and Intelligence):
  - $\Omega =$ all possible students
  - What are events
    - Grade_A = all students with grade A
    - Grade_B = all students with grade A
    - Intelligence_High = … with high intelligence
  - Very cumbersome
  - We need "functions" that maps from Ω to an attribute space.

Random Variables



Random Variables



P(I = high) = P({all students whose intelligence is high})

Probability Review

- Events and Event spaces
- Random variables
- Joint probability distributions
  - Marginalization, conditioning, chain rule, Bayes Rule, law of total probability, etc.
- Structural properties
  - Independence, conditional independence
- Examples
- Moments

Joint Probability Distribution

- Random variables encodes attributes
- Not all possible combination of attributes are equally likely
  - Joint probability distributions quantify this
- $P(X= x, Y= y) = P(x, y)$
  - How probable is it to observe these two attributes together?
  - Generalizes to N-RVs
  - How can we manipulate Joint probability distributions?

Chain Rule

- Always true
  - P(x,y,z) = p(x) p(y|x) p(z|x, y)
    = p(z) p(y|z) p(x|y, z)
    =…

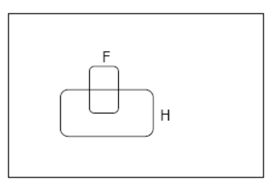
Conditional Probability

events

$$\mathrm{P}\left(\mathrm{X}=x \middle| \mathrm{Y}=y\right)=\frac{\mathrm{P}\left(\mathrm{X}=x \cap \mathrm{Y}=y\right)}{\mathrm{P}\left(\mathrm{Y}=y\right)}$$

But we will always write it this way:

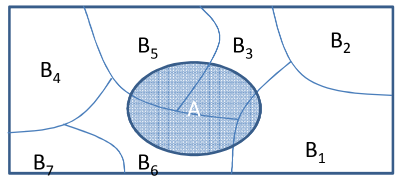
$$P(x \mid y)=\frac{p(x, y)}{p(y)}$$



Marginalization

- We know p(X,Y), what is P(X=x)?
- We can use the low of total probability, why?

$$p(x) = \sum_y P(x, y)$$
$$= \sum_y P(y) P(x \mid y)$$



Marginalization Cont.

- Another example

$$p(x) = \sum_{y,z} P(x, y, z)$$
$$= \sum_{z,y} P(y, z) P(x \mid y, z)$$

Bayes Rule

- We know that P(smart) = .7
  - If we also know that the students grade is A+, then how this affects our belief about his intelligence?

$$P(x \mid y) = \frac{P(x)P(y \mid x)}{P(y)}$$

  - Where this comes from?

Bayes Rule cont.

- You can condition on more variables

$$P(x \mid y, z) = \frac{P(x \mid z)P(y \mid x, z)}{P(y \mid z)}$$

Probability Review

- Events and Event spaces
- Random variables
- Joint probability distributions
  - Marginalization, conditioning, chain rule, Bayes Rule, law of total probability, etc.
- Structural properties
  - Independence, conditional independence
- Examples
- Moments

Independence

- X is independent of Y means that knowing Y does not change our belief about X.
  - $P(X|Y=y) = P(X)$
  - $P(X=x, Y=y) = P(X=x)\, P(Y=y)$
    - Why this is true?
  - The above should hold for all x, y
  - It is symmetric and written as $X \perp Y$

CI: Conditional Independence

- RV are rarely independent but we can still leverage local structural properties like CI.
- $X \perp Y \mid Z$ if once Z is observed, knowing the value of Y does not change our belief about X
  - The following should hold for all x,y,z
  - $P(X=x \mid Z=z, Y=y) = P(X=x \mid Z=z)$
  - $P(Y=y \mid Z=z, X=x) = P(Y=y \mid Z=z)$
  - $P(X=x, Y=y \mid Z=z) = P(X=x \mid Z=z)\, P(Y=y \mid Z=z)$

We call these factors : very useful concept !!

Properties of CI

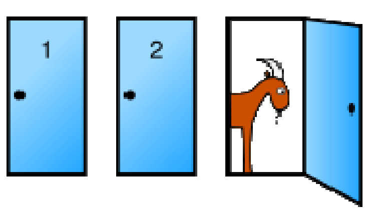
- **Symmetry:**
  - $(X \perp Y \mid Z) \Rightarrow (Y \perp X \mid Z)$
- **Decomposition:**
  - $(X \perp Y,W \mid Z) \Rightarrow (X \perp Y \mid Z)$
- **Weak union:**
  - $(X \perp Y,W \mid Z) \Rightarrow (X \perp Y \mid Z,W)$
- **Contraction:**
  - $(X \perp W \mid Y,Z)$ & $(X \perp Y \mid Z) \Rightarrow (X \perp Y,W \mid Z)$
- **Intersection:**
  - $(X \perp Y \mid W,Z)$ & $(X \perp W \mid Y,Z) \Rightarrow (X \perp Y,W \mid Z)$
  - Only for positive distributions!
  - $P(\alpha)>0, \forall\alpha, \alpha\neq\emptyset$
- **You will have more fun in your HW1 !!**

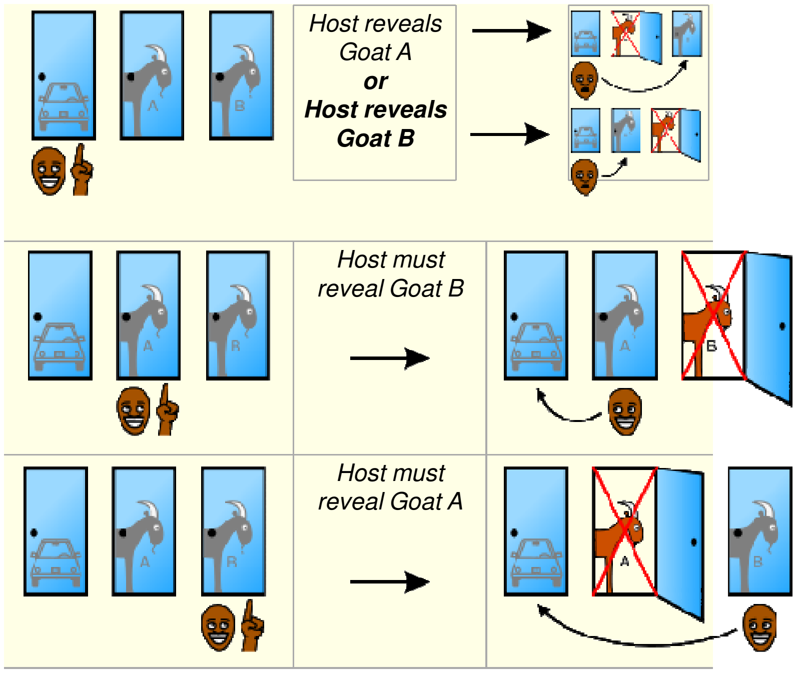
Probability Review

- Events and Event spaces
- Random variables
- Joint probability distributions
  - Marginalization, conditioning, chain rule, Bayes Rule, law of total probability, etc.
- Structural properties
  - Independence, conditional independence
- Examples
- Moments

Monty Hall Problem

- You're given the choice of three doors: Behind one door is a car; behind the others, goats.
- You pick a door, say No. 1
- The host, who knows what's behind the doors, opens another door, say No. 3, which has a goat.
- Do you want to pick door No. 2 instead?





Monty Hall Problem: Bayes Rule

- C_i : the car is behind door *i*, *i* = 1, 2, 3
- $P(C_i) = 1/3$
- H_{ij}: the host opens door *j* after you pick door *i*
- $$P(H_{ij} | C_k) = \begin{cases} 0 & i = j \\ 0 & j = k \\ 1/2 & i = k \\ 1 & i \neq k, j \neq k \end{cases}$$

Monty Hall Problem: Bayes Rule cont.

- WLOG, i=1, j=3
- $P(C_1|H_{13}) = \frac{P(H_{13}|C_1)P(C_1)}{P(H_{13})}$
- $P(H_{13}|C_1)P(C_1) = \frac{1}{2}\cdot\frac{1}{3} = \frac{1}{6}$

Monty Hall Problem: Bayes Rule cont.

- $$\begin{aligned} P(H_{13}) &= P(H_{13},C_1) + P(H_{13},C_2) + P(H_{13},C_3) \\ &= P(H_{13}|C_1)P(C_1) + P(H_{13}|C_2)P(C_2) \\ &= \frac{1}{6} + 1\cdot\frac{1}{3} \\ &= \frac{1}{2} \end{aligned}$$
- $P(C_1|H_{13}) = \frac{1/6}{1/2} = \frac{1}{3}$

Monty Hall Problem: Bayes Rule cont.

- $P(C_1|H_{13}) = \frac{1/6}{1/2} = \frac{1}{3}$
- $P(C_2|H_{13}) = 1 - \frac{1}{3} = \frac{2}{3} > P(C_1|H_{13})$
- *You should switch!*

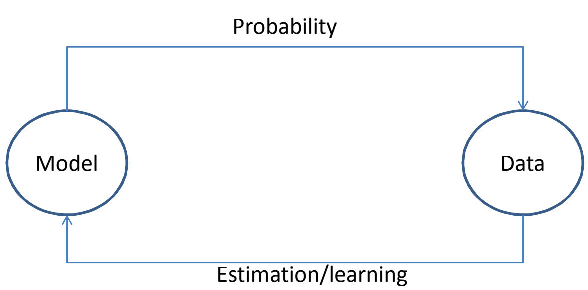
Moments

- Mean (Expectation): $\mu = E(\mathrm{X})$
  - Discrete RVs: $E(\mathrm{X}) = \sum_{v_i} v_i \mathrm{P}(\mathrm{X} = v_i)$
  - Continuous RVs: $E(\mathrm{X}) = \int_{-\infty}^{+\infty} xf(x)\,dx$
- Variance: $V(\mathrm{X}) = E(\mathrm{X} - \mu)^2$
  - Discrete RVs: $V(\mathrm{X}) = \sum_{v_i} (v_i - \mu)^2 \mathrm{P}(\mathrm{X} = v_i)$
  - Continuous RVs: $V(\mathrm{X}) = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x)dx$

Properties of Moments

- Mean
  - $E(\mathrm{X}+\mathrm{Y})=E(\mathrm{X})+E(\mathrm{Y})$
  - $E(a\mathrm{X})=aE(\mathrm{X})$
  - If X and Y are independent, $E(\mathrm{XY})=E(\mathrm{X})\cdot E(\mathrm{Y})$
- Variance
  - $V(a\mathrm{X}+b)=a^2V(\mathrm{X})$
  - If X and Y are independent, $V(\mathrm{X}+\mathrm{Y})=V(\mathrm{X})+V(\mathrm{Y})$

The Big Picture



Statistical Inference

- Given observations from a model
  - What (conditional) independence assumptions hold?
    - Structure learning
  - If you know the family of the model (ex, multinomial), What are the value of the parameters: MLE, Bayesian estimation.
    - Parameter learning

MLE

- Maximum Likelihood estimation
  - Example on board
    - Given N coin tosses, what is the coin bias (θ)?
- Sufficient Statistics: SS
  - Useful concept that we will make use later
  - In solving the above estimation problem, we only cared about N_h, N_t , these are called the SS of this model.
    - All coin tosses that have the same SS will result in the same value of θ
    - Why this is useful?

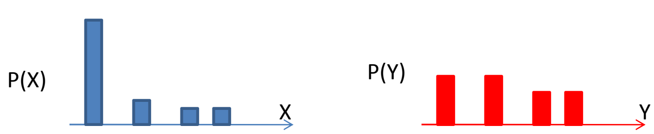
Statistical Inference

- Given observation from a model
  - What (conditional) independence assumptions holds?
    - Structure learning
  - If you know the family of the model (ex, multinomial), What are the value of the parameters: MLE, Bayesian estimation.
    - Parameter learning

We need some concepts from information theory

Information Theory

- P(X) encodes our uncertainty about X
  - Some variables are more uncertain that others



  - How can we quantify this intuition?
    - Entropy: average number of bits required to encode X

$$H_p(X) = E\left[\log \frac{1}{p(x)}\right] = \sum_x P(x) \log \frac{1}{P(x)}$$

Information Theory cont.

- Entropy: average number of bits required to encode X

$$H_P(X) = E\left[\log \frac{1}{p(x)}\right] = \sum_x P(x) \log \frac{1}{P(x)}$$

- We can define conditional entropy similarly

$$H_P(X \mid Y) = E\left[\log \frac{1}{p(x \mid y)}\right] = H_P(X, Y) - H_P(Y)$$

- We can also define chain rule for entropies (not surprising)

$$H_P(X, Y, Z) = H_P(X) + H_P(Y \mid X) + H_P(Z \mid X, Y)$$

Mutual Information: MI

- Remember independence?
  - If X⊥Y then knowing Y won't change our belief about X
  - Mutual information can help quantify this! (not the only way though)
- MI: $I_P(X;Y) = H_P(X) - H_P(X \mid Y)$
  - Symmetric
  - I(X;Y) = 0 iff, X and Y are independent!

Continuous Random Variables

- What if X is continuous?
- Probability density function (pdf) instead of probability mass function (pmf)
- A pdf is any function $f(x)$ that describes the probability density in terms of the input variable *x*.

PDF

- Properties of pdf
  - $f(x) \geq 0, \forall x$
  - $\int_{-\infty}^{+\infty} f(x) = 1$
  - $f(x) \leq 1$???
- Actual probability can be obtained by taking the integral of pdf
  - E.g. the probability of X being between 0 and 1 is

$$\mathrm{P}(0 \leq X \leq 1) = \int_0^1 f(x)dx$$

Cumulative Distribution Function

- $F_X(v) = P(X \le v)$
- Discrete RVs
  - $F_X(v) = \sum_{v_i} P(X = v_i)$
- Continuous RVs
  - $F_X(v) = \int_{-\infty}^{v} f(x)\,dx$
  - $\frac{d}{dx} F_X(x) = f(x)$

Acknowledgment

- Andrew Moore Tutorial: http://www.autonlab.org/tutorials/prob.html
- Monty hall problem: http://en.wikipedia.org/wiki/Monty_Hall_problem
- http://www.cs.cmu.edu/~guestrin/Class/10701-F07/recitation_schedule.html