UVA CS 6316: Machine Learning

Lecture 12: Probability Review

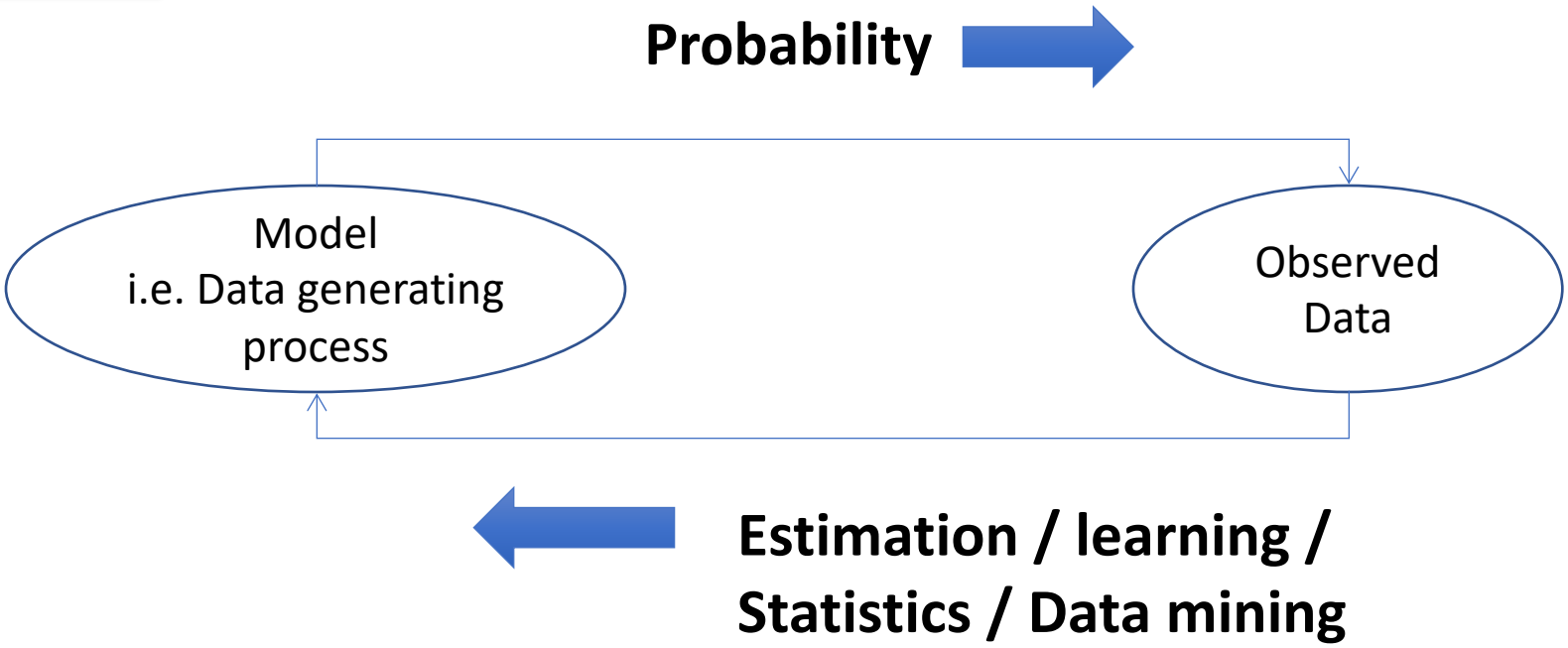
Dr. Yanjun Qi

University of Virginia
Department of Computer Science

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, e.g., Independence, conditional independence
- Maximum Likelihood Estimation

The Big Picture



Probability

- Counting
- Basics of probability
- Conditional probability
- Random variables
- Discrete and continuous distributions
- Expectation and variance
- Tail bounds and central limit theorem
- ……

Statistics

- Maximum likelihood estimation
- Bayesian estimation
- Hypothesis testing
- Linear regression
- [Machine learning]
- ……

Probability as frequency

- Consider the following questions:
  - 1. What is the probability that when I flip a coin it is "heads"?
  - 2. why ?
  - 3. What is the probability of Blue Ridge Mountains to have an erupting volcano in the near future ?

We can count ➔ ~1/2

➔ could not count

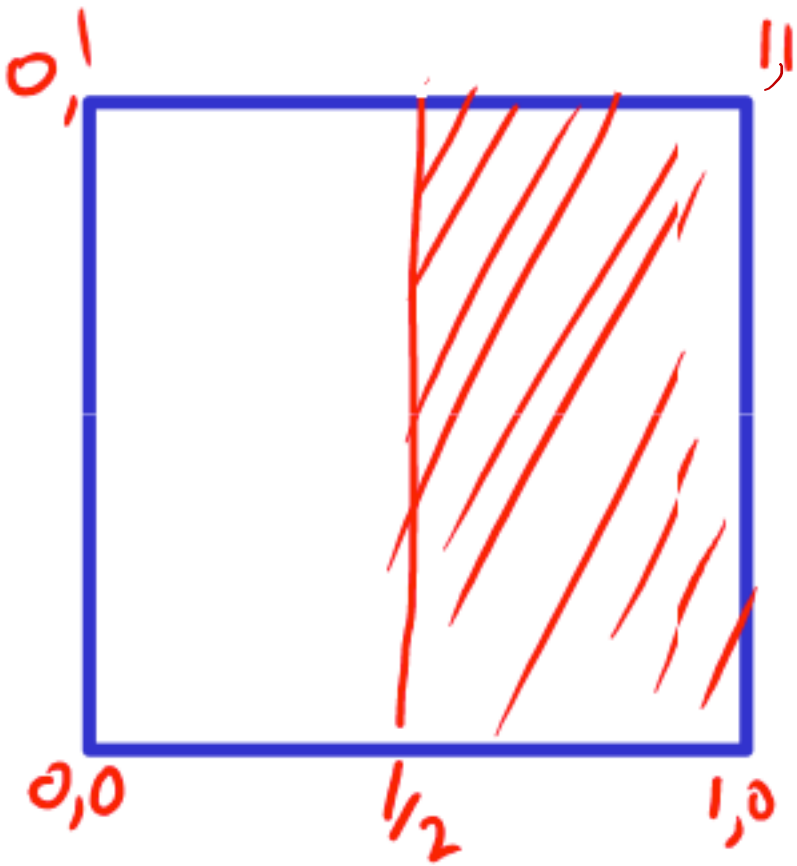
Message: *The frequentist view is very useful, but it seems that we can also use domain knowledge to come up with probabilities.*

Adapt from Prof. Nando de Freitas's review slides

Probability as a measure of uncertainty

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

- What is the probability that a dart wi
  the shaded area?

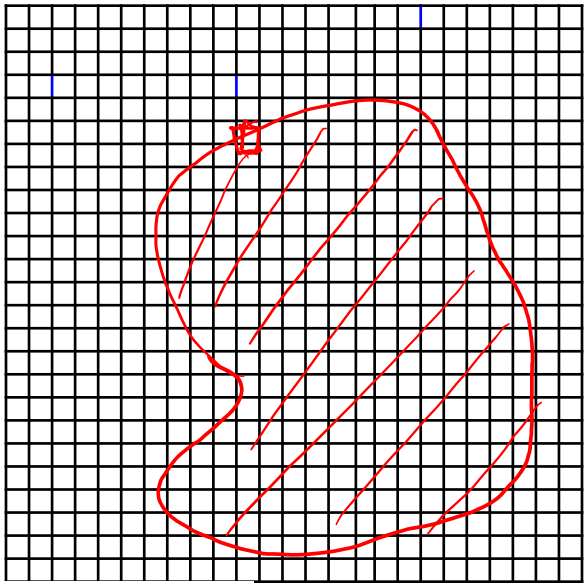


Adapt from Prof. Nando de Freitas's review slides

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

$$prob = \frac{\#\,\mathrm{Re}\,dBoxes}{\#\,Boxes}$$

Adapt from Prof. Nando de Freitas's review slides

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, e.g., Independence, conditional independence
- Maximum Likelihood Estimation

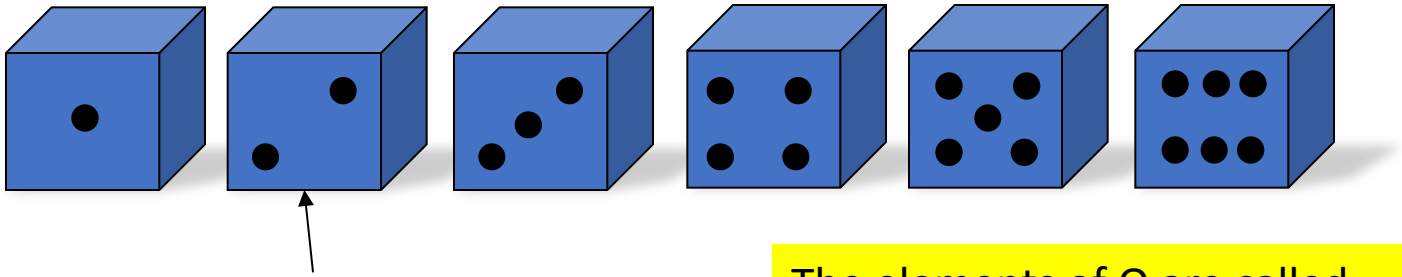
Probability

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 2"

The elements of O 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.,***

$$P(\text{EVEN}) = 1/2$$

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 $\mathrm{O} = \{\mathrm{HH},\mathrm{HT},\mathrm{TH},\mathrm{TT}\}$
- Event: a subset of O
  - First toss is head = {HH,HT}
- S: event space, a set of events:
  - Contains the empty event and O

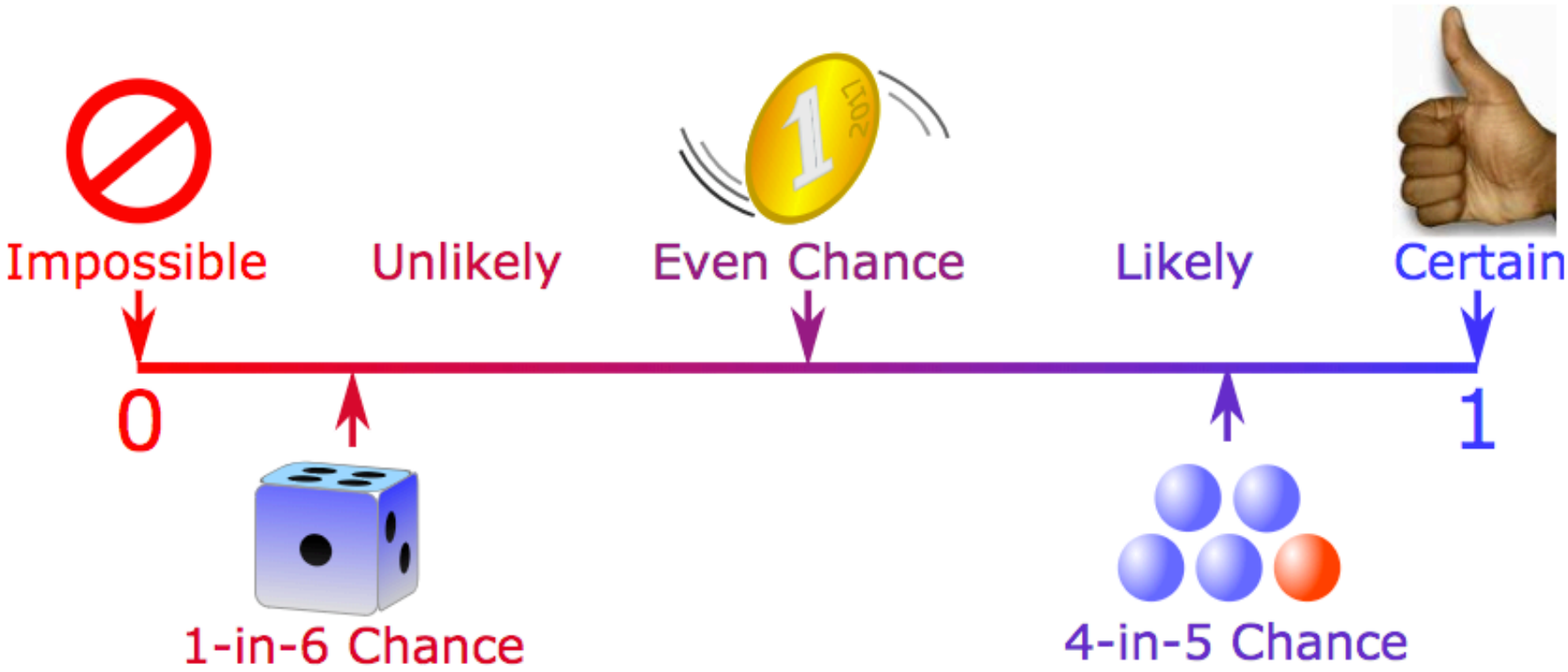
Axioms for Probability

sample space

event space

- Defined over (O,S) s.t.
  - 1 >= P(α) >= 0 for all α in S
  - P(O) = 1

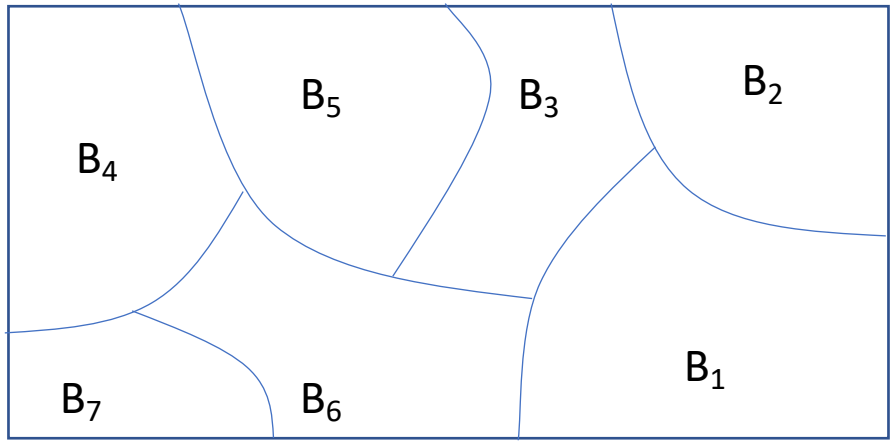
  - If A, B are disjoint, then
    - P(A U B) = p(A) + p(B)



Probability is always between 0 and 1

Axioms for Probability

- $P(O) = \sum P(B_i)$ $= 1$

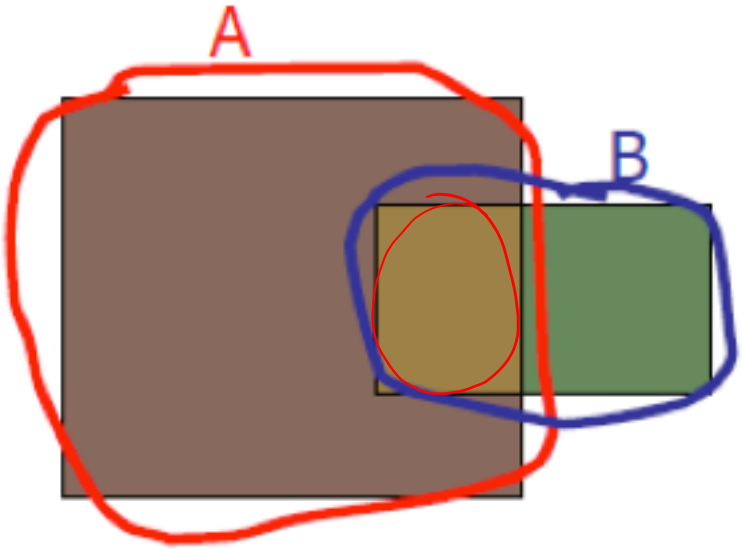


OR operation for Probability

- 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 set and B set)

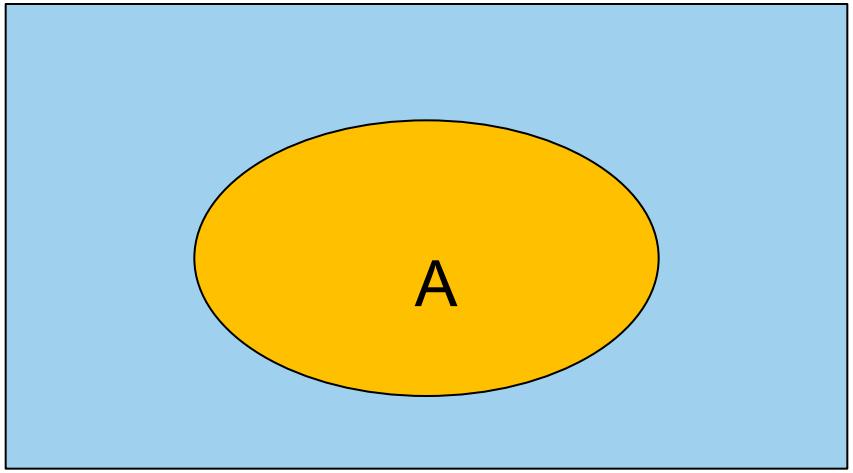


NOT operation for Probability

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

From these we can prove:

P(not A) = P(~A) = 1-P(A)



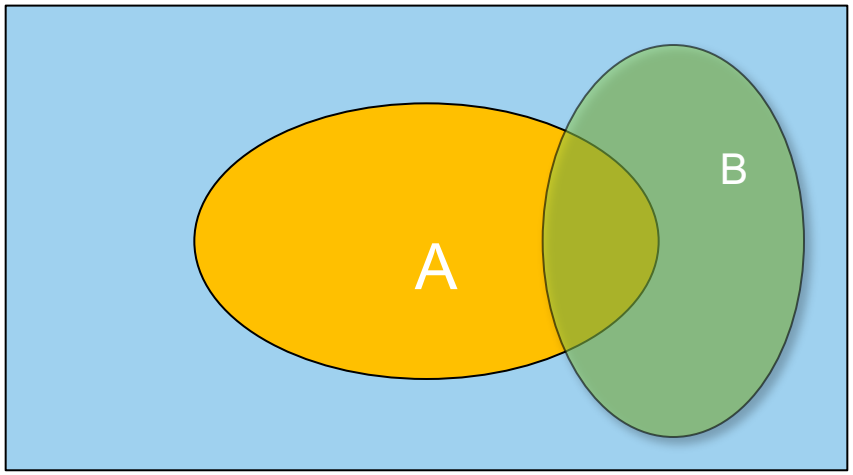
Law of Total Probability

- 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)



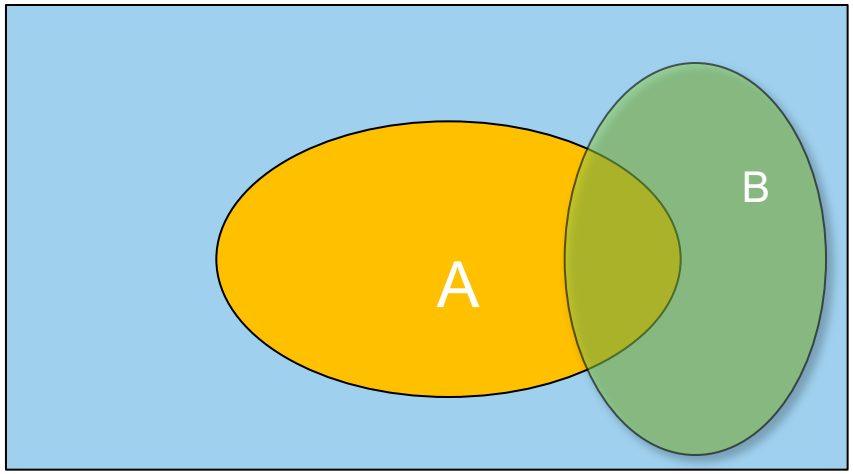
Law of Total Probability

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

From these we can prove:

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

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



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, e.g., Independence, conditional independence
- Maximum Likelihood Estimation

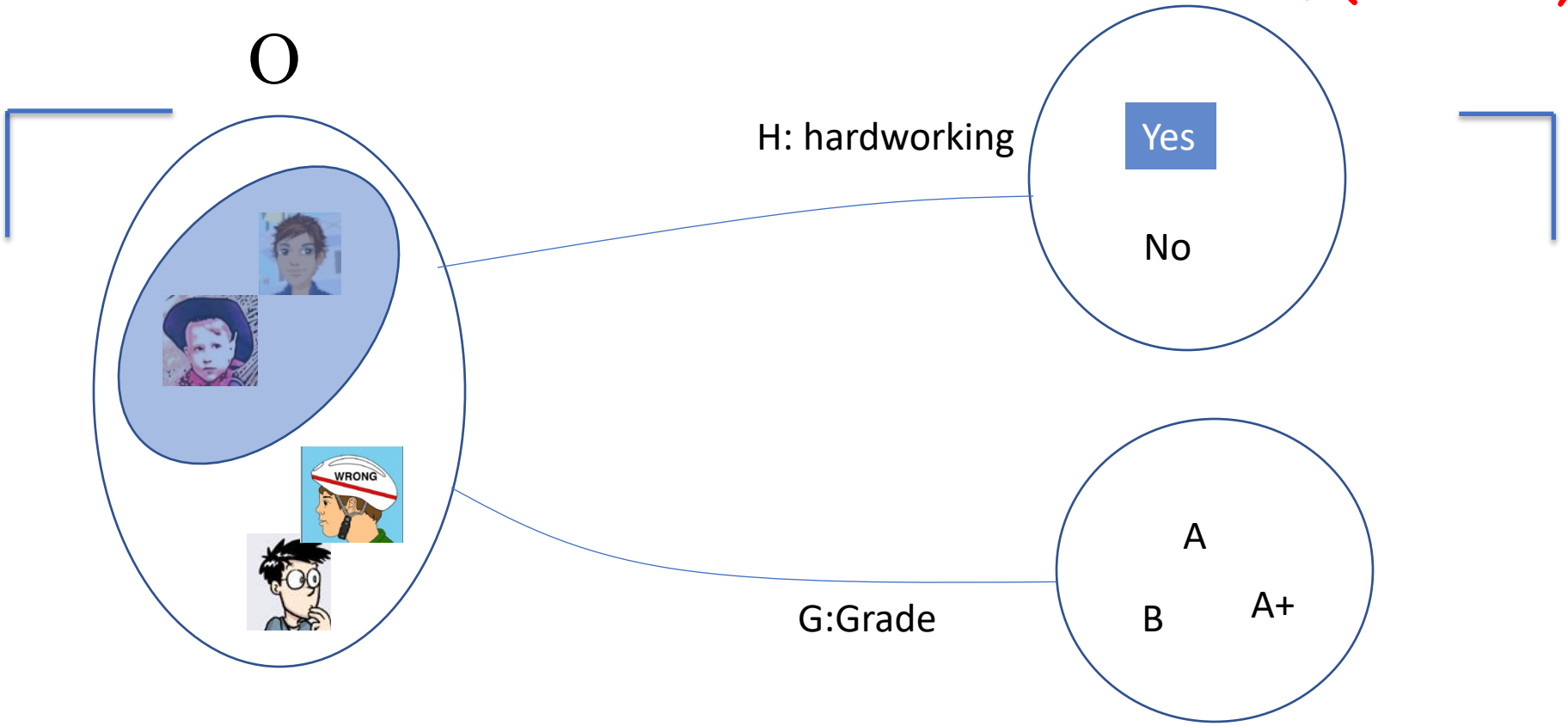
From Events to Random Variable (RV)

- Concise way of specifying attributes of outcomes
- Modeling students (Grade and Intelligence):
  - $\mathrm{O} =$ 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(H = Yes) = P({all students who is working hard on the course})

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

Notations

- 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

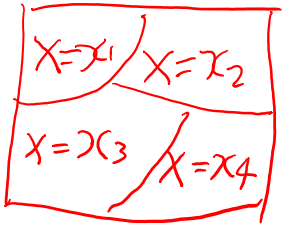
- Random variables (RVs) which may take on only a **countable** number of **distinct** values

- 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): $P(X = x_i)$
- 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 $i \neq j$
  - $P(X = x_1 \cup X = x_2 \cup \ldots \cup X = x_k) = 1$

$$\sum_{i=1}^{4} P(X = x_i) = 1$$



e.g. Coin Flips

- You flip a coin
  - Head with probability p, e.g. =0.5

- You flip a coin for *k*, e.g., =100 times
  - How many heads would you expect

e.g. Coin Flips cont.

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

Binary = {H, T}

P(#Heads)

Integer {1, 2, ..., k}

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
  - E.g. the possible values that X can take on are 0, 1, 2,…, 100

$$\{x_1, \ldots, x_k\}$$

e.g., two Common Distributions

- Uniform
  - X takes values 1, 2, …, *N*
  - E.g. picking balls of different colors from a box

$$X \sim U[1,\ldots,N]$$

$$P(X=i)=1/N$$

- Binomial
  - X takes values 0, 1, …, *k*
  - E.g. coin flips k times

$$X \sim Bin(k,p)$$

$$P(X=i)=\binom{k}{i}p^{i}(1-p)^{k-i}$$

: heads out k

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

If hard to directly estimate from data, most likely we can estimate

- 1. Joint probability
  - Use Chain Rule

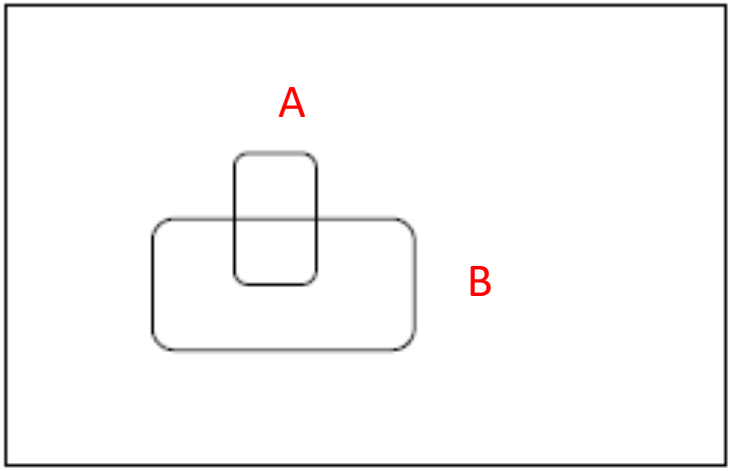
$$P(A, B) = P(B)\, P(A|B)$$

- 2. Marginal probability
  - Use the total law of probability

$$P(B) = P(B, A) + P(B, \sim A)$$
$$= P(B, A \cup \sim A)$$

- 3. Conditional probability
  - Use the Bayes Rule

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



Conditional / Joint / Marginal 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.

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:

Chain rule

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

from Prof. Nando de Freitas's review

(1). To calculate Joint Probability: Use Chain Rule

- Two ways to use chain rules on joint probability

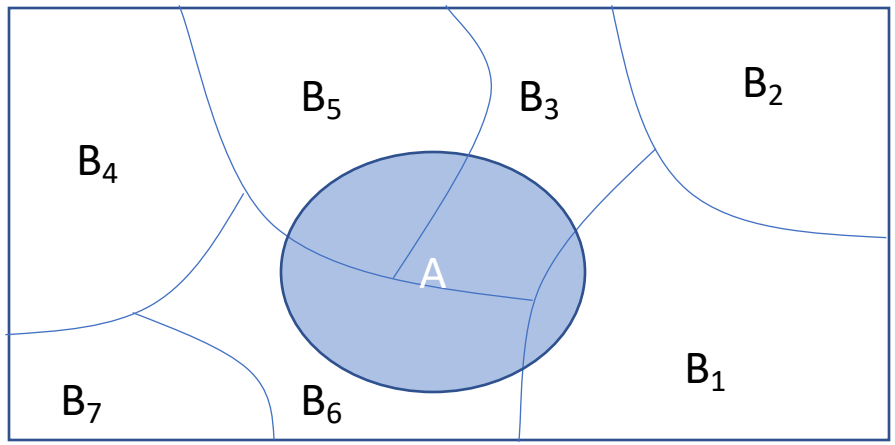
joint conditional

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

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

(2). To calculate Marginal Probability:

Use Rule of total probability (e.g. event version)



$P(A) =$

$P(A \wedge B) +$

$P(A \wedge \sim B)$

$$\Rightarrow p(A) = \sum P(B_i)\overbrace{P(A \mid B_i)}^{P(B_i \wedge A)}$$

WHY ???

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

$= P(A \cap (B_1 \cup B_2 \cdots \cup B_k))$

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

(2). To calculate Marginal Probability:
Use Rule of total probability (e.g. RV version)

- Given two discrete RVs X and Y, which take values in:

$$\{x_1,\ldots,x_k\} \qquad \{y_1,\ldots,y_m\}$$

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

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

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

(3). To calculate Conditional Probability: Use Bayes Rule (e.g. RV version)

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

One 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?*

$$\boldsymbol{P(B_1 = r, B_2 = r)} = \underbrace{P(B_1=r)}_{\frac{3}{4}} \underbrace{P(B_2=r \mid B_1=r)}_{\times \frac{2}{3}} = \frac{1}{2}$$

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

$$P(B_1 = r | B_2 = r) = \frac{P(B_1=r, B_2=r)}{P(B_2=r)}$$

One Example: Joint

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?*

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

Adapt from Prof. Nando de Freitas's review slides

One Example: Joint

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

One Example: Joint

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

One Example: Marginal

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

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

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

One Example: Marginal

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

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

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

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

One Example: Conditional

Chain Rule
total law Prob

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

$$= \frac{P(B_2 = r \mid B_1 = r)\, P(B_1 = r)}{P(B_2 = r)}$$

last
last
page

Last

$$= \frac{P(B_2 = r \mid B_1 = r)\, P(B_1 = r)}{P(B_2 = r, B_1 = r) + P(B_2 = r, B_1 = b)}$$

Bayes Rule

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

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**

if $P(Y=yes|x) > P(Y=No|x)$

$$\frac{P(Y=yes)\,P(x|Y=yes)}{P(x)}$$

$$\frac{P(Y=No)\,P(x|Y=No)}{P(x)}$$

$\Rightarrow \hat{y} = yes$

More General Forms of Bayes Rule

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

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

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

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

E.g.: Use both Bayes Rule and Marginal

- X and Y are discrete RVs…

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

$$\{x_1, \ldots, x_k\}$$

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

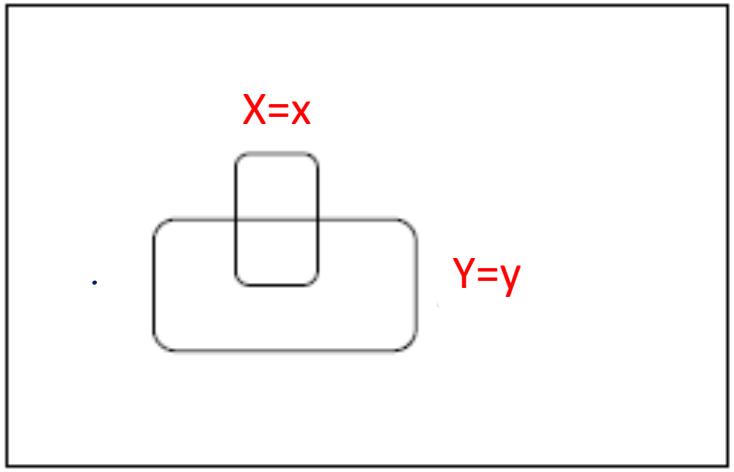
Simplify Notation: Conditional Probability

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)

$P(x) \leftarrow P(X=x) \leftarrow P(X=x \text{ true})$

value RV event

Simplify Notation:
An Example of estimating conditional

- 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?

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

W= G=

Simplify Notation:
An Example of estimating conditional

- 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=

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

$P(W=S \mid wet)$

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

Simplify Notation: Conditional

- Bayes Rule

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

- You can condition on more variables

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

Simplify Notation: Marginal

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

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

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

$$\{y_1, \ldots, y_m\}$$

all possible Y values

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

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

$$\sum_{y} \sum_{z} p(y, z) = 1$$

Simplify Notation:
An Example

- 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?

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

0.5 1

Weather Grass

{rain, sunny} {wet, dry}

P(wet, rain)

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

P(rain) P(wet | rain) + P(sunny) P(wet | sunny)

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, e.g., Independence, conditional independence
- Maximum Likelihood Estimation

Independent RVs

- Definition: X and Y are independent ***iff***

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

More on Independence

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

$$P(X = x | Y = y) = P(X = x)$$

$$P(Y = y | X = x) = P(Y = y)$$

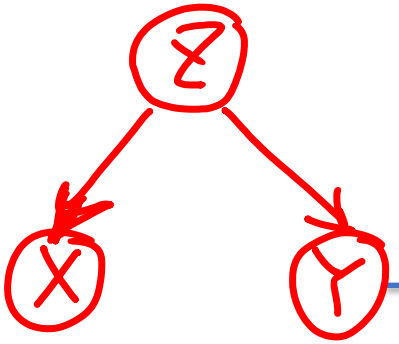
- E.g. no matter how many heads you get, your friend will not be affected, and vice versa

More on Independence

- X is independent of Y means that knowing Y does not change our belief about X.
- The following forms are equivalent:
  - $P(X=x, Y=y) = P(X=x)\, P(Y=y)$
  - $P(X=x|Y=y) = P(X=x)$

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

Conditionally Independent RVs

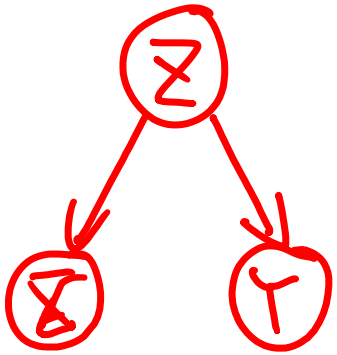


- Intuition: X and Y are conditionally independent given Z means that once Z is **known**, the value of X does not add any **additional** information about Y
- Definition: X and Y are conditionally independent given Z ***iff***

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

If holding for all x_i, y_j, z_k

$$X \perp Y \mid Z$$



More on Conditional Independence

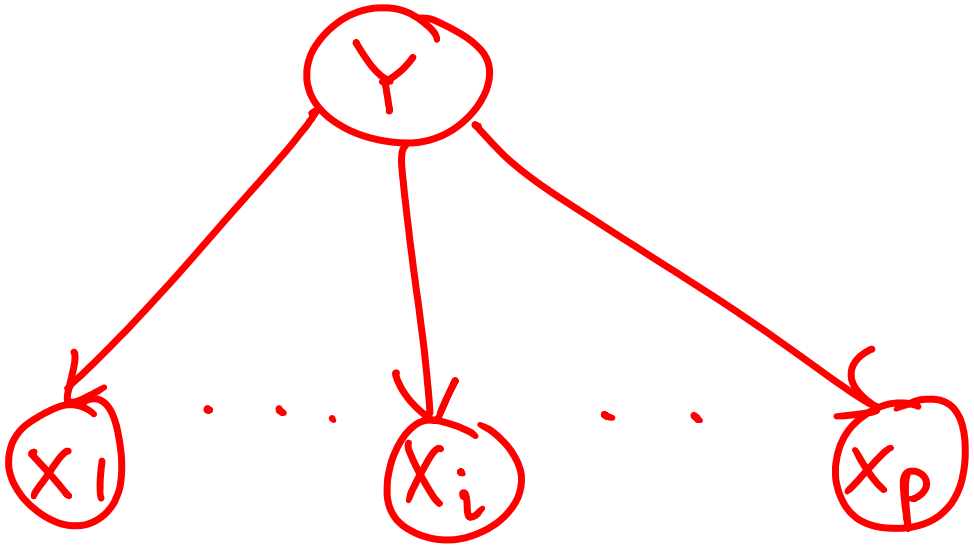
$$P\left(X = x \cap Y = y \middle| Z = z\right) = P\left(X = x \middle| Z = z\right) P\left(Y = y \middle| Z = z\right)$$

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

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

independence and conditional independence

- Independence does not imply conditional independence.
- Conditional independence does not imply independence.



Today Recap: 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, e.g., Independence, conditional independence
- Maximum Likelihood Estimation (next class)

References

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