UVA CS 6316/4501 – Fall 2016 Machine Learning

Lecture 11: Probability Review

Dr. Yanjun Qi

University of Virginia

Department of Computer Science

Announcements: Schedule

– Midterm – Nov. 26 Wed / 3:30pm – 4:45pm / open notes
– HW4 includes sample midterm questions

– Grading of HW1 is available on Collab
– Solution of HW1 is available on Collab
– Grading of HW2 will be available next week
– Solution of HW2 will be available next week

Where are we ? → Five major sections of this course

- ❑ Regression (supervised)
- ❑ Classification (supervised)
- ❑ Unsupervised models
- ❑ Learning theory
- ❑ Graphical models

Where are we ? → Three major sections for classification

- We can divide the large variety of classification approaches into roughly three major types

1. Discriminative
   - directly estimate a decision rule/boundary
   - e.g., support vector machine, decision tree

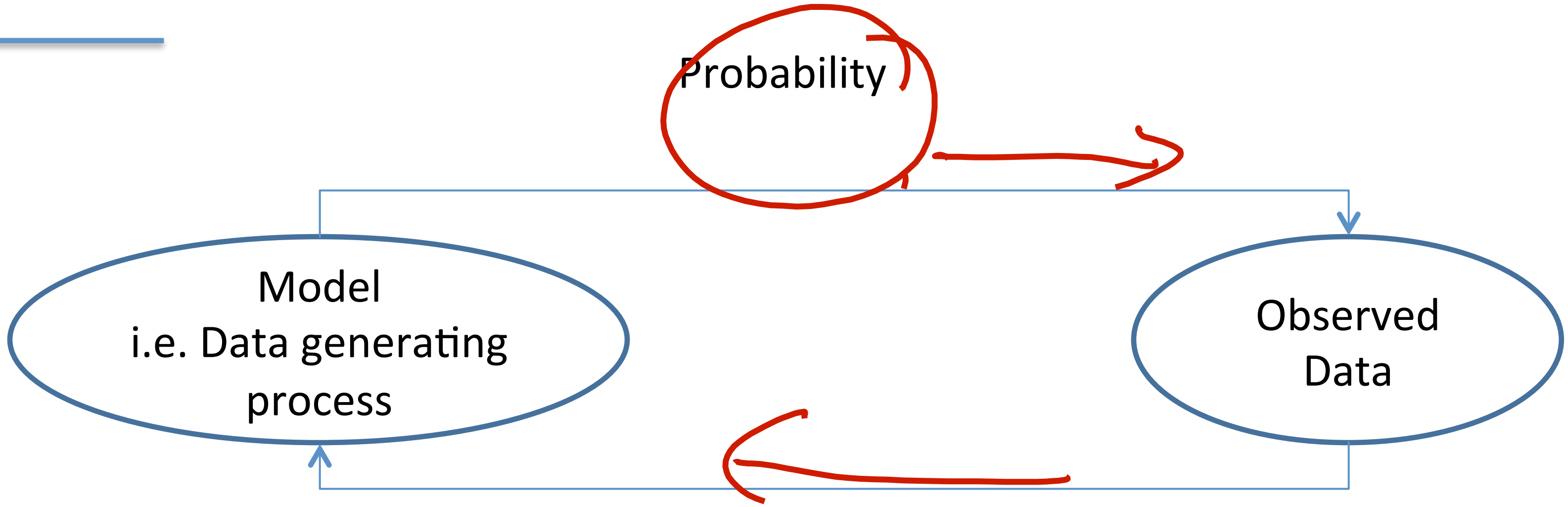
2. Generative:
   - build a generative statistical model
   - e.g., naïve bayes classifier, Bayesian networks

3. Instance based classifiers
   - Use observation directly (no models)
   - e.g. K nearest neighbors

Today : Probability Review

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

The Big Picture



Estimation / learning / Inference / Data mining

But how to specify a model?

Probability as frequency

- Consider the following questions:
  - 1. What is the probability that when I flip a coin it is "heads"?
  - 2. why ?

We can count → ~1/2

  - 3. What is the probability of Blue Ridge Mountains to have an erupting volcano in the near future ?

→ could not count

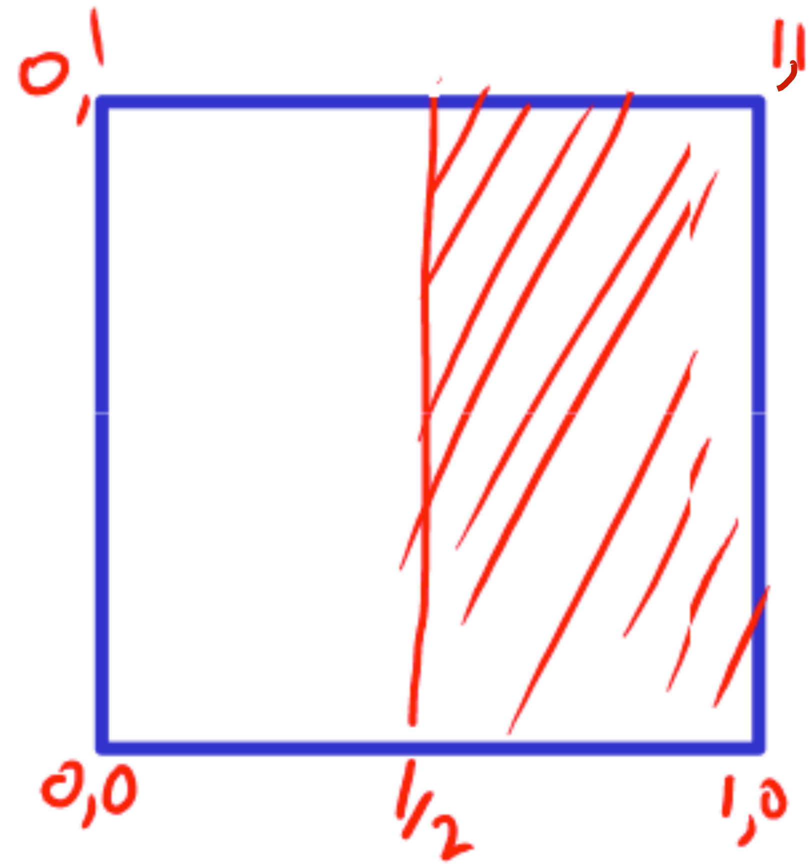
Message: *The frequentist view is very useful, but it seems that we also use domain knowledge to come up with probabilities.*

Adapt from Prof. Nando de Freitas's review slides

Probability as a measure of uncertainty

- Imagine we are throwing darts at a wall of size 1x1 and that all darts are guaranteed to fall within this 1x1 wall.

- What is the probability that a dart will hit the shaded area?

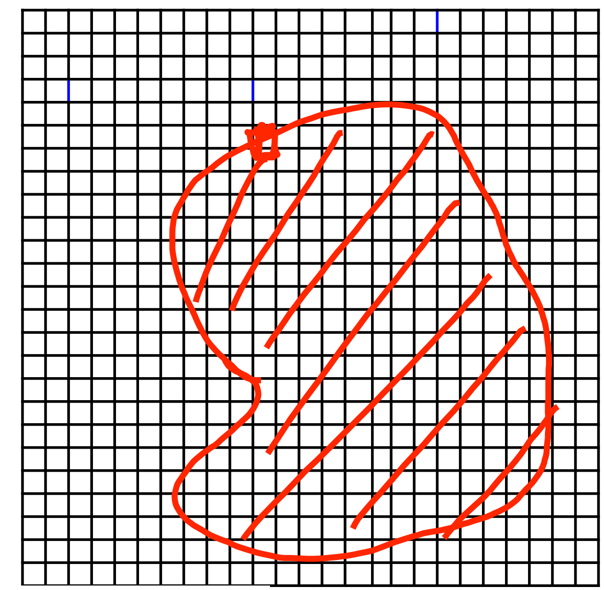


Adapt from Prof. Nando de Freitas's review slides

Probability as a measure of uncertainty

- *Probability is a measure of certainty of an event taking place.*

- *i.e. in the example, we were measuring the chances of hitting the shaded area.*



Its area is 1

$$prob = \frac{\#\,RedBoxes}{\#\,Boxes}$$

Adapt from Prof. Nando de Freitas's review slides

Today : Probability Review

- The big picture
- Sample space, Event and Event spaces
- Random variables
- Joint probability, Marginalization, conditioning, chain rule, Bayes Rule, law of total probability, etc.
- Structural properties
    - Independence, conditional independence

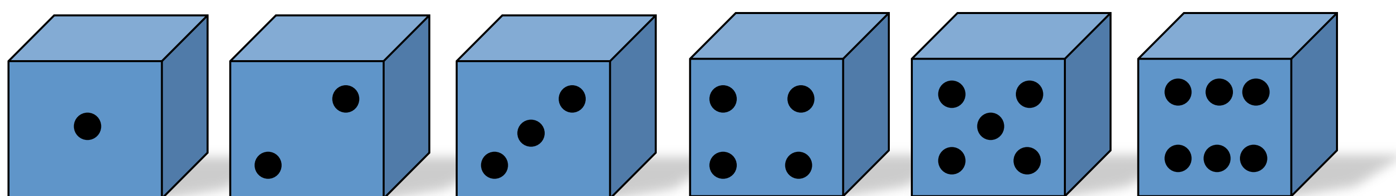
Probability

Probability *is the formal study of the laws of chance. Probability allows us to* ***manage uncertainty****.*

The ***sample space*** *is the set of all* ***outcomes****. For example, for a die we have 6 outcomes:*

$$O_{die} = \{1,2,3,4,5,6\}$$

O:



Elementary Event “Throw a 2”

The elements of W are called *elementary events*.

Probability

- *Probability allows us to measure many* ***events.***
- ***The events are subsets of the sample space O.*** *For example, for a die we may consider the following events: e.g.,*

GREATER = {5, 6}

EVEN = {2, 4, 6}

- ***Assign probabilities to these events: e.g.,***

P(EVEN) = 1/2

Adapt from Prof. Nando de Freitas's review slides

Sample space and Events

- O : Sample Space,
  - result of an experiment / set of all outcomes
  - If you toss a coin twice O = {HH,HT,TH,TT}
- Event: a subset of O
  - First toss is head = {HH,HT}
- S: event space, a set of events:
  - Contains the empty event and O

$P(\emptyset) = 0$
$P(O) = 1$

Axioms for Probability

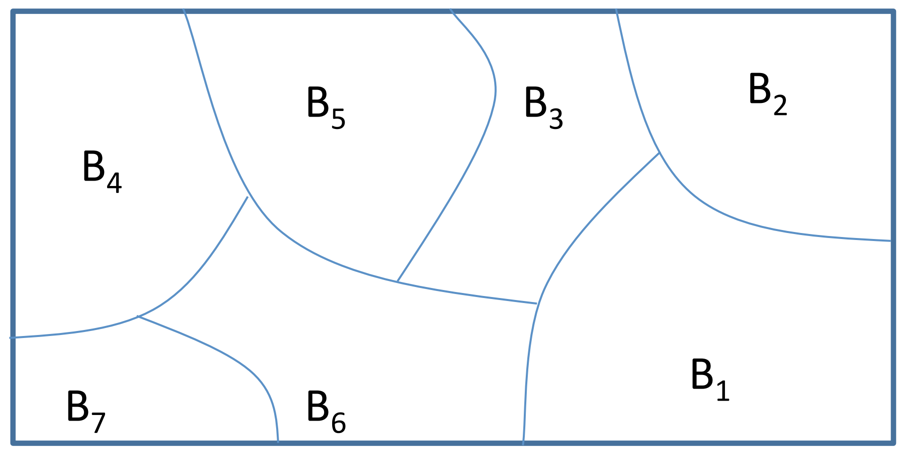
sample space

event space

- Defined over (O,S) s.t.
  - 1 >= P(a) >= 0 for all a in S
  - P(O) = 1
  - If A, B are disjoint, then
    - P(A U B) = p(A) + p(B)

Axioms for Probability

- $P(O) = \sum P(B_i)$

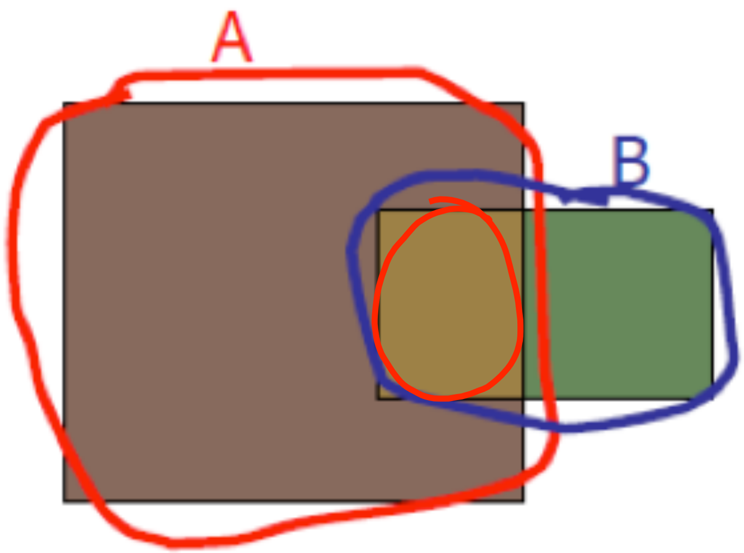


OR operation for Probability

- We can deduce other axioms from the above ones
  - Ex: P(A U B) for non-disjoint events

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

P(Union of A and B)



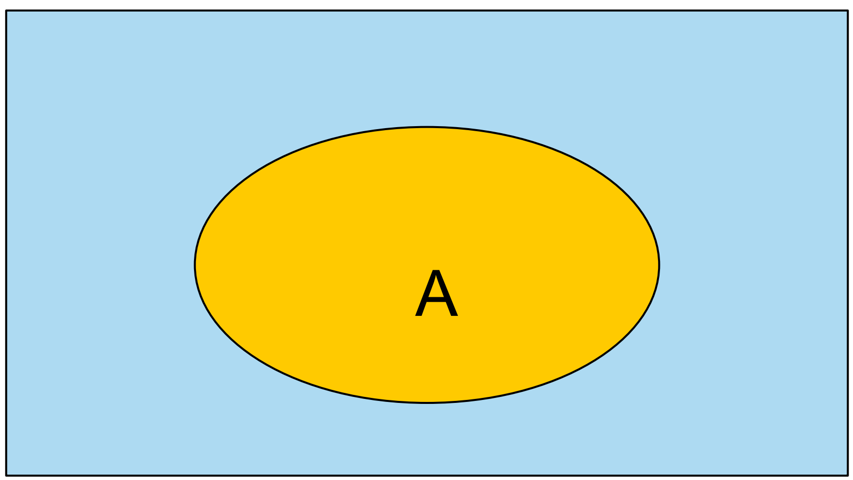
Theorems from the Axioms

- $0 <= P(A) <= 1$,
- P(A or B) = P(A) + P(B) - P(A and B)

From these we can prove:

$$P(\cancel{\text{not}}\ A) = P(\sim A) = 1\text{-}P(A)$$

Complement



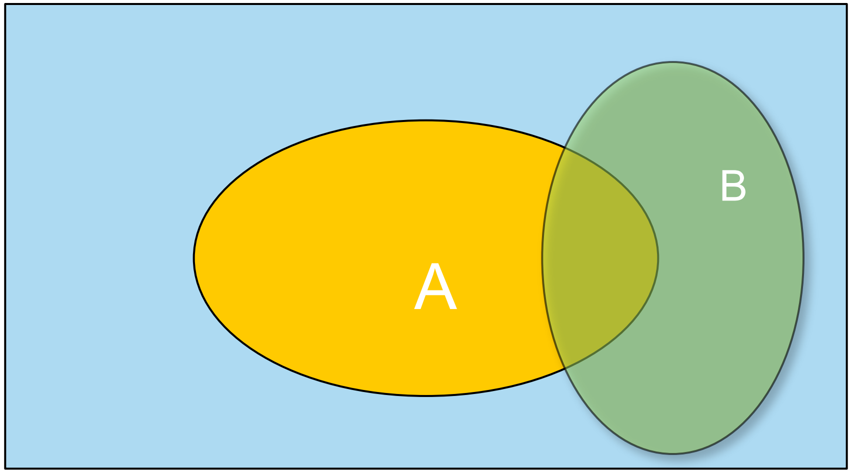
Another important theorem

- 0 <= P(A) <= 1,
- P(A or B) = P(A) + P(B) - P(A and B)

From these we can prove:

$$P(A) = P(A \wedge B) + P(A \wedge \sim B)$$

P(Intersection of A and B)



Another important theorem

- 0 <= P(A) <= 1,
- P(A or B) = P(A) + P(B) - P(A and B)

From these we can prove:

P(A) = P(A ^ B) + P(A ^ ~B)

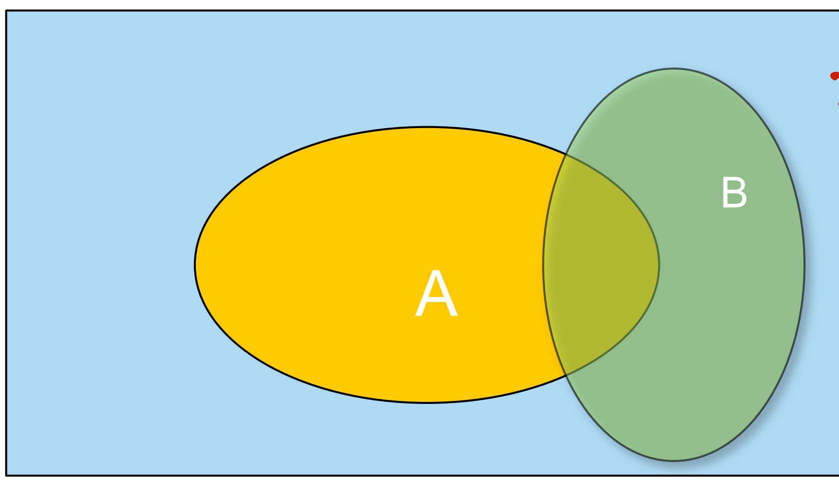
$P(A)$

$= P(A \cap \Omega)$

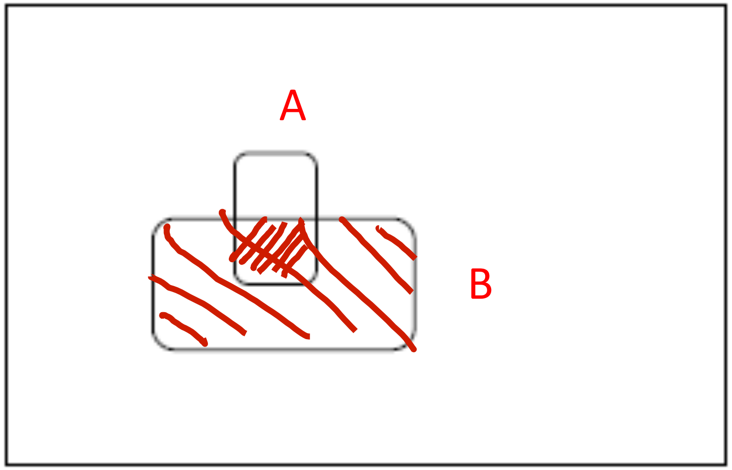
$= P(A \cap (B \cup \neg B))$

$= P((A \cap B) \cup (A \cap \neg B))$

$= P(A \cap B) + P(A \cap \neg B)$



Conditional Probability



$$P(A \text{ given } B) = P(A \text{ and } B) \,/\, P(B)$$

That is, in the frequentist interpretation, we calculate the ratio of the number of times both A and B occurred and divide it by the number of times B occurred.

Chain rule

For short we write: $P(A|B) = P(AB)/P(B)$; *or* $P(AB)=P(A|B)P(B)$, *where* $P(A|B)$ *is the* ***conditional*** *probability,* $P(AB)$ *is the* ***joint****, and* $P(B)$ *is the* ***marginal****.*

If we have more events, we use the chain rule:

$$P(ABC) = P(A|BC)\,P(B|C)\,P(C)$$

$P(ABC)$
$= P(A|BC)P(BC)$
$= P(A|BC)P(B|C)P(C)$

from Prof. Nando de Freitas's review

Conditional Probability / Chain Rule

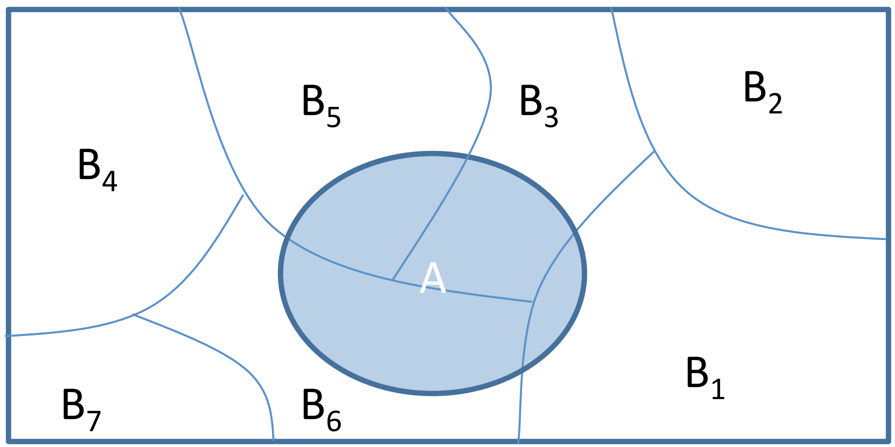
- More ways to write out chain rule …

joint conditional

$$P(A,B) = p(B|A)p(A)$$ → marginal

$$P(A,B) = p(A|B)p(B)$$

Rule of total probability => Marginalization



$P(A) =$

$P(A \wedge B) +$

$P(A \wedge \neg B)$

$$\Rightarrow p(A) = \sum \underbrace{P(B_i)P(A \mid B_i)}_{P(B_i \wedge A)}$$

WHY ???

$P(A) = P(A \cap \Omega)$

$= P(A \cap (B1 \cup B2 \cdots \cup Bk))$

$= \sum P(A \cap Bi)$

Today : Probability Review

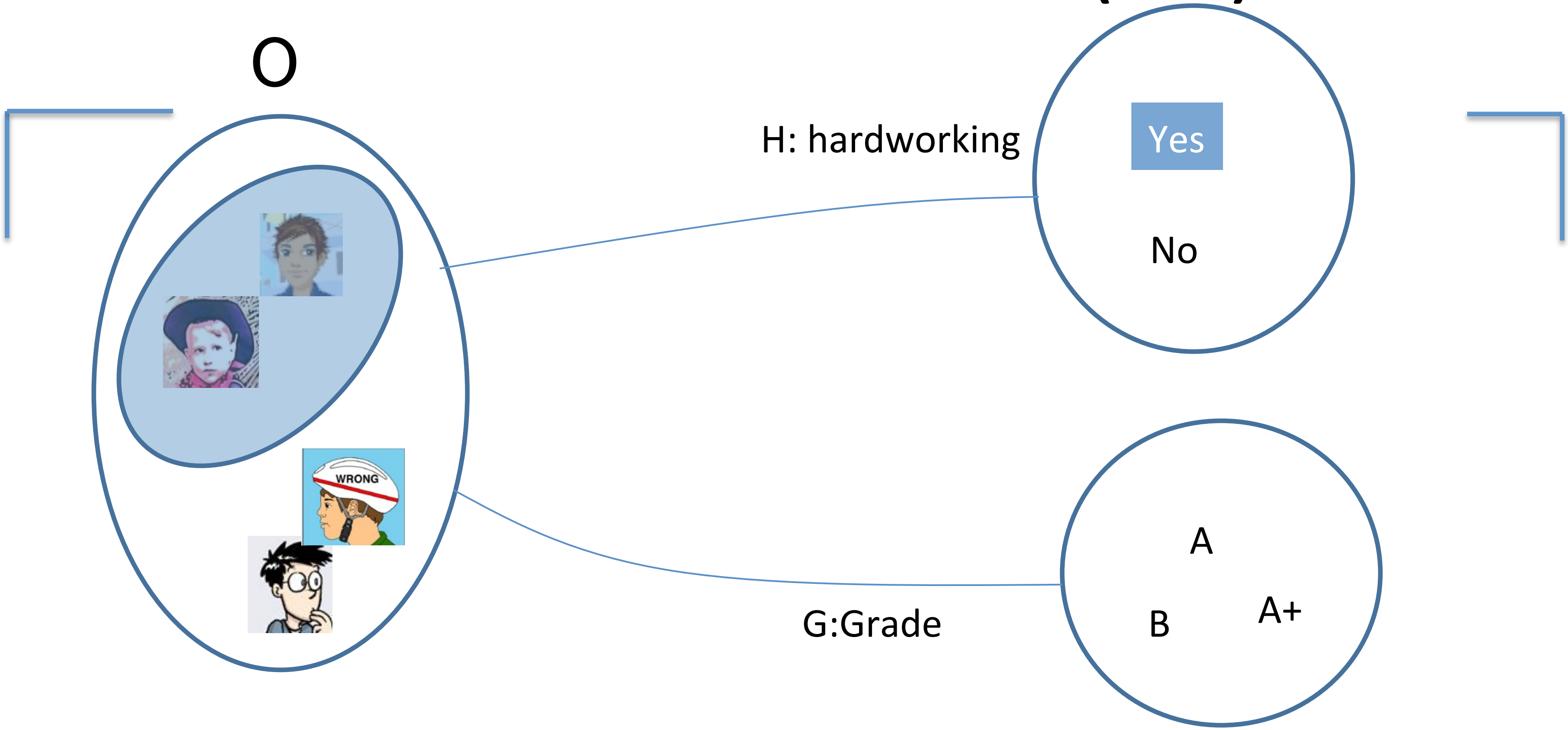
- The big picture
- Events and Event spaces
- Random variables
- Joint probability, Marginalization, conditioning, chain rule, Bayes Rule, law of total probability, etc.
- Structural properties
  - Independence, conditional independence

From Events to Random Variable

- Concise way of specifying attributes of outcomes
- Modeling students (Grade and Intelligence):
  - O = all possible students (sample space)
  - What are events (subset of sample space)
    - Grade_A = all students with grade A
    - Grade_B = all students with grade B
    - HardWorking_Yes = … who works hard
  - Very cumbersome

  - Need "functions" that maps from O to an attribute space T.
  - P(H = YES) = P({student ϵ O : H(student) = YES})

Random Variables (RV)



P(H = Yes) = P({all students who is working hard on the course})

- "functions" that maps from O to an attribute space T.

Notation Digression

- P(A) is shorthand for P(A=true)
- P(~A) is shorthand for P(A=false)
- Same notation applies to other binary RVs: P(Gender=M), P(Gender=F)
- Same notation applies to *multivalued* RVs: P(Major=history), P(Age=19), P(Q=c)
- Note: upper case letters/names for *variables*, lower case letters/names for *values*

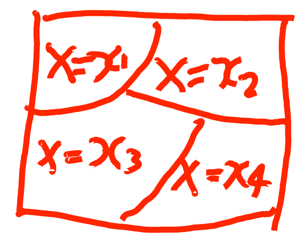
Discrete Random Variables

- Random variables (RVs) which may take on only a **countable** number of **distinct** values

- X is a RV with arity k if it can take on exactly one value out of $\{x_1, \ldots, x_k\}$

Probability of Discrete RV

- Probability mass function (pmf): $\boxed{P(X = x_i)}$
- Easy facts about pmf
  - $\Sigma_i P(X = x_i) = 1$
  - $P(X = x_i \cap X = x_j) = 0$ if $i \neq j$
  - $P(X = x_i \cup X = x_j) = P(X = x_i) + P(X = x_j)$ if $i \neq j$
  - $P(X = x_1 \cup X = x_2 \cup \ldots \cup X = x_k) = 1$



e.g. Coin Flips

- You flip a coin
  - Head with probability 0.5

- You flip 100 coins
  - How many heads would you expect

e.g. Coin Flips cont.

$\{H, T\}$

- You flip a coin
  - Head with probability *p*
  - Binary random variable
  - Bernoulli trial with success probability *p*
- You flip *k* coins
  - How many heads would you expect
  - Number of heads X: discrete random variable
  - Binomial distribution with parameters *k* and *p*

Discrete Random Variables

- Random variables (RVs) which may take on only a **countable** number of **distinct** values
  - E.g. the total number of heads X you get if you flip 100 coins
- X is a RV with arity *k* if it can take on exactly one value out of $\left\{x_1,\ldots,x_k\right\}$
  - E.g. the possible values that X can take on are 0, 1, 2,…, 100

e.g., two Common Distributions

- Uniform $X \sim U[1,\ldots,N]$
  - X takes values 1, 2, ..., *N*
  - $P(X=i)=1/N$
  - E.g. picking balls of different colors from a box

- Binomial $X \sim Bin(k,p)$
  - X takes values 0, 1, ..., *k*
  - $P(X=i)=\binom{k}{i} p^i (1-p)^{k-i}$
  - E.g. coin flips k times

Today : Probability Review

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

e.g., Coin Flips by Two Persons

- Your friend and you both flip coins
  - Head with probability 0.5
  - You flip 50 times; your friend flip 100 times
  - How many heads will both of you get

Joint Distribution

- Given two discrete RVs X and Y, their **joint distribution** is the distribution of X and Y together $P((X=21) \wedge (Y=70))$
  - E.g. P(You get 21 heads AND you friend get 70 heads)

Joint Distribution

- Given two discrete RVs X and Y, their **joint distribution** is the distribution of X and Y together $P((X=21) \wedge (Y=70))$
  - E.g. P(You get 21 heads AND you friend get 70 heads)
-
  - E.g. sum $$\sum_x \sum_y P(X = x \cap Y = y) = 1$$

$$\sum_{i=0}^{50} \sum_{j=0}^{100} P(\text{You get } i \text{ heads AND your friend get } j \text{ heads}) = 1$$

Conditional Probability

- $\mathrm{P}(\mathrm{X}=x \mid \mathrm{Y}=y)$ is the probability of $\mathrm{X}=x$, given the occurrence of $\mathrm{Y}=y$
  - E.g. you get 0 heads, given that your friend gets 61 heads

- $$\mathrm{P}(\underbrace{\mathrm{X}=}_{}x \mid \underbrace{\mathrm{Y}=y}_{}) = \frac{\mathrm{P}(\mathrm{X}=x \cap \mathrm{Y}=y)}{\mathrm{P}(\mathrm{Y}=y)}$$

Law of Total Probability

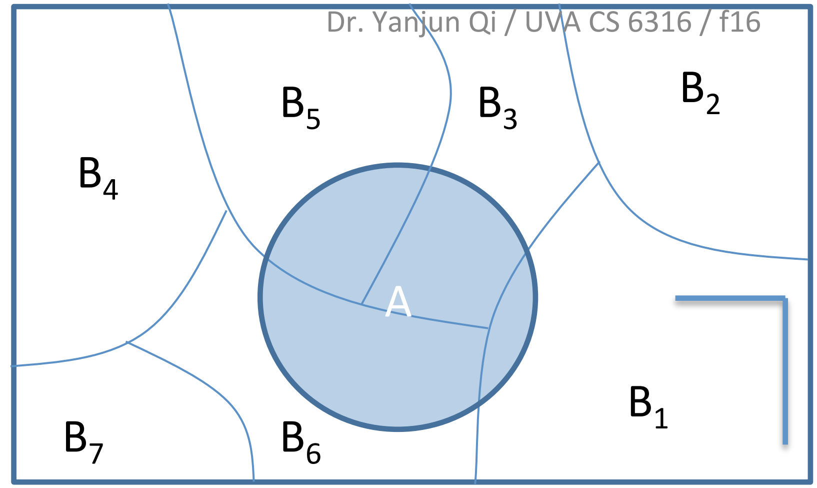
- Given two discrete RVs X and Y, which take values in $\{x_1, \ldots, x_m\}$ and $\{y_1, \ldots, y_n\}$, We have

$$= P(X = x_i \cap \Omega)$$

$$\mathrm{P}(\mathrm{X} = x_i) = \sum_j \mathrm{P}(\mathrm{X} = x_i \cap \mathrm{Y} = y_j)$$

$$= \sum_j \mathrm{P}(\mathrm{X} = x_i \mid \mathrm{Y} = y_j)\,\mathrm{P}(\mathrm{Y} = y_j)$$

Marginalization



Marginal Probability

Law of Total Probability

$$P(X = x_i) = \sum_j P(X = x_i \cap Y = y_j)$$

$\Downarrow$ chain rule

$$= \sum_j P(X = x_i \mid Y = y_j) P(Y = y_j)$$

Conditional Probability

Marginal Probability

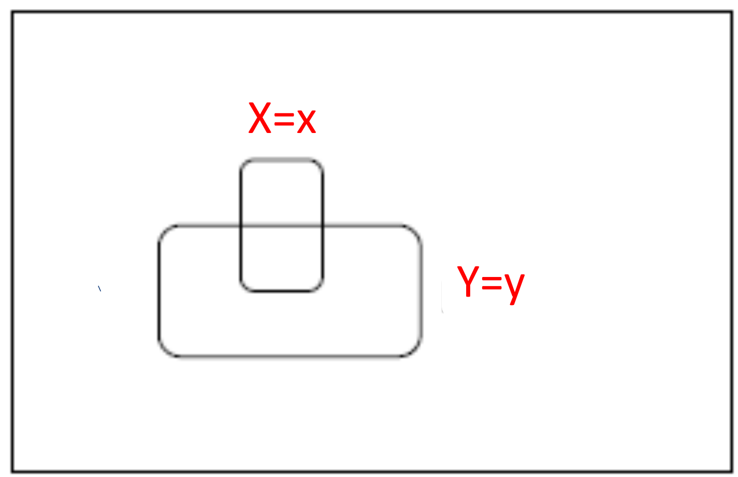
Simplify Notation: Conditional Probability

events

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

But we will always write it this way:

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



P(X=x true) -> P(X=x) -> P(x)

$\underset{\text{value}}{P(x)} \leftarrow \underset{\text{RV}}{P(X=x)} \leftarrow \underset{\text{event}}{P(X=x\ \text{true})}$

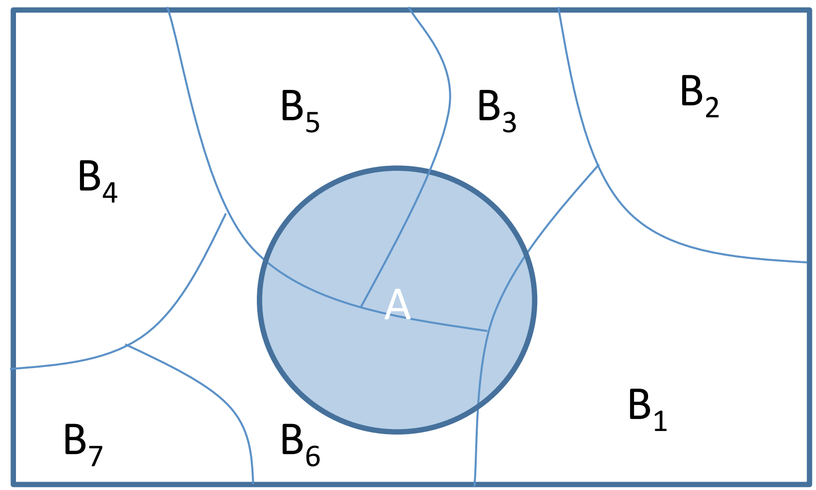
Simplify Notation: Marginalization

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

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

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

all possible Y values



Marginalization Cont.

- Another example

$$\sum_y \sum_z p(y,z) = 1$$

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

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

Bayes Rule

- We know that P(rain) = 0.5
  - If we also know that the grass is wet, then how this affects our belief about whether it rains or not?

0.5

1

$$P(rain \mid wet) = \frac{P(rain)P(wet \mid rain)}{P(wet)}$$

P(wet, rain)

P(wet, rain) + P(wet, sunny)

Weather

Grass

{rain, Sunny}

{wet, dry}

P(rain) P(Wet | rain) + P(Sunny) P(Wet | sunny)

Bayes Rule

- We know that P(rain) = 0.5
  - If we also know that the grass is wet, then how this affects our belief about whether it rains or not?

W= G=

$$P(rain \mid wet) = \frac{P(rain)P(wet \mid rain)}{P(wet)}$$

$P(W=S \mid wet)$

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

Bayes Rule

- X and Y are discrete RVs…

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

$$P\left(X = x_i \middle| Y = y_j\right) = \frac{P\left(Y = y_j \middle| X = x_i\right) P\left(X = x_i\right)}{\sum_k P\left(Y = y_j \middle| X = x_k\right) P\left(X = x_k\right)}$$

X_k

What we just did…

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

This is Bayes Rule

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London,* **53:370-418**

More General Forms of Bayes Rule

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

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

$$P(A = a_1 | B) = \frac{P(B | A = a_1)P(A = a_1)}{\sum_i P(B | A = a_i)P(A = a_i)}$$

Bayes Rule cont.

- You can condition on more variables

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

Conditional Probability Example

Assume we have a dark box with 3 red balls and 1 blue ball. That is, we have the ***set*** *{**r**,**r**,**r**,**b**}. What is the probability of drawing 2 red balls in the first 2 tries?*

$$P(B_1 = r, B_2 = r) = P(B_1 = r)\underbrace{P(B_2 = r \mid B_1 = r)}_{\frac{2}{3}}$$

$$P(B_1 = r) = \frac{3}{4}$$

$$P(B_1 = b) = \frac{1}{4}$$

Adapt from Prof. Nando de Freitas's review slides

Conditional Probability Example

Assume we have a dark box with 3 red balls and 1 blue ball. That is, we have the ***set*** *{r,r,r,b}. What is the probability of drawing 2 red balls in the first 2 tries?*

$$P(B_1 = r, B_2 = r) = P(B_1 = r)\, P(B_2 = r \mid B_1 = r)$$

$$= \frac{3}{4} \times \frac{2}{3} = \frac{1}{2}$$

Adapt from Prof. Nando de Freitas's review slides

Conditional Probability Example

What is the probability that the 2nd ball drawn from the **set** *{r,r,r,b} will be red?*

Using marginalization, $\boldsymbol{P(B_2 = r)}$ $=$ $P(B_2=r,\ B_1=r)$

$+\ P(B_2=r,\ B_1=b)$

Conditional Probability Example

What is the probability that the 2nd ball drawn from the set {r,r,r,b} will be red?

Joint

Using marginalization, $\boldsymbol{P(B_2 = r)} = P(B_2 = r \wedge B_1 = r) + P(B_2 = r \wedge B_1 = b)$

$$= \underbrace{P(B_1 = r)}_{\text{marginal}} \underbrace{P(B_2 = r \mid B_1 = r)}_{\text{Conditional}} + P(B_1 = b) P(B_2 = r \mid B_1 = b)$$

$$\begin{bmatrix} P(B_2=r) \\ P(B_2=b) \end{bmatrix} = \begin{bmatrix} P(B_2=r|B_1=r)P(B_1=r) + P(B_2=r|B_1=b)P(B_1=b) \\ P(B_2=b|B_1=r)P(B_1=r) + P(B_2=b|B_1=b)P(B_1=b) \end{bmatrix}$$

$$= \begin{bmatrix} P(B_2=r|B_1=r), & P(B_2=r|B_1=b) \\ P(B_2=b|B_1=r), & P(B_2=b|B_1=b) \end{bmatrix} \begin{bmatrix} P(B_1=r) \\ P(B_1=b) \end{bmatrix}$$

→ Matrix Notation

$P(B_2)$ $\quad P(B_2|B_1)$ $\quad P(B_1)$

$\Rightarrow$ matrix notation form $\Pi_2 = G^T \Pi_1$

For short, we write this using vectors and a stochastic matrix:

Today : Probability Review

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

Independent RVs

- Intuition: X and Y are independent means that $X = x$ **neither** makes it **more or less** probable that $Y = y$
- Definition: X and Y are independent ***iff***

$$P(X = x \cap Y = y) = P(X = x)P(Y = y)$$

$$P(X=x \mid Y=y)\,P(Y=y)$$

More on Independence

- $$P(X = x \cap Y = y) = P(X = x)P(Y = y)$$

$$P(X = x \mid Y = y) = P(X = x) \qquad P(Y = y \mid X = x) = P(Y = y)$$

- E.g. no matter how many heads you get, your friend will not be affected, and vice versa

More on 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)

- The above should hold for all x_i, y_j
- It is symmetric and written as $X \perp Y$

$X \perp Y$

Conditionally Independent RVs

- Intuition: X and Y are conditionally independent given Z means that once Z is **known**, the value of X does not add any **additional** information about Y
- Definition: X and Y are conditionally independent given Z ***iff***

X : sale ice-cream
Y : rate of Dro
Z : Weather

$$P\left(X = x \cap Y = y \middle| Z = z\right) = P\left(X = x \middle| Z = z\right) P\left(Y = y \middle| Z = z\right)$$

If holding for all x_i, y_j, z_k

$$X \perp Y \mid Z$$

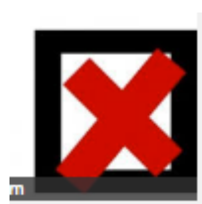
Conditionally Independent RVs

X: sale ice-cream (High vs. Low)

Y: rate of Drawning (High vs. Low)

Z: Weather is Hot (or not)

$X \perp Y \mid Z$ 

$X \perp Y$ 

More on Conditional Independence

$$P(X = x \cap Y = y \mid Z = z) = P(X = x \mid Z = z)\,P(Y = y \mid Z = z)$$

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

$$P(X = x \mid Y = y, Z = z) = P(X = x \mid Z = z)$$

$$P(Y = y \mid X = x, Z = z) = P(Y = y \mid Z = z)$$

Today Recap : Probability Review

- The big picture : data <-> probabilistic model
- Sample space, Events and Event spaces
- Random variables
- Joint probability, Marginal probability, conditional probability,
- Chain rule, Bayes Rule, Law of total probability, etc.
- Independence, conditional independence

References

- Prof. Andrew Moore's review tutorial
- Prof. Nando de Freitas's review slides
- Prof. Carlos Guestrin recitation slides