

Probability Review

CMSC 473/673
UMBC

Frank Ferraro – ferraro@umbc.edu

Probability Prerequisites

Basic probability axioms and definitions Probabilistic Independence

Expected Value (of a function) of a Random Variable

Definition of conditional probability

Joint probability

Bayes rule

Marginal probability

Probability chain rule

(Most) Probability Axioms

$$p(\text{everything}) = 1$$

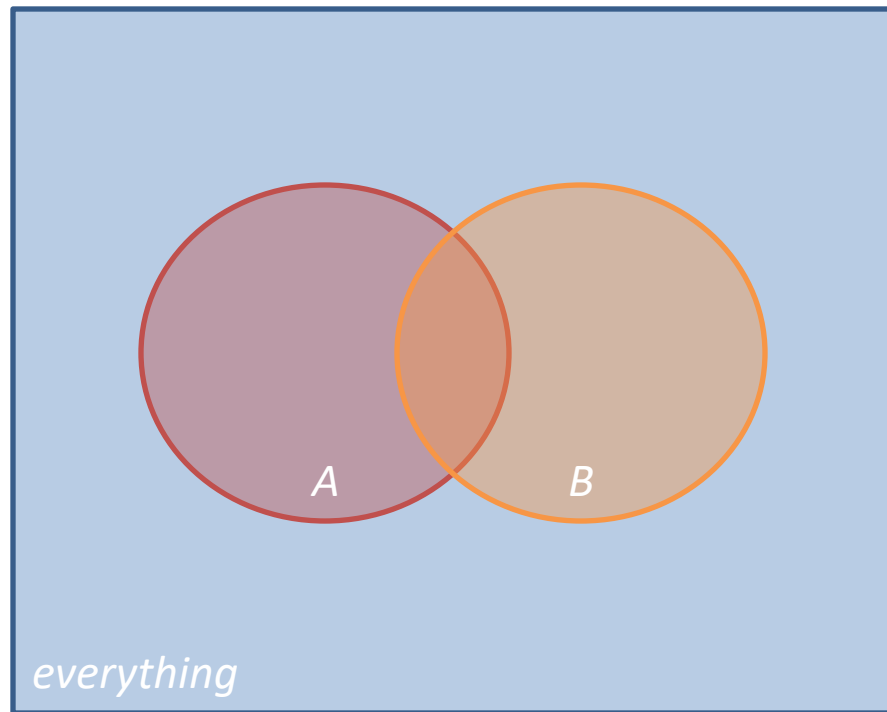
$$p(\text{nothing}) = p(\phi) = 0$$

empty set

$$p(A) \leq p(B), \text{ when } A \subseteq B$$

$$p(A \cup B) = p(A) + p(B),$$

when $A \cap B = \phi$



$$p(A \cup B) \neq p(A) + p(B)$$

$$p(A \cup B) = p(A) + p(B) - p(A \cap B)$$

Examining $p(\text{everything}) = 1$

If $p(\text{everything}) = 1 \dots$

Examining $p(\text{everything}) = 1$

If $p(\text{everything}) = 1 \dots$

and you can break *everything* into M unique items

$x_1, x_2, \dots, x_M \dots$

Examining $p(\text{everything}) = 1$

If $p(\text{everything}) = 1 \dots$

and you can break **everything** into M unique items
 $x_1, x_2, \dots, x_M \dots$

then each pair x_i and x_j are disjoint ($x_i \cap x_j = \phi$)...

Examining $p(\text{everything}) = 1$

If $p(\text{everything}) = 1 \dots$

and you can break **everything** into M unique items

$x_1, x_2, \dots, x_M \dots$

then each pair x_i and x_j are disjoint ($x_i \cap x_j = \phi$)...

and because **everything** is the union of

$x_1, x_2, \dots, x_M \dots$

Examining $p(\text{everything}) = 1$

If $p(\text{everything}) = 1 \dots$

and you can break **everything** into M unique items

$x_1, x_2, \dots, x_M \dots$

then each pair x_i and x_j are disjoint ($x_i \cap x_j = \phi$)...

and because **everything** is the union of $x_1, x_2, \dots, x_M \dots$

$$p(\text{everything}) = \sum_{i=1}^M p(x_i) = 1$$

A Very Important Concept to Remember

The probabilities of all unique (disjoint) items $x_1, x_2, \dots, x_M$ *must* sum to 1:

$$p(\text{everything}) = \sum_{i=1}^M p(x_i) = 1$$

Probabilities and Random Variables

Random variables: variables that represent the possible outcomes of some random “process”

Probabilities and Random Variables

Random variables: variables that represent the possible outcomes of some random “process”

Example #1: A (weighted) coin that can come up heads or tails

X is a random variable denoting the possible outcomes

$X = \text{HEADS}$ or $X = \text{TAILS}$

Distribution Notation

If X is a R.V. and G is a distribution:

- $X \sim G$ means X is distributed according to (“sampled from”) G

Distribution Notation

If X is a R.V. and G is a distribution:

- $X \sim G$ means X is distributed according to (“sampled from”) G
- G often has parameters $\rho = (\rho_1, \rho_2, \dots, \rho_M)$ that govern its “shape”
- Formally written as $X \sim G(\rho)$

Distribution Notation

If X is a R.V. and G is a distribution:

- $X \sim G$ means X is distributed according to (“sampled from”) G
- G often has parameters $\rho = (\rho_1, \rho_2, \dots, \rho_M)$ that govern its “shape”
- Formally written as $X \sim G(\rho)$

i.i.d. If $X_1, X_2, \dots, X_N$ are all independently sampled from $G(\rho)$, they are independently and identically distributed

Probability Prerequisites

Basic probability axioms and Probabilistic Independence definitions

Expected Value (of a function) of a Random Variable

Definition of conditional probability

Bayes rule

Joint probability

Probability chain rule

Marginal probability

Expected Value of a Random Variable

random variable


$$X \sim p(\cdot)$$

Expected Value of a Random Variable

random variable

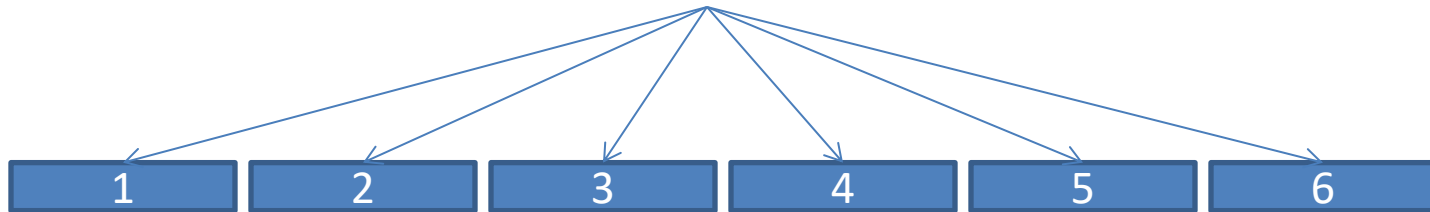
$$X \sim p(\cdot)$$

$$\mathbb{E}[X] = \sum_x x p(x)$$

*expected value
(distribution p is
implicit)*

Expected Value: Example

uniform distribution of number of cats I have

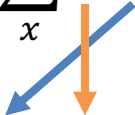


$$\mathbb{E}[X] = \sum_x x p(x)$$
$$\begin{aligned} & 1/6 * 1 + \\ & 1/6 * 2 + \\ & 1/6 * 3 + \\ & 1/6 * 4 + \\ & 1/6 * 5 + \\ & 1/6 * 6 \end{aligned} = 3.5$$

Expected Value: Example 2

non-uniform distribution of number of cats a normal cat person has



$$\mathbb{E}[X] = \sum_x x p(x)$$


$$\begin{aligned} &1/2 * 1 + \\ &1/10 * 2 + \\ &1/10 * 3 + \\ &1/10 * 4 + \\ &1/10 * 5 + \\ &1/10 * 6 \end{aligned} = 2.5$$

Expected Value of a Function of a Random Variable

$$X \sim p(\cdot)$$

$$\mathbb{E}[X] = \sum_x x p(x)$$

$$\mathbb{E}[f(X)] = ???$$

Expected Value of a Function of a Random Variable

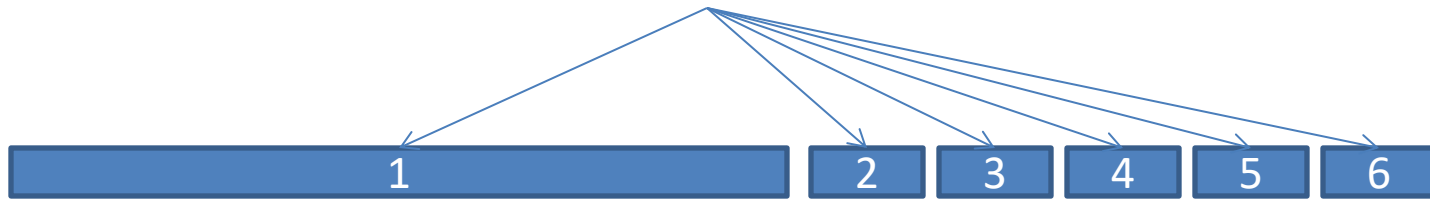
$$X \sim p(\cdot)$$

$$\mathbb{E}[X] = \sum_x x p(x)$$

$$\mathbb{E}[f(X)] = \sum_x f(x) p(x)$$

Expected Value of Function: Example

non-uniform distribution of number of cats I start with



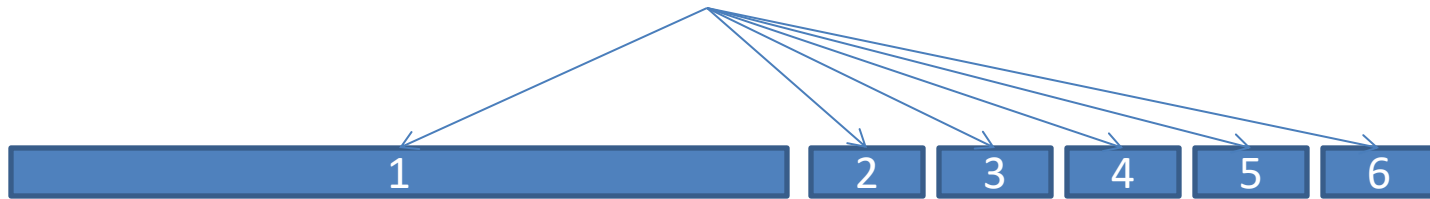
What if each cat magically becomes two?

$$f(k) = 2^k$$

$$\mathbb{E}[f(X)] = \sum_x f(x) p(x)$$

Expected Value of Function: Example

non-uniform distribution of number of cats I start with



What if each cat magically becomes two?

$$f(k) = 2^k$$

$$\mathbb{E}[f(X)] = \sum_x f(x) p(x) = \sum_x 2^x p(x)$$

$$\begin{aligned} & 1/2 * 2^1 + \\ & 1/10 * 2^2 + \\ & 1/10 * 2^3 + \\ & 1/10 * 2^4 + \\ & 1/10 * 2^5 + \\ & 1/10 * 2^6 \end{aligned} = 13.4$$

Probability Prerequisites

Basic probability axioms and Probabilistic Independence definitions

Expected Value (of a function) of a Random Variable

Definition of conditional probability

Bayes rule

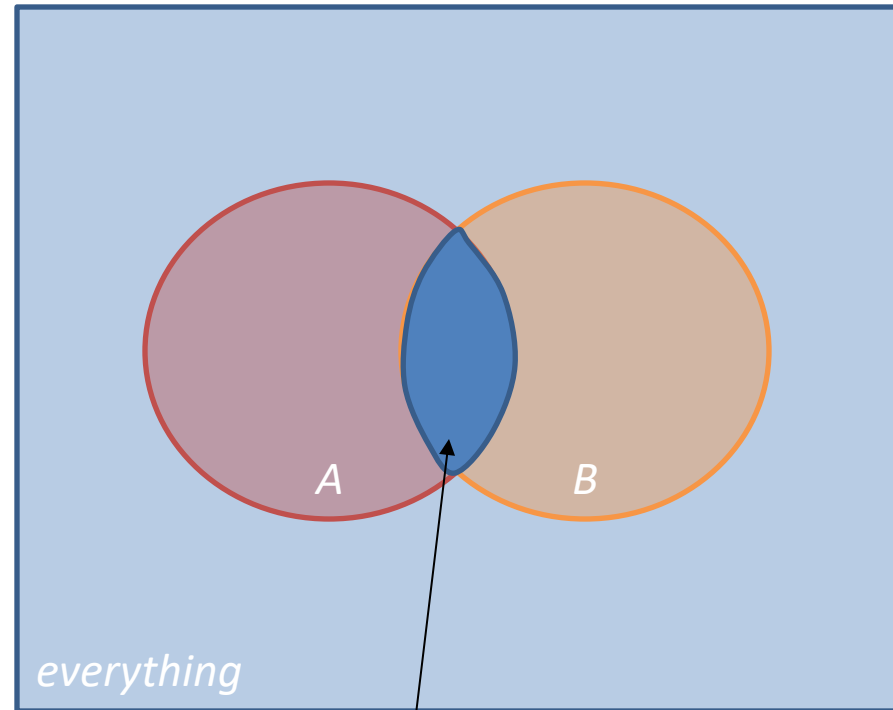
Joint probability

Probability chain rule

Marginal probability

Joint Probability

Probability that multiple things “happen together”



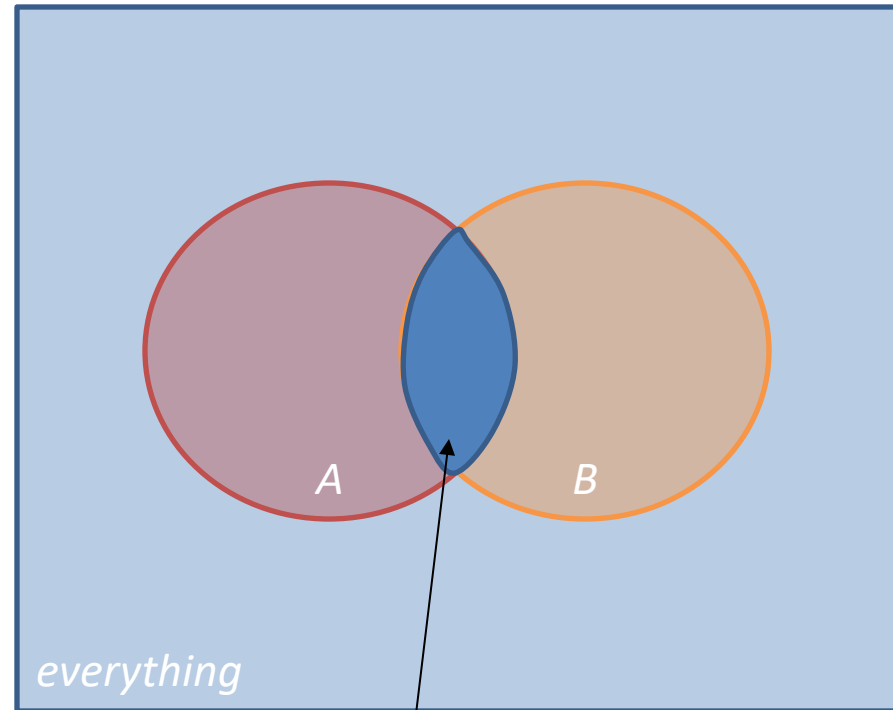
Joint
probability

Joint Probability

Probability that multiple things “happen together”

$p(x,y)$, $p(x,y,z)$, $p(x,y,w,z)$

Symmetric: $p(x,y) = p(y,x)$



Joint
probability

Joint Probability

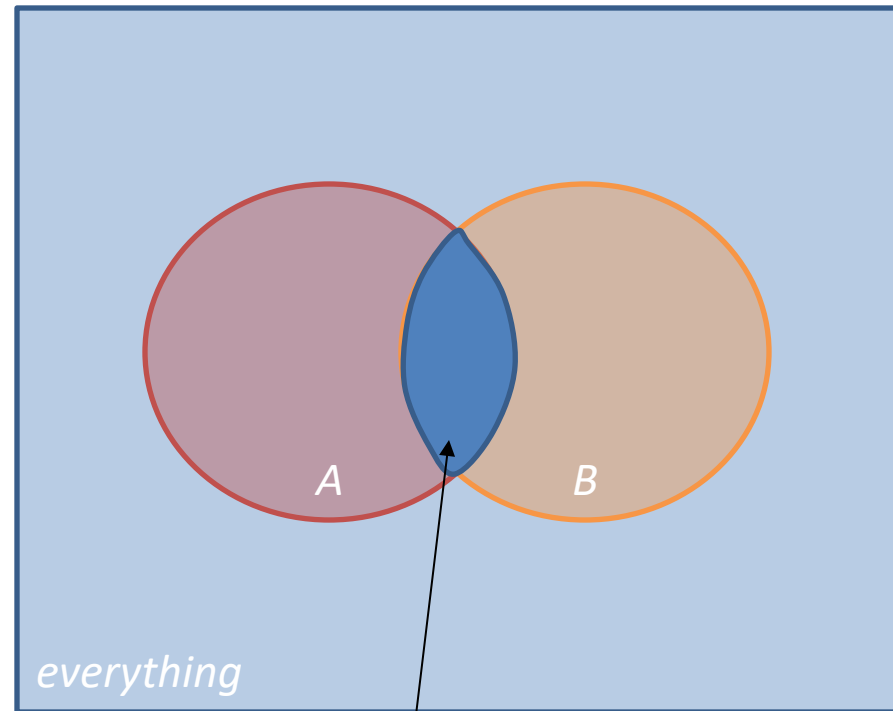
Probability that multiple things
“happen together”

$p(x,y)$, $p(x,y,z)$, $p(x,y,w,z)$

Symmetric: $p(x,y) = p(y,x)$

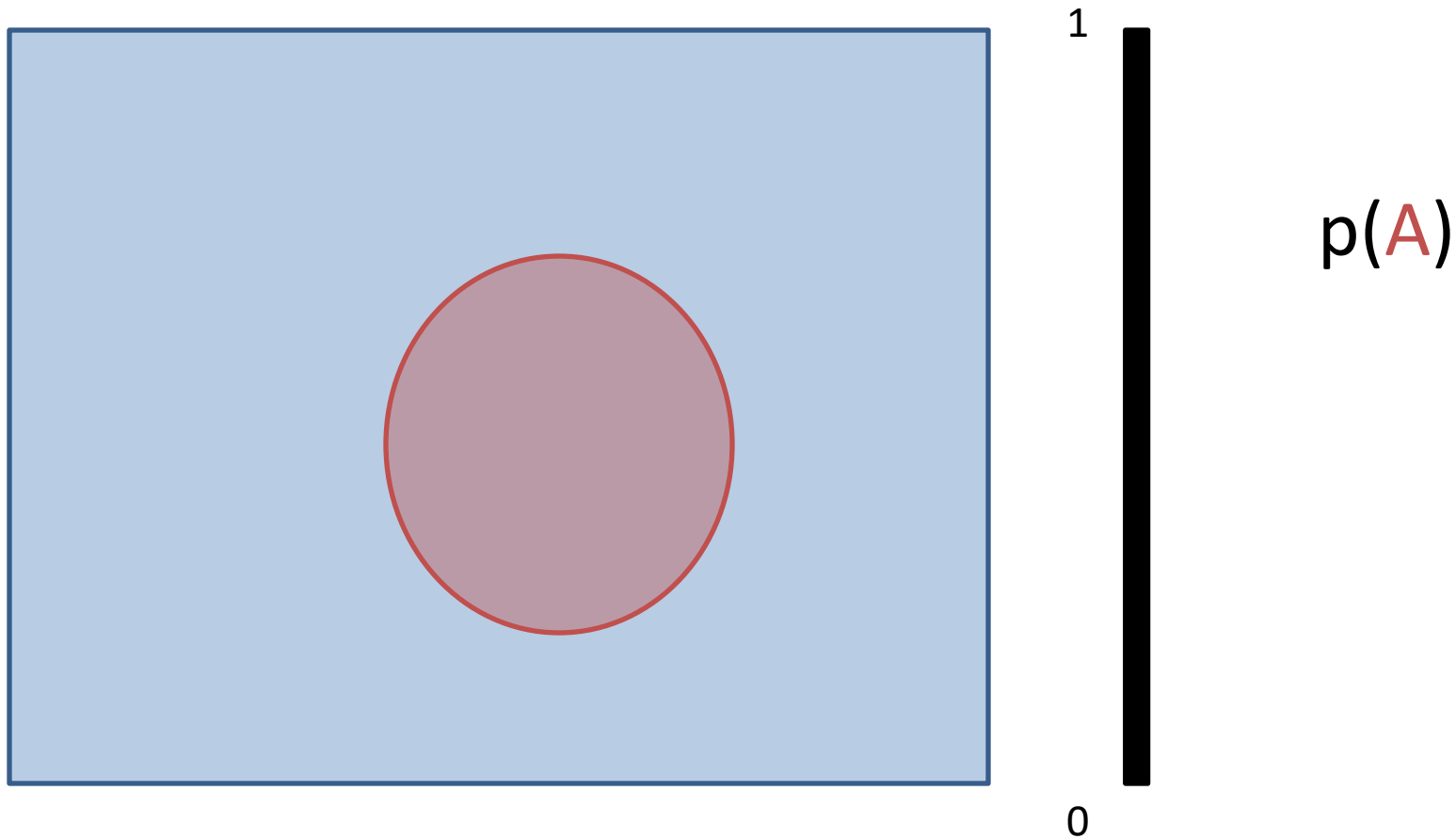
Form a table based of
outcomes: sum across cells = 1

$p(x,y)$	Y=0	Y=1
X="cat"	.04	.32
X="dog"	.2	.04
X="bird"	.1	.1
X="human"	.1	.1



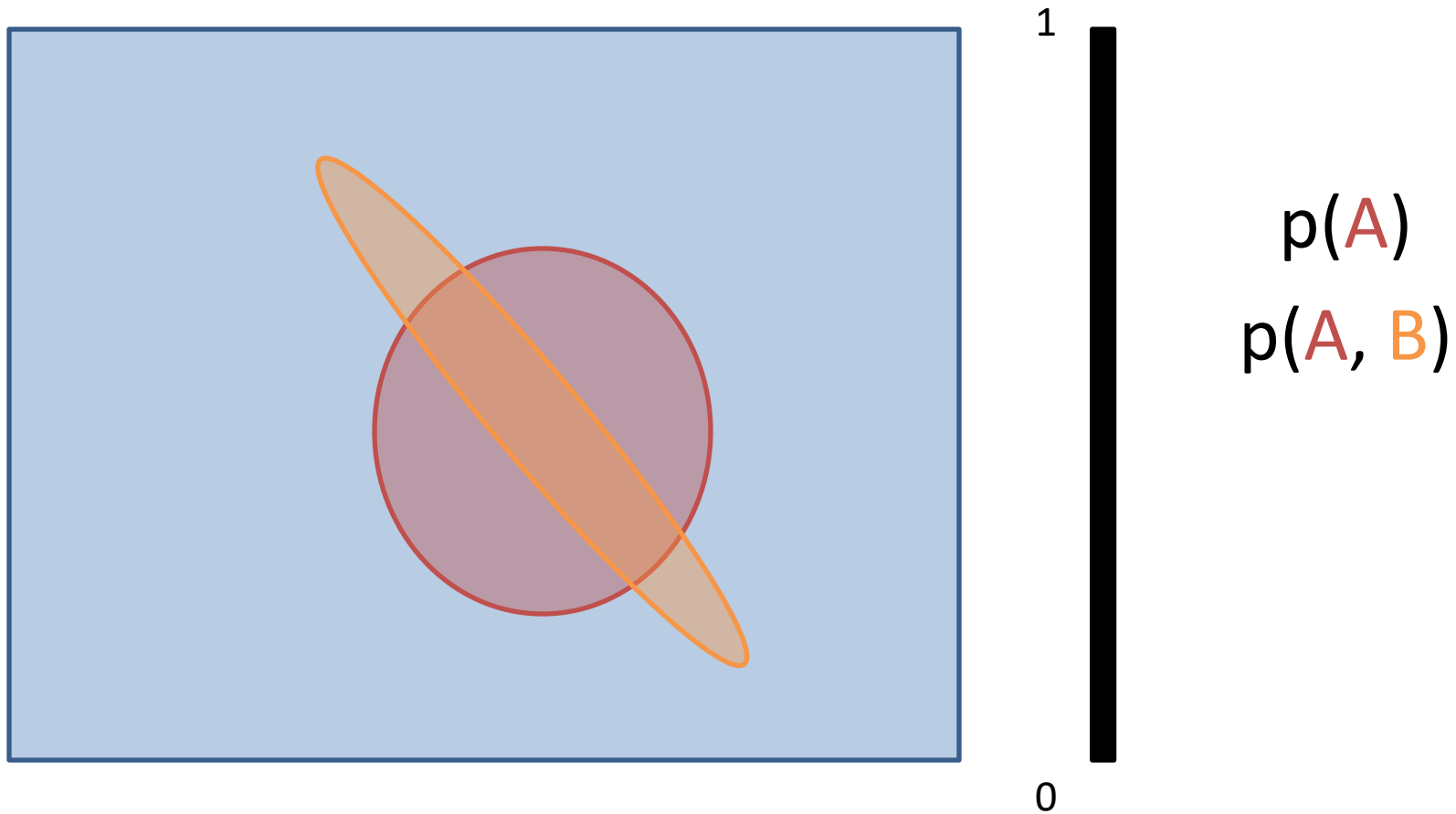
Joint
probability

Joint Probabilities



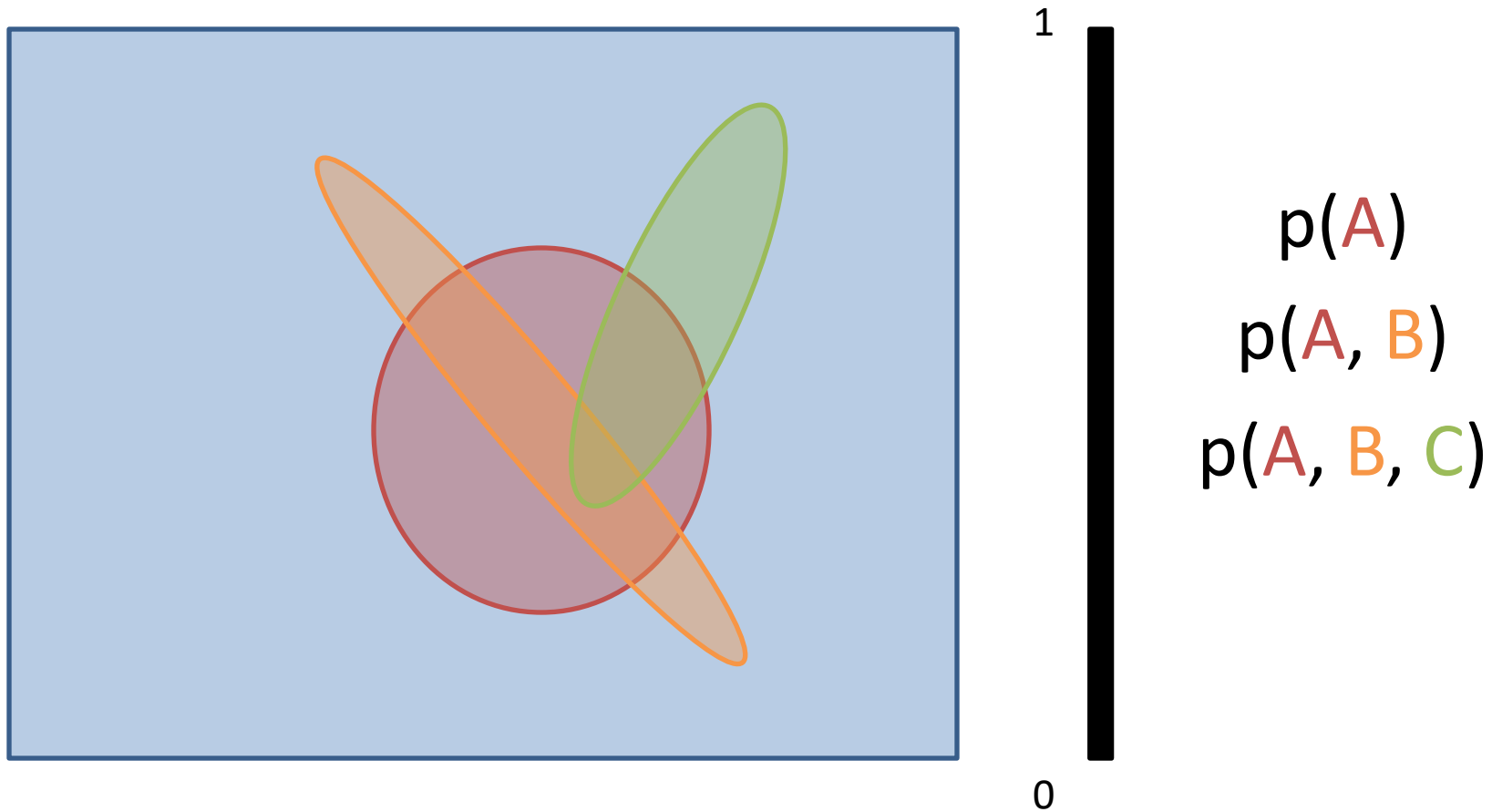
what happens as we add conjuncts?

Joint Probabilities



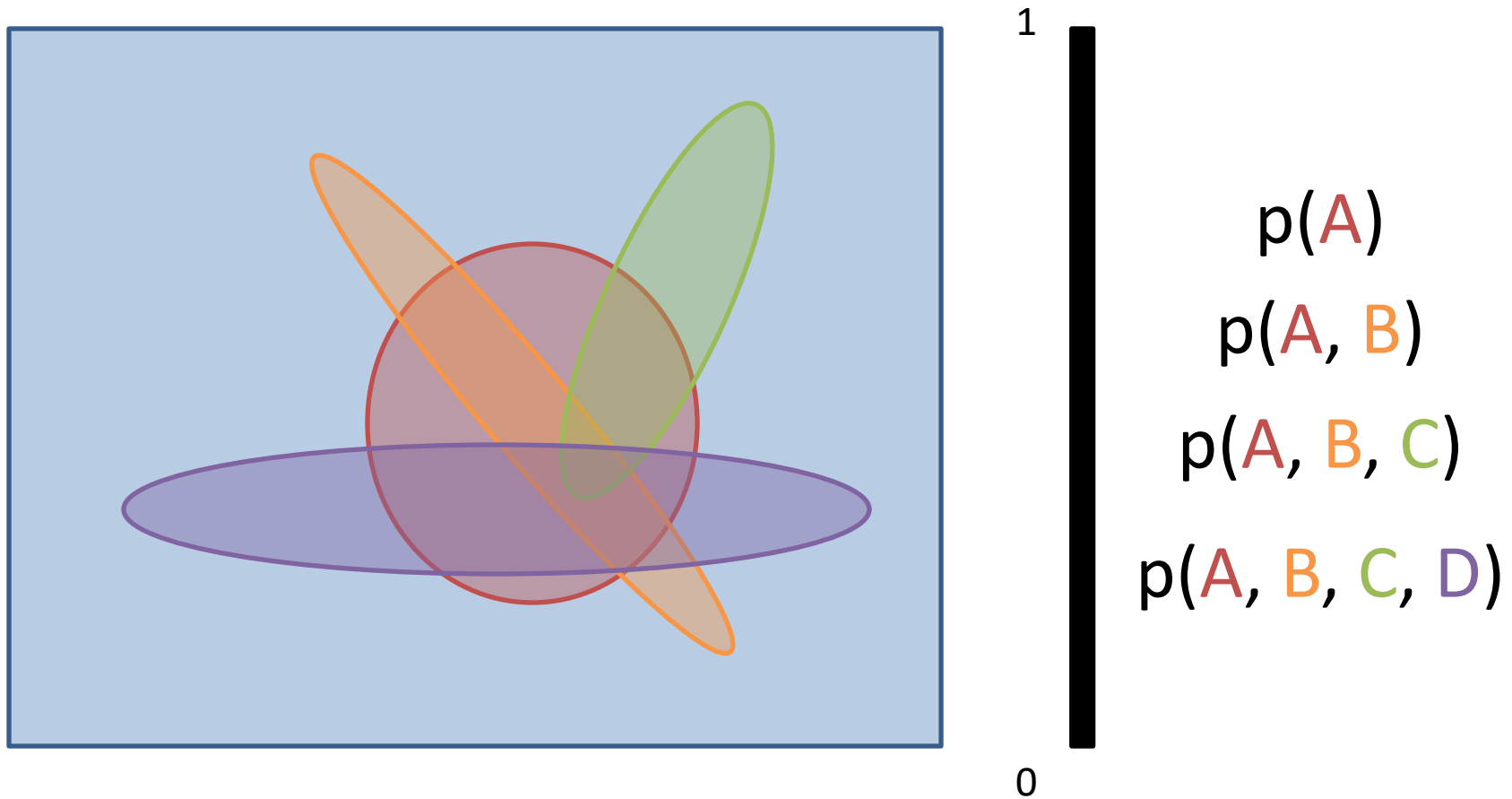
what happens as we add conjuncts?

Joint Probabilities



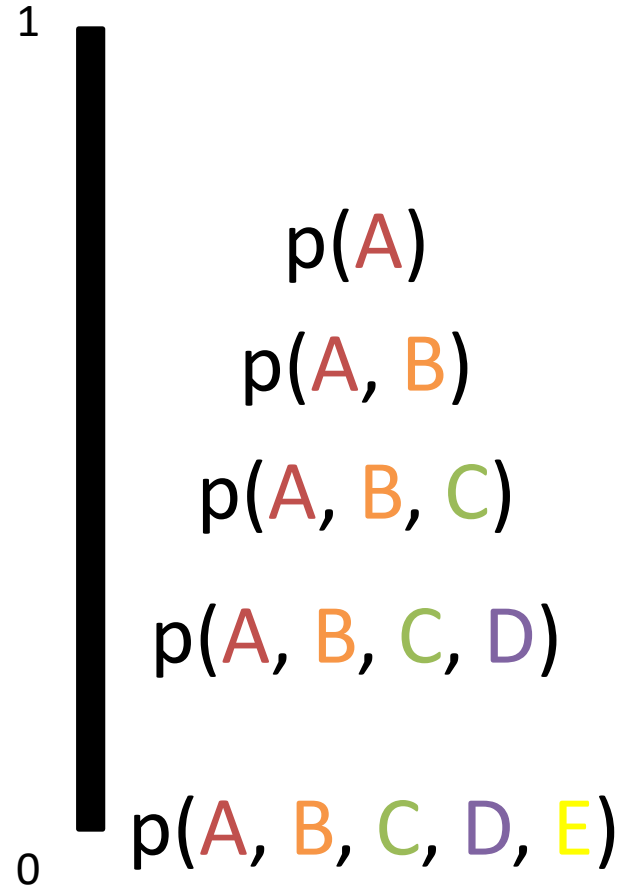
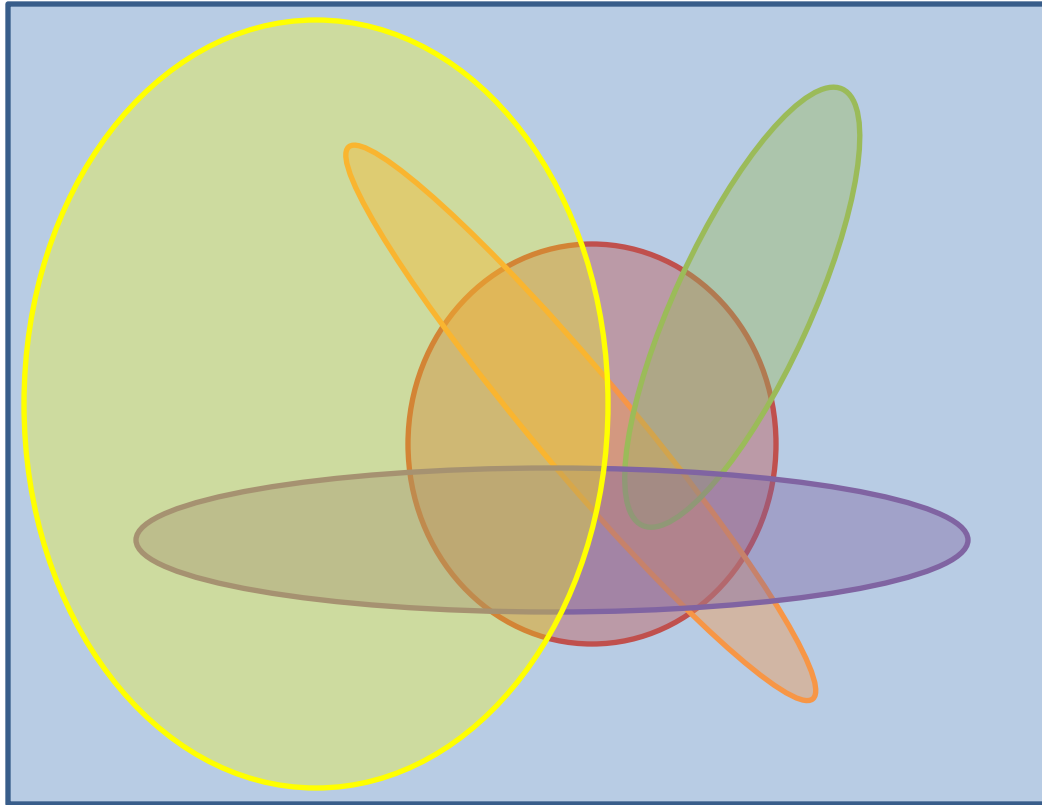
what happens as we add conjuncts?

Joint Probabilities



what happens as we add conjuncts?

Joint Probabilities



what happens as we add conjuncts?

A Note on Notation

$$p(X \text{ *INCLUSIVE_OR* } Y) \Leftrightarrow p(X \cup Y)$$

$$p(X \text{ *AND* } Y) \Leftrightarrow p(X, Y)$$

$$p(X, Y) = p(Y, X)$$

- except when order matters (should be obvious from context)

Probability Prerequisites

Basic probability axioms and Probabilistic Independence definitions

Expected Value (of a function) of a Random Variable

Definition of conditional probability

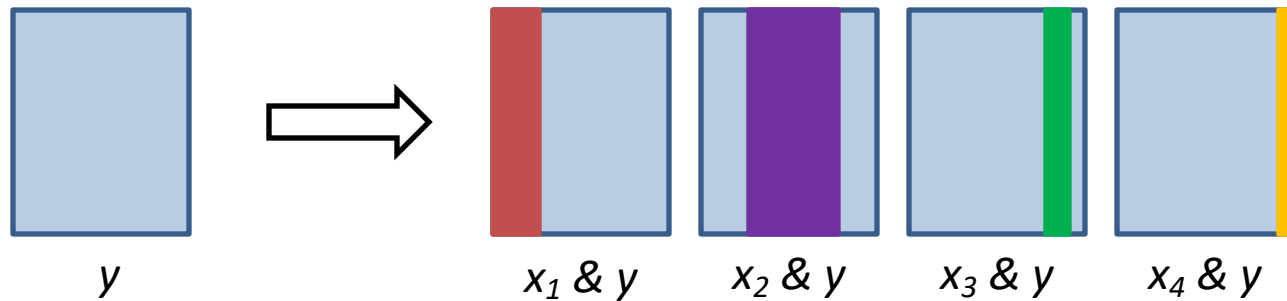
Bayes rule

Joint probability

Probability chain rule

Marginal probability

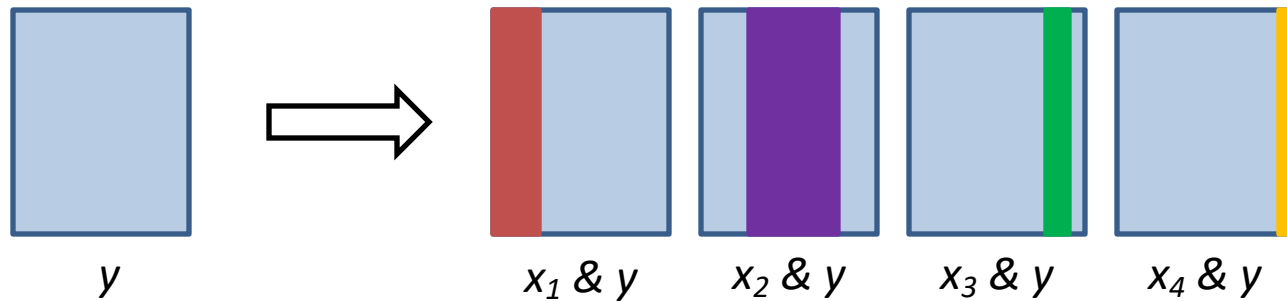
Marginal(ized) Probability: The Discrete Case



Consider the **mutually exclusive** ways
that different values of x could occur
with y

Q: How do write this in
terms of joint probabilities?

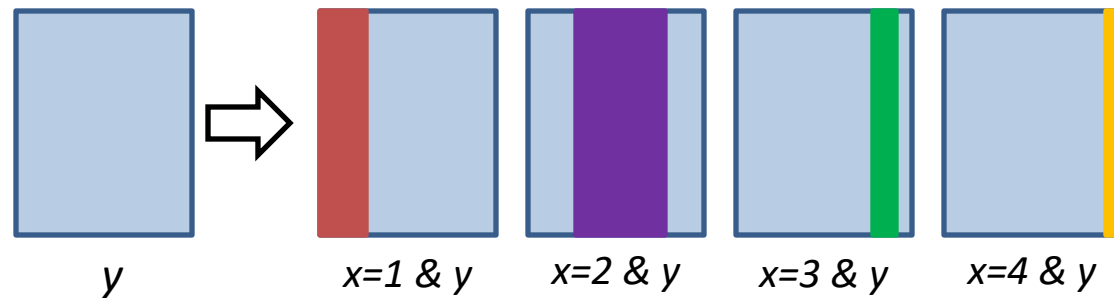
Marginal(ized) Probability: The Discrete Case



Consider the **mutually exclusive** ways
that different values of x could occur
with y

$$p(y) = \sum_x p(x, y)$$

Marginal(ized) Probability: The Discrete Case



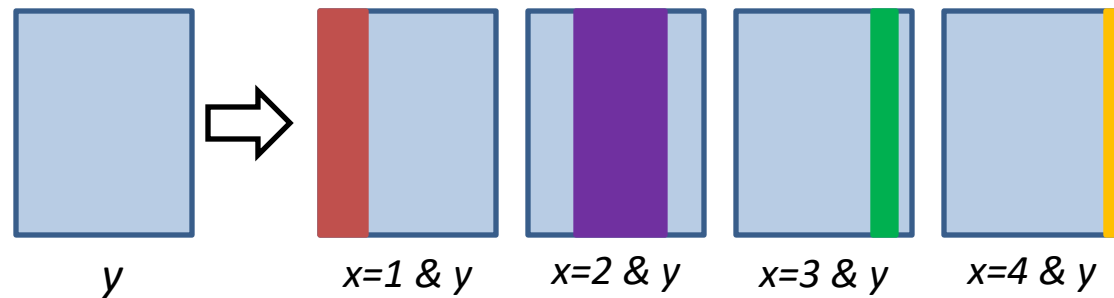
Consider the **mutually exclusive** ways
that different values of x could occur
with y

$$p(y) = \sum_x p(x, y)$$

$p(x,y)$	Y=0	Y=1
X="cat"	.04	.32
X="dog"	.2	.04
X="bird"	.1	.1
X="human"	.1	.1

Q: What is $p(y=1)$?

Marginal(ized) Probability: The Discrete Case



Consider the **mutually exclusive** ways
that different values of x could occur
with y

$$p(y) = \sum_x p(x, y)$$

$p(x,y)$	Y=0	Y=1
X="cat"	.04	.32
X="dog"	.2	.04
X="bird"	.1	.1
X="human"	.1	.1

Q: What is $p(y=1)$?

A: 0.56

Probability Prerequisites

Basic probability axioms and **Probabilistic Independence**
definitions

Expected Value (of a
function) of a Random
Variable

Definition of conditional
probability

Bayes rule

Joint probability

Probability chain rule

Marginal probability

Probabilistic Independence

Independence: when events can occur and not impact the probability of other events

Q: Are the results of flipping the same coin twice in succession independent?

Formally: $p(x,y) = p(x)*p(y)$

Generalizable to > 2 random variables

Probabilistic Independence

Independence: when events can occur and not impact the probability of other events

Formally: $p(x,y) = p(x)*p(y)$

Generalizable to > 2 random variables

Q: Are the results of flipping the same coin twice in succession independent?

A: Yes (assuming no weird effects)

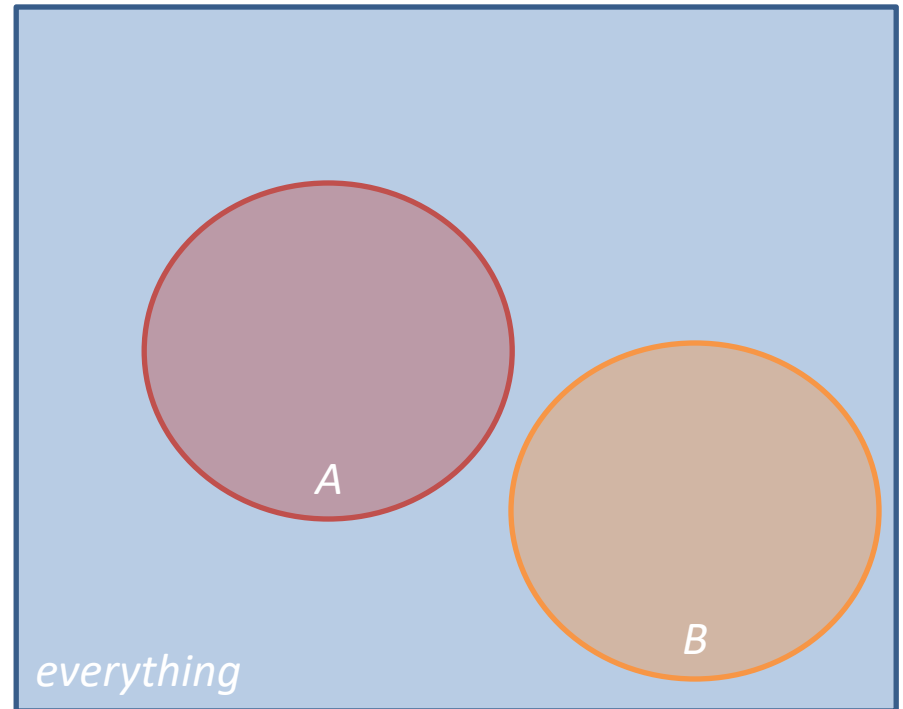
Probabilistic Independence

Independence: when events can occur and not impact the probability of other events

Formally: $p(x,y) = p(x)*p(y)$

Generalizable to > 2 random variables

Q: Are A and B independent?



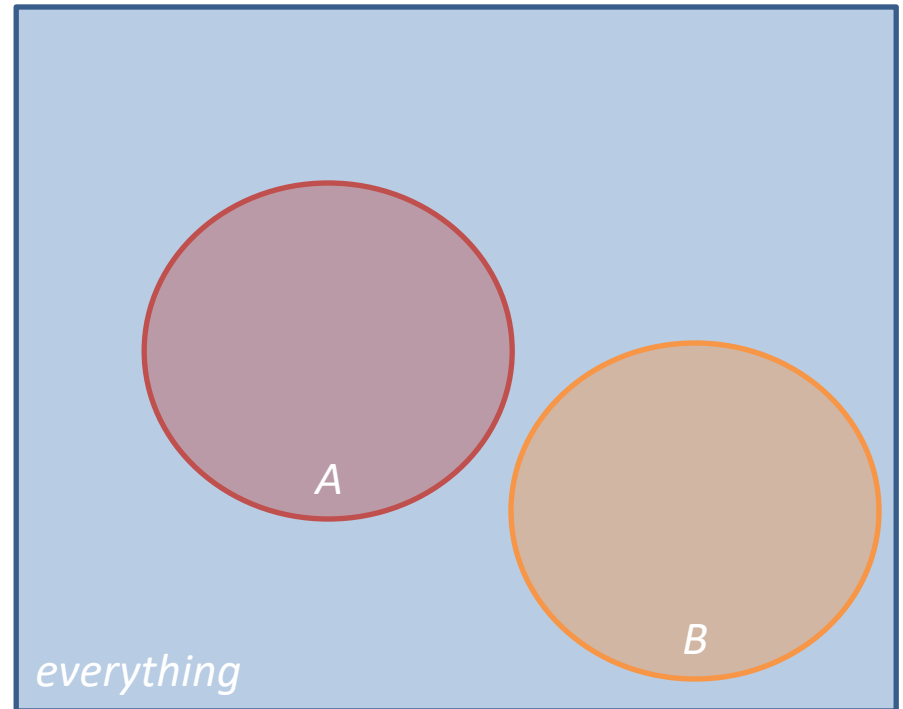
Probabilistic Independence

Independence: when events can occur and not impact the probability of other events

Formally: $p(x,y) = p(x) * p(y)$

Generalizable to > 2 random variables

Q: Are A and B independent?



A: No (work it out from $p(A,B)$ and the axioms)

Probabilistic Independence

Independence: when events can occur and not impact the probability of other events

Formally: $p(x,y) = p(x) * p(y)$

Generalizable to > 2 random variables

Q: Are X and Y independent?

$p(x,y)$	Y=0	Y=1
X="cat"	.04	.32
X="dog"	.2	.04
X="bird"	.1	.1
X="human"	.1	.1

Probabilistic Independence

Independence: when events can occur and not impact the probability of other events

Formally: $p(x,y) = p(x)*p(y)$

Generalizable to > 2 random variables

Q: Are X and Y independent?

$p(x,y)$	Y=0	Y=1
X="cat"	.04	.32
X="dog"	.2	.04
X="bird"	.1	.1
X="human"	.1	.1

A: No (find the marginal probabilities of $p(x)$ and $p(y)$)

Probability Prerequisites

Basic probability axioms and Probabilistic Independence definitions

Expected Value (of a function) of a Random Variable

Definition of conditional probability

Joint probability

Bayes rule

Marginal probability

Probability chain rule

Conditional Probability

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

Conditional Probabilities
are Probabilities

Conditional Probability

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

Why do we care?

It formulates some outcome within some specific context:

Language Modeling: “cat” in the context of “The fluffy ...”

Prediction: Label prediction/document classification

Conditional Probability

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

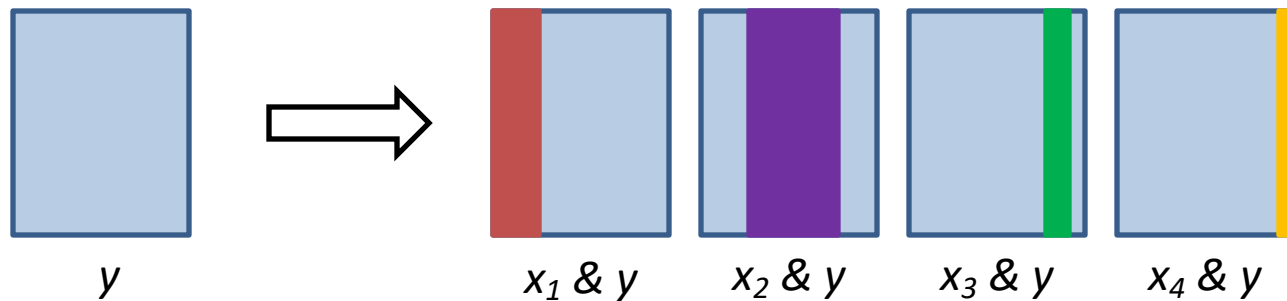
$p(Y)$ = marginal probability of Y

Conditional Probability

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

$$p(Y) = \sum_x p(X = x, Y)$$

Revisiting Marginal Probability: The Discrete Case



$$\begin{aligned} p(y) &= \sum_x p(x, y) \\ &= \sum_x p(x) p(y \mid x) \end{aligned}$$

Probability Prerequisites

Basic probability axioms and Probabilistic Independence definitions

Expected Value (of a function) of a Random Variable

Definition of conditional probability

Bayes rule

Joint probability

Probability chain rule

Marginal probability

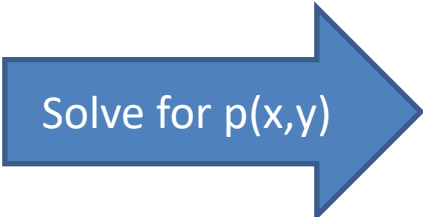
Deriving Bayes Rule

Start with conditional

$$p(Y | X)$$

Deriving Bayes Rule

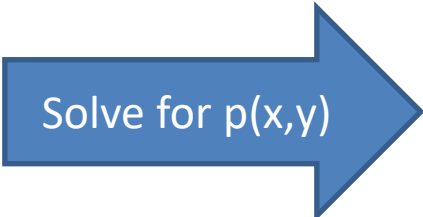
$$p(Y | X) = \frac{p(X, Y)}{p(X)}$$



Solve for $p(x,y)$

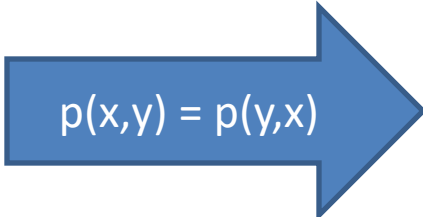
Deriving Bayes Rule

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



Solve for $p(x, y)$

$$p(X, Y) = p(X | Y)p(Y)$$



$p(x, y) = p(y, x)$

$$p(Y | X) = \frac{p(X | Y) * p(Y)}{p(X)}$$

Bayes Rule

$$\underset{\text{posterior probability}}{p(Y | X)} = \frac{\overset{\text{likelihood}}{p(X | Y)} * \overset{\text{prior probability}}{p(Y)}}{\underset{\substack{\text{marginal likelihood} \\ \text{(probability)}}}{p(X)}}$$

Bayes Rule

class/
label

observed
data

$$p(Y | X) = \frac{p(X | Y) * p(Y)}{p(X)}$$

class-based likelihood
(language model)

prior
probability of
class label

observation likelihood (averaged over all classes)

Why do we care?

It provides a rigorous approach for making predictions (Y) based off of observed data (X).

Probability Prerequisites

Basic probability axioms and Probabilistic Independence definitions

Expected Value (of a function) of a Random Variable

Definition of conditional probability

Bayes rule

Joint probability

Probability chain rule

Marginal probability

Probability Chain Rule

$$p(x_1, x_2) = p(x_1)p(x_2 | x_1)$$

Bayes rule

Probability Chain Rule

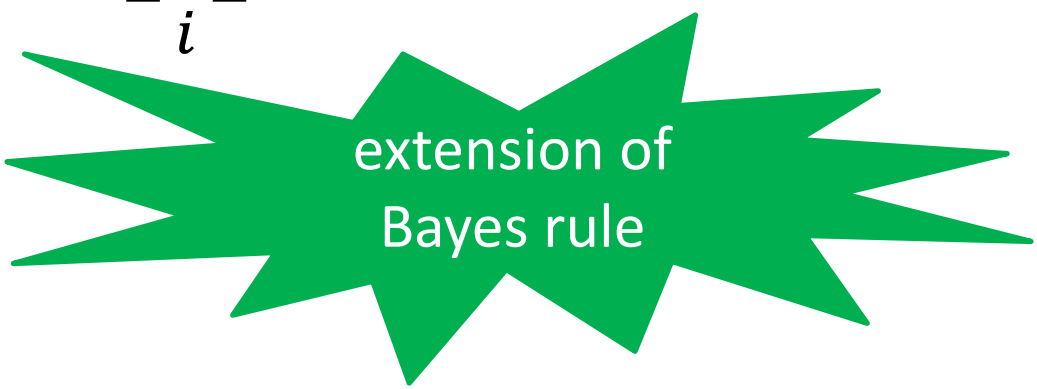
$$p(x_1, x_2, \dots, x_S) = \\ p(x_1)p(x_2 | x_1)p(x_3 | x_1, x_2) \cdots p(x_S | x_1, \dots, x_i)$$

Probability Chain Rule

$$\begin{aligned} p(x_1, x_2, \dots, x_S) &= \\ p(x_1)p(x_2 | x_1)p(x_3 | x_1, x_2) \cdots p(x_S | x_1, \dots, x_i) &= \\ \prod_i^S p(x_i | x_1, \dots, x_{i-1}) \end{aligned}$$

Probability Chain Rule

$$\begin{aligned} p(x_1, x_2, \dots, x_S) &= \\ p(x_1)p(x_2 | x_1)p(x_3 | x_1, x_2) \cdots p(x_S | x_1, \dots, x_{i-1}) &= \\ \prod_i^S p(x_i | x_1, \dots, x_{i-1}) \end{aligned}$$



extension of
Bayes rule

Probability Chain Rule

$$\begin{aligned} p(x_1, x_2, \dots, x_S) &= \\ p(x_1)p(x_2 | x_1)p(x_3 | x_1, x_2) \cdots p(x_S | x_1, \dots, x_{i-1}) &= \\ \prod_i^S p(x_i | x_1, \dots, x_{i-1}) \end{aligned}$$

Why do we care?

It gives an easier way of computing the probability of a sequence of words.