SPRING 2018

MAT 331

COURSE LINK

FINAL EXAM: TUESDAY MAY 15, 11:15AM-1:45PM

GENERAL INFORMATION ABOUT THE CLASS

WHEN & WHERE & WHO

MAT331

- Tuesday 11:30am - 12:50pm, Mathematics S235
  Location may change depending on if we need chalks…
  Theoretical aspects

- Thursday 11:30am - 12:50pm, Mathematics S235
  Programming aspects

- Instructor: Dr. Letao Zhang
  office: math 4-116
  office hour: Tuesday 4:20pm-5:20pm
  email: letao (dot) zhang (at) stonybrook.edu

- Grader: Mu Zhao
  email: mu (dot) zhao (at) stonybrook.edu

STRUCTURE OF THE CLASS

MAT 331

- Lecture notes, codes, homework problems will all be posted here https://sites.google.com/site/letaoedu/teaching/spr18mat331

- Grades will be posted on Blackboard

- On Tuesdays, we will mainly cover mathematical aspects of our class.

- On Thursdays
  I will show some codes first. And then I will give you some exercise to try in class. So that you can conquer the programming homework easily.

HOMEWORK

MAT 331

- Homework will be posted every Friday
  under "**Class Material and Homework**" on my website

- Each homework will mainly have two parts:

  - written part: you will hand in in class
    due the following Thursday in class

  - programming part: you will submit through Blackboard
    due at 5pm on Friday

- Make sure to check if your grade(s) are uploaded correctly on Blackboard. If not, please email me as soon as possible.

- No grade modification will be given after the last day of class.

GRADE

MAT 331

- Your grade of both written homework and programming homework will be posted on Blackboard.
- Your grade will be based on:
- (1) In-class quizzes.
  There will be 5 quizzes in total.
- (2) Homework (written part and programming part):
  Written part: Usually involving mathematical proofs and calculation by hand.

GRADE – CONTINUE

MAT 331

Programming part:
usually involving writing a short report stating the problem, describing how to solve it, giving the Python code with comments you used, and results.
Note that each programming assignment contributes to a portion of a project.

- (3) One midterm exam: in class
  (similar to quizzes, but longer)

- (4) One final exam
  (similar to quizzes, but longer)

GRADE

CALCULATION

- General Rule:
  - no late homework will be accepted
    But your lowest TWO homework grades will be dropped
  - no make-up quiz will be given
    But your lowest ONE quiz grade will be dropped
  - no make-up midterm or final will be given

GRADE
CALCULATION

- Your grade is comprised of:
  - 20% Homework.
  - 30% Quizzes
  - 20% Midterm.
    There will ONE in-class midterm exam
    The time will be determined later
  - 30% Final.
    Final Exam: Tuesday, May 15, 11:15am-1:45pm

PREREQUISITES

MAT 331

- Basic probability
  Probability covered by Chapter 1 through 7 of the following book should be sufficient:
  https://math.dartmouth.edu/~prob/prob/prob.pdf

  Although we will review some concepts whenever required, it is better that you are reviewing concepts rather than learning it for the first time. Things can get harder quickly.

- eg: random variables (continuous/discrete), density, distributions, expectation, variance, conditional probability, etc.

PREREQUISITES

MAT 331

You should be comfortable with matrix manipulations and calculus.
You should have a passing knowledge of multivariate calculus.

- Calculus
  Multivariate calculus or above should be sufficient.
  eg: taking derivatives, plot basic functions (sin, cos, exponential, etc.), find local minimum/maximum, gradient, Green's theorem, convex functions, etc.

- Linear Algebra
  Matrix operations, eigen-values/spaces, rank/basis of a vector space, linearly (in)dependency, orthogonality, QR-decomposition, etc.

PREREQUISITES

MAT 331

- Programming experience
  No previous programming experience in Python is required. But you should have some knowledge of writing code in some language.
  eg. Latex, Maple, Mathematica, Matlab, R, C, C++, Java, etc.

- We will use Python 3.6/Anaconda 3 — **this is required!**
  Some of the functions in python 2.X are not supported in 3.X, while some new features in 3.X cannot be used in 2.X.
  eg: dataframe.ix[] is not supported in 3.X

- **Next class, we will install everything together to avoid version issues.**

GOAL

Bayesian

- Understand mathematics behind naive Bayesian
  eg: Conjugate priors
  eg: Markov Chains, random walks

- Mathematical world is ideal, but reality is cruel…
  eg: Law of Large Numbers vs. How large is large?
  eg: Central Limit vs. Do I.I.D. random variables exist in reality?

- Plot and meaning of Plot

- Write some models, and then improve these models
  eg: linear regression vs. metropolis hastings
  eg: for loop vs. numpy.dot

RESOURCES

MAT 331

- Best place to find programming related answers
  https://stackoverflow.com/

- Basic Python3.x (required)
  Chapter 1 - 10 in the tutorial:
  https://www.py4e.com/html3/01-intro

- Probability (prerequisite Chapter 1 - 7; required chapter 8-12)
  https://math.dartmouth.edu/~prob/prob/prob.pdf

RESOURCES

MAT 331

- Highly Recommended :
  we will NOT use R/Stan, but the stats part is useful

  Statistical Rethinking: A Bayesian Course with Examples in R and Stan by Richard McElreath.

  Bayesian Data Analysis, by Andrew Gelman, John Carlin, Hal Stern, David Dunson, Aki Vehtari, and Donald Rubin. 3rd Edition.

RESOURCES – FOR YOUR OWN INTERESTS

MAT 331

- If time allows, we will cover some Bayesian Networks. The following summary is a great resource.
  M. Jordan, An Introduction to Graphical Models
  (link: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.116.7467&rep=rep1&type=pdf)

- To learn Python, the best way is to practice and use Google search…
  (just my experience)

INTRODUCTION

MODEL REPRESENTATION

- Intelligent software :
- Goal: software that can adapt, learn, and reason

Player skill → **Game result**

Movie preferences → **Ratings**

Words → **Ink**

Can be described by a *model*

MODEL REPRESENTATION

- Intelligent software :
- Goal: software that can adapt, learn, and reason

Player skill → **Game result**

Movie preferences → **Ratings**

Words → **Ink**

Reasoning backwards

HANDLING UNCERTAINTY

- We are uncertain about a player's skill. Each result provides relevant information. But we are never completely certain
- How can we compute with uncertainty in a principled way?

UNCERTAINTY EVERYWHERE

- Which movie should the user watch next?
- Which did the user really write?
  ZOO or 200?
- If the user clicked a link, is it by accident? or on purpose?
- What kind of product does the user wish to buy?
  Beers next to diapers…
- Which stock to buy?
  (95% sure GOOG goes up by 1%; 90% sure AMZN goes up by 2%)
  Which one to buy?

REDUCE UNCERTAINTY

- Inference
  ⍰ How do I answers questions/queries according to my model and/or based given data?

- Conditional probability. P(A|D)

- eg:
  Given day1's average temperature 10F, what day2 temperature will be? (10F +/- 10F probably)
  But if you have no idea about day1's temperature, day2's temperature could 0F to 100F….

LEARNING

- To find the best model for my data

- eg:
  Maximal likelihood
  Random Forest
  Linear regression
  etc...

- measure how well my model works

RECAP OF BASIC PROBABILITY

WHAT IS PROBABILITY

MAT 331

- Everyone starts with a coin…

WHAT IS PROBABILITY

PHYSICAL MODEL

- Suppose you were to flip a coin.
  - You expect not to be able to say next toss — tail or head?
  - But you are able to tell a friend that the odds of getting a head is equal to to the odds of getting a tail, and that both are 1/2
- This intuitive notion of odds is a probability.

“

WHERE DOES THIS 1/2 COME FROM?

”

WHAT IS PROBABILITY

PHYSICAL MODEL

- Our physical model of the world: U.S. Mint

- Physical model: Because of our faith in the U.S. Mint, we might be willing to, **without having seen any tosses**, say that the coin is fair.
  i.e. Both of Head(H) and Tail(T) are equally likely.

WHAT IS PROBABILITY

SYMMETRY

- Everyone then goes with an example of dice

WHAT IS PROBABILITY

SYMMETRY

- If we were tossing a ‘fair’ six-sided die, we may thus equivalently say that the odds of the die falling on any one of its sides is 1/6

- Tossing a ‘N’ sided die
  The odds of the die falling on any one of its sides is 1/N

- This notion of probability springing from symmetry assuming **Fairness**

BUT THE WORLD IS NOT PERFECT

MAT 331

- The following is an unfair coin
- It has a 60% chance of coming up heads
- It has a 1-60% chance of coming up tails

60%

40%

“

WHERE DO YOU THINK THIS 60% COMES FROM?

”

WHAT IS PROBABILITY

EXPERIMENT BY FREQUENCY

- We see a bended coin
- We assume it is not a fair coin
- We may toss it, say 10,000,000 times, and 6,000,000 come up with heads
- We conclude that there is 60% chance coming up heads
- The above conclusion is from frequencies

PROBABILITY

Always start with ASSUMPTION or it is somehow always "Conditional"

- from symmetry
  fair dice (assuming fairness)

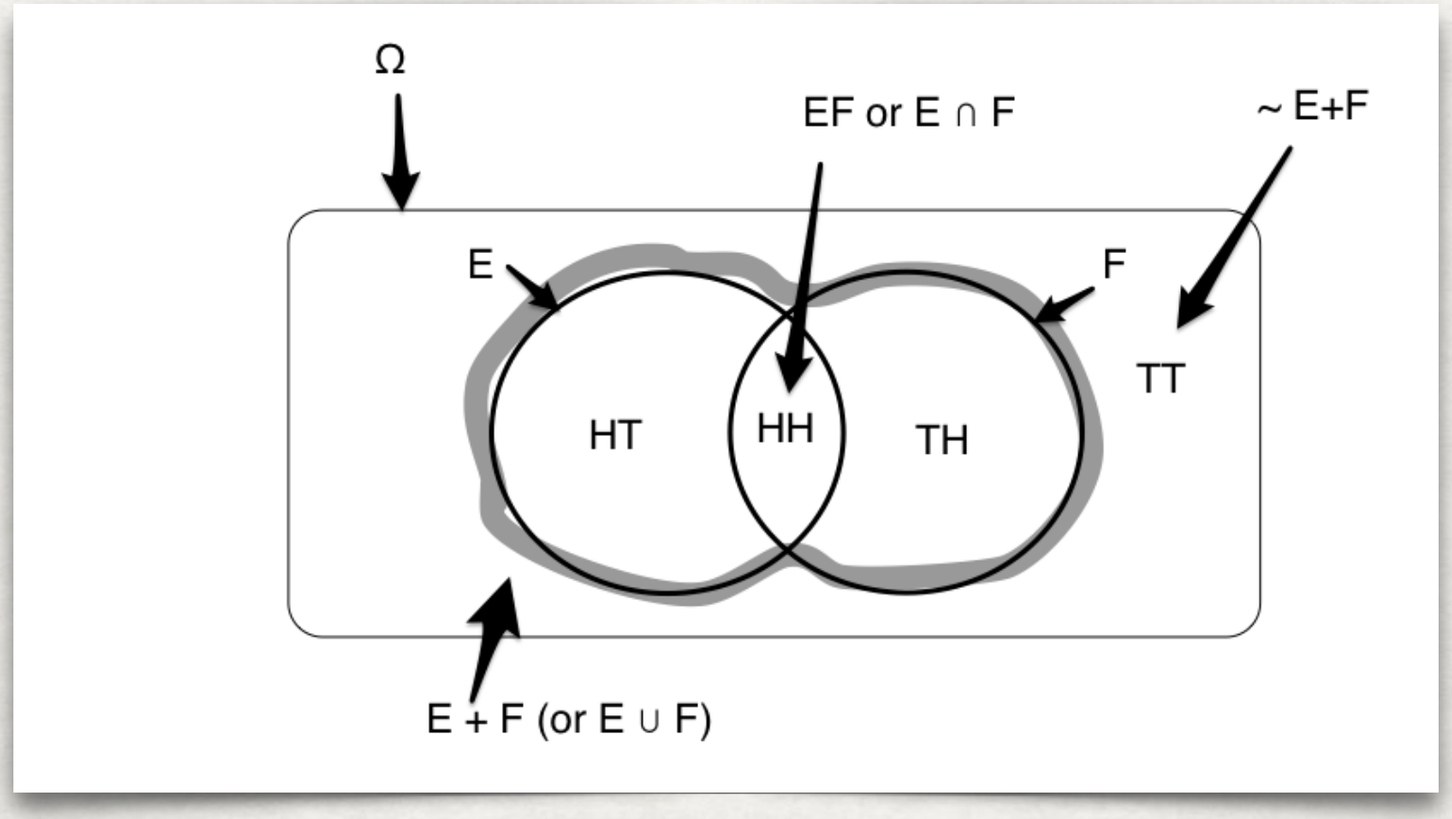
- from a model, and combining beliefs and data:
  See an example later — bayesian approach

- from long run frequency
  10,000,000 tosses
  Assuming independent tosses

MATHEMATICAL EXPRESSION OF PROBABILITY

RANDOM VARIABLE

- E is the event of getting a head in a first coin toss
- F is the same for a second coin toss
- Ω is the set of all possibilities that can happen when you toss two coins: {HH,HT,TH,TT}



RANDOM VARIABLE

Definition. A random variable is a mapping

$$X : \Omega \to \mathbb{R}$$

that assigns a real number $X(\omega)$ to each outcome ω.
- Ω is the sample space. Points
- ω in Ω are called sample outcomes, realizations, or elements.

- Note X only assigns ONE and only ONE real number to each element in the sample space Ω.
- The set of all possible values of the random variable X is called the support, or space, of X.

DISCRETE RANDOM VARIABLE

- Definition.
  A random variable X is a discrete random variable if:

  1. there are a finite number of possible outcomes of X, or
  2. there are a countably infinite number of possible outcomes of X

- eg:
  finite: Coin tosses one time
  countable + infinite: coin tosses infinite number of times

DISCRETE RANDOM VARIABLE

FAIR COIN

- $\Omega = \{H,T\}$
- say: $X(H) = 0$, $X(T) = 1$
- Note you can also define:
  say: $X(H) = 100$, $X(T) = 200$
- All you need to do is to assign H and T (all out comes for this case) different real numbers

- Say $\omega = HHTTTTHTT$ then $X(\omega) = 3$ if defined as number of heads in the sequence ω.

- We will assign a real number P(A) to every event A, called the probability of A.
- We also call P a probability distribution or a probability measure.

DIE EXAMPLE

- A die is rolled once. We let X denote the outcome of this experiment. Then the sample space for this experiment is the 6-element set

- $\Omega = \{1,2,3,4,5,6\}$
  where each outcome i, for i = 1, . . . , 6, corresponds to the number of dots on the face which turns up.

- Since X denotes the outcome, the map is naturally X(i) = i

- Event: E = {2,4,6}
  The event E can also be described by saying that X is even.

- $P(E) = P(X=2 \cup X=4 \cup X=6) = 1/2$

PROBABILITY MASS FUNCTIONS

DISCRETE RANDOM VARIABLE

- The probability that a discrete random variable X takes on a particular value x,
  that is, P(X = x), is frequently denoted f(x).
- The function f(x) is typically called the probability mass function,
- Note: some authors also refer to it as
  - probability function,
  - the frequency function,
  - probability density function (pdf)
- **We will use pdf or pmf for probability density function or prob. mass. func.**

PROBABILITY MASS FUNCTIONS

DEFINITION

- Definition. The probability mass function, P(X = x) = f(x), of a discrete random variable X is a function that satisfies the following properties:

$$(1)\ P(X = x) = f(x) > 0 \quad \text{if } x \in \text{the support } S$$

$$(2)\ \sum_{x \in S} f(x) = 1$$

$$(3)\ P(X \in A) = \sum_{x \in A} f(x)$$

- Or any function defined over a countable subset of real numbers is a probability mass function for some random variable if it satisfies the above (1) - (3)

PROBABILITY MASS FUNCTIONS

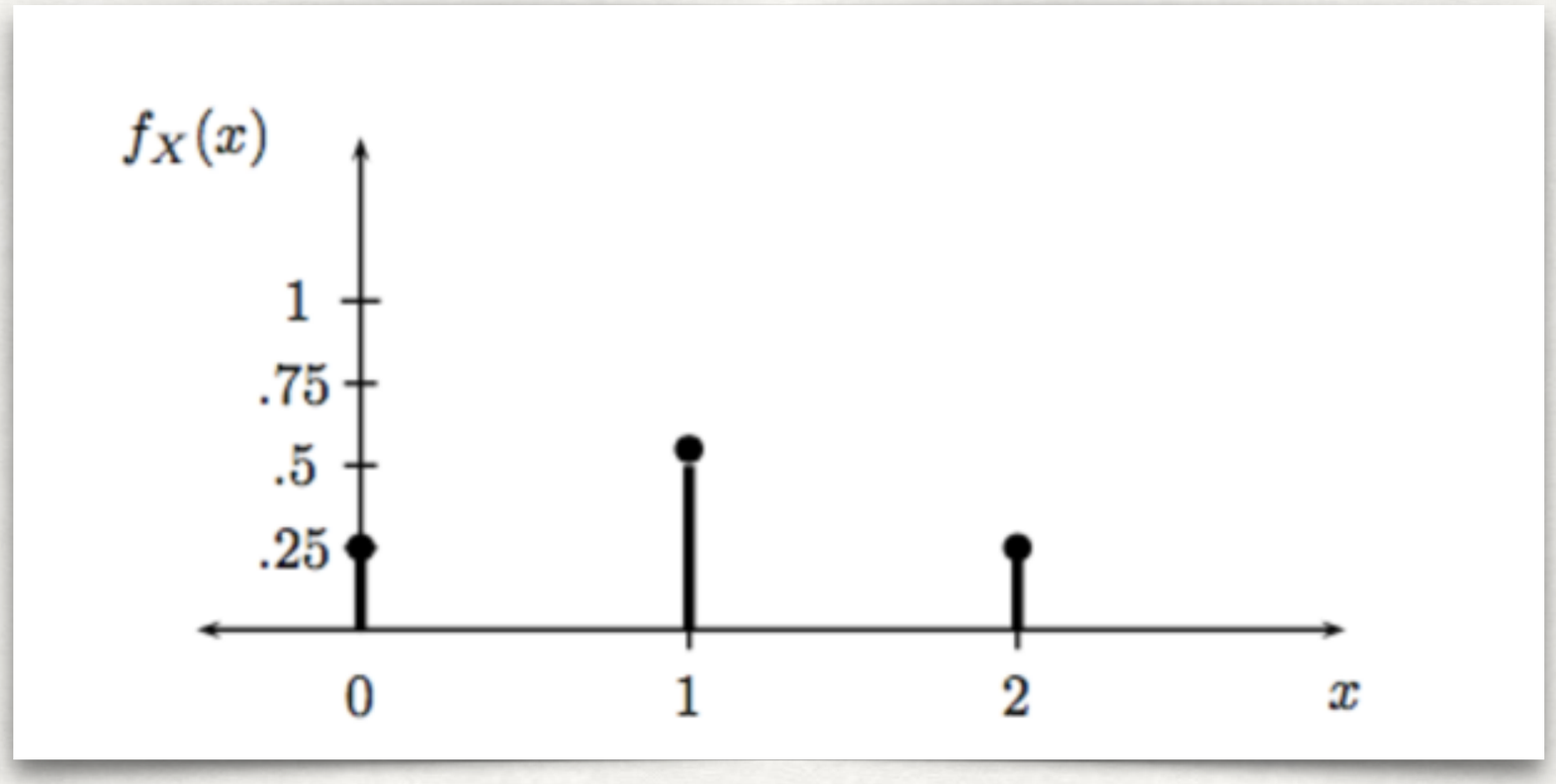
ONE COIN TOSS

- $\Omega = \{H,T\}$
- say: $X(H) = 0$, $X(T) = 1$
- Probability mass function here is a function P, where:
  P(head) is really $P(X = 0) = 1/2$
  P(tail) is really $P(X = 1) = 1/2$
  $P(X = t) = 0$ for all real numbers t except t = 0 and 1
- eg: $P(X = 3.13) = ?$

PROBABILITY MASS FUNCTIONS

TWO COIN TOSSES

- Example:
  The pmf for the number of heads in two coin tosses



INDEPENDENCE

X and Y are independent if:

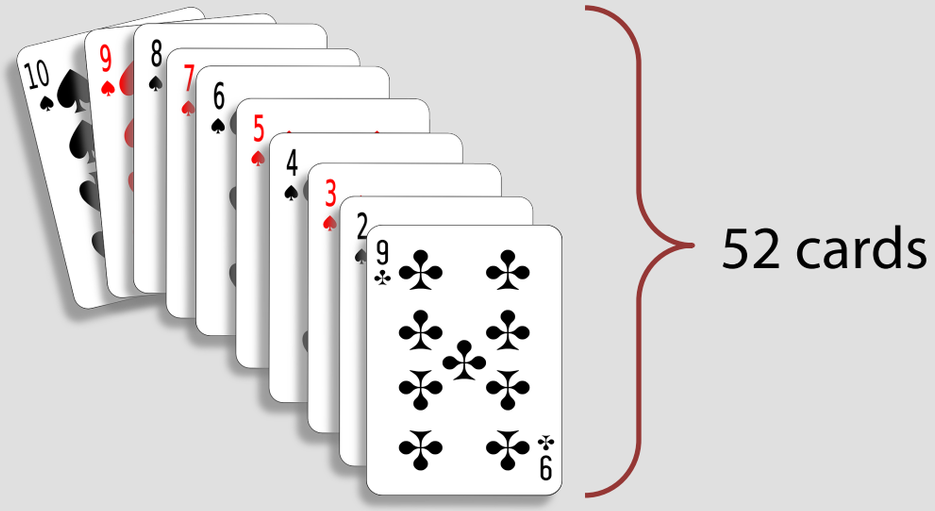
Marginals

$$P(X, Y) = P(X)P(Y)$$

Joint

INDEPENDENCE

DRAWS FROM THE DECK



$$P(X_1 = 9\clubsuit, X_2 = 9\clubsuit) = 0$$

$$P(X_1 = 9\clubsuit)P(X_2 = 9\clubsuit) = \frac{1}{52^2}$$

CONDITIONAL PROBABILITY

Probability of **X** given that **Y** happened:

Joint

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

Conditional

Marginal

FUNDAMENTAL RULES OF PROBABILITY

- p(X) >= 0
  probability must be non-negative

- 0<=p(X) <= 1

- p(X) + p(not X) = 1
  either happen or not happen

- p(X+Y) = p(X) + p(Y) - p(XY)

CHAIN RULE

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

$$P(X, Y, Z) = P(X|Y, Z)P(Y|Z)P(Z)$$

$$P(X_1,\ldots,X_N) = ?$$

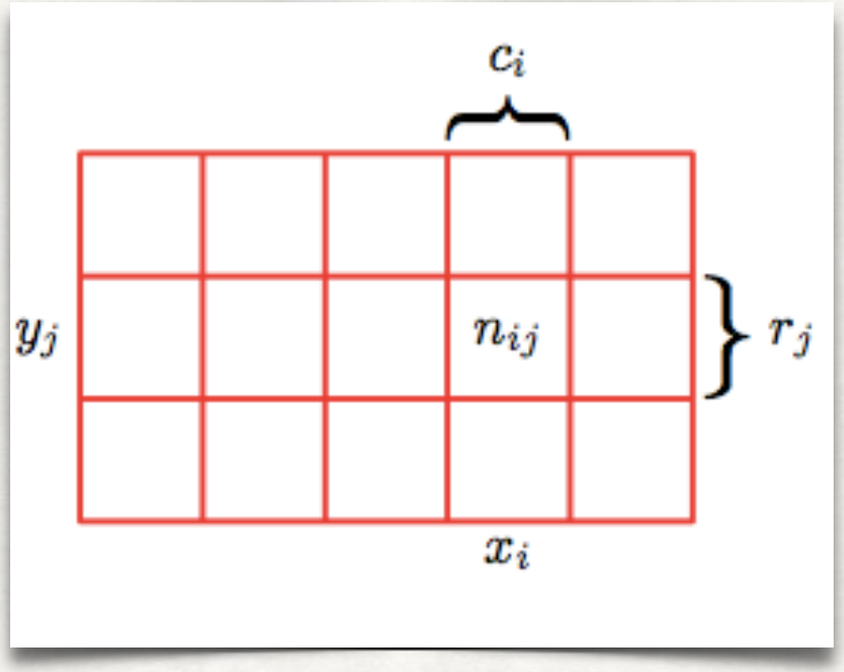
CHAIN RULE

$$P(X_1, \ldots, X_N) = \prod_{i=1}^{N} P(X_i | X_1, \ldots, X_{i-1})$$

- If X_i (i = i … N) are binary, and you know nothing about relations among them. The number of values required to find the joint distribution?

- What if X_i (i = i … N) are binary and independent?
  $P(X_1,\ldots,X_N) = P(X_1)\ldots P(X_N)$
  Each $P(X_i)$ contributes two values, in total we have 2N values to determine the joint distribution table.

MARTINGALE AND CONDITIONAL PROBABILITY



- Marginalize for X sum of all rows (integral against y)

$$p(X = x_i) = \sum_j p(X = x_i, Y = y_j)$$

$$p(Y = y_j \mid X = x_i) \times p(X = x_i) = p(X = x_i, Y = y_j).$$

BAYES THEOREM

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

BAYES THEOREM

STATISTICALLY: THINK BAYESIAN

θ — parameters

X — observations

Posterior Likelihood Prior

$$P(\theta|X) = \frac{P(X,\theta)}{P(X)} = \frac{P(X|\theta)P(\theta)}{P(X)}$$

Evidence

BAYES THEOREM

READ MEDICAL RESULTS

row as condition	Has disease A	Does not have A
Tested A (positive)	True Positive	False Positive
Tested Not A (negative)	False Negative	True Negative

- P(test result | actual have disease)

A murder mystery

A fiendish murder has been committed

Whodunit?

There are two suspects:

– the **Butler**

– the **Cook**

There are three possible murder weapons:

– a butcher's **Knife**

– a **Pistol**

– a fireplace **Poker**

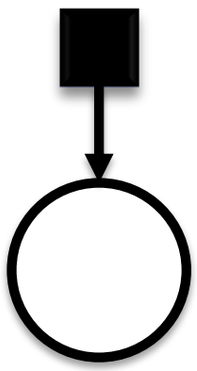
Prior distribution

Butler has served family well for many years
Cook hired recently, rumours of dodgy history

P(Culprit = **Butler**) = 20%

P(Culprit = **Cook**) = 80%

Probabilities add to 100%



P(Culprit)

This is called a *factor graph*
(we'll see why later)

Culprit = {**Butler**, **Cook**}

Conditional distribution

Butler is ex-army, keeps a gun in a locked drawer

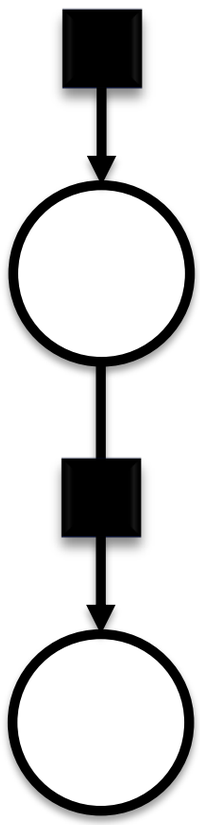
Cook has access to lots of knives

Butler is older and getting frail

	Pistol	Knife	Poker	
Cook	5%	65%	30%	**= 100%**
Butler	80%	10%	10%	**= 100%**

P(Weapon | Culprit)

Factor graph



Prior distribution

P(Culprit)

Culprit = {**Butler**, **Cook**}

Conditional distribution

P(Weapon | Culprit)

Weapon = {**Pistol**, **Knife, Poker**}

Generative viewpoint

Murderer	Weapon
Cook	Knife
Butler	Knife
Cook	Pistol
Cook	Poker
Cook	Knife
Butler	Pistol
Cook	Poker
Cook	Knife
Butler	Pistol
Cook	Knife
…	…

Joint distribution

What is the probability that the **Cook** committed the murder using the **Pistol**?

P(Culprit = **Cook**) = 80%

P(Weapon = **Pistol** | Culprit = **Cook**) = 5%

P(Weapon = **Pistol** , Culprit = **Cook**) = 80% x 5% = 4%

Likewise for the other five combinations of Culprit and Weapon

Joint distribution

	Pistol	Knife	Poker	
Cook	4%	52%	24%	**= 100%**
Butler	16%	2%	2%	

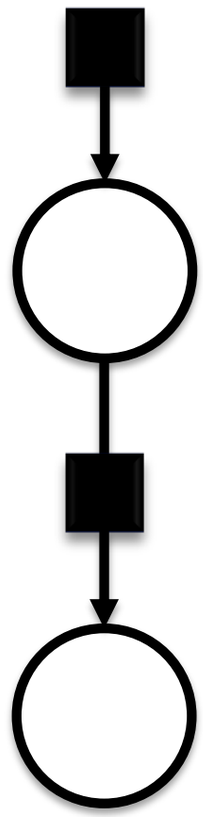
P(Weapon, Culprit) = P(Weapon | Culprit) P(Culprit)

$$P(x, y) = P(y|x)P(x)$$

Product rule

Factor graphs

Generative model



P(Culprit)

Culprit = {**Butler**, **Cook**}

P(Weapon | Culprit)

Weapon = {**Pistol**, **Knife, Poker**}

P(Weapon, Culprit) = P(Weapon | Culprit) P(Culprit)

Generative viewpoint

Murderer	Weapon
Cook	Knife
Butler	Knife
Cook	Pistol
Cook	Poker
Cook	Knife
Butler	Pistol
Cook	Poker
Cook	Knife
Butler	Pistol
Cook	Knife
…	…

Marginal distribution of Culprit

	Pistol	Knife	Poker	
Cook	4%	52%	24%	**= 80%**
Butler	16%	2%	2%	**= 20%**

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

Sum rule

Marginal distribution of Weapon

	Pistol	Knife	Poker
Cook	4%	52%	24%
Butler	16%	2%	2%
	= 20%	**= 54%**	**= 26%**

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

Sum rule

Posterior distribution

We discover a **Pistol** at the scene of the crime

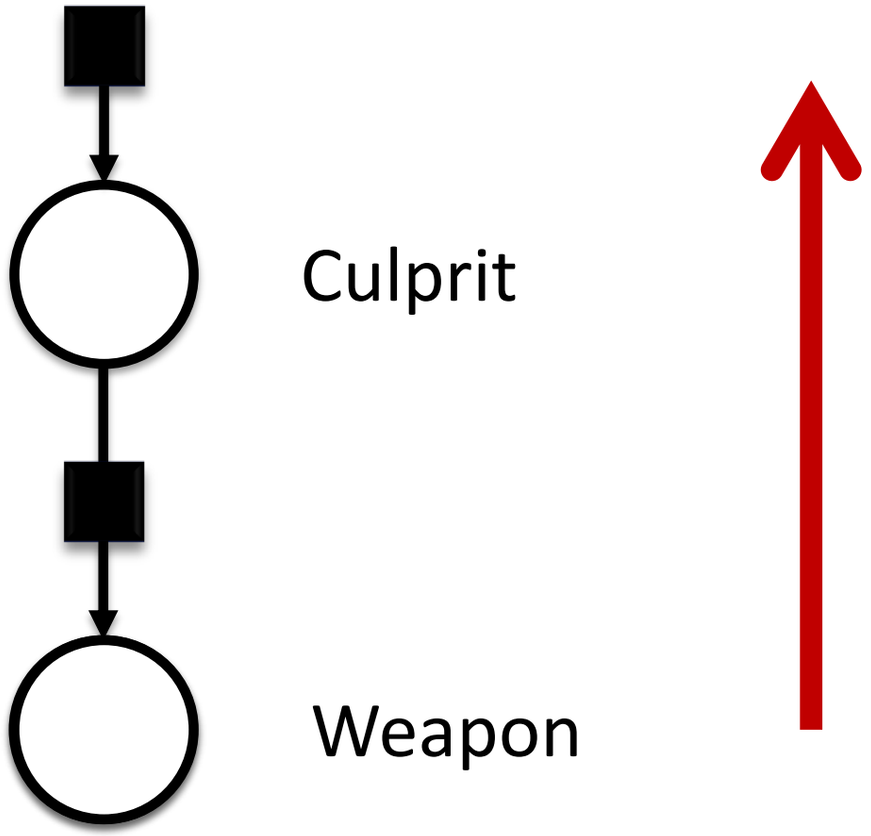
	Pistol	**Knife**	**Poker**	
Cook	4%	~~52%~~	~~24%~~	= 20%
Butler	16%	~~2%~~	~~2%~~	= 80%

This looks bad for the Butler!

Generative viewpoint

Murderer	Weapon
Cook	Knife
Butler	Knife
Cook	Pistol
Cook	Poker
Cook	Knife
Butler	Pistol
Cook	Poker
Cook	Knife
Butler	Pistol
Cook	Knife
…	…

Reasoning backwards



Bayes' theorem

$$P(x, y) = P(y|x)P(x)$$

likelihood **prior**

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

posterior

Prior – belief before making a particular obs.

Posterior – belief after making the obs.

Posterior is the prior for the next observation

– Intrinsically incremental

The Rules of Probability

Sum rule

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

Product rule

$$P(x, y) = P(y|x)P(x)$$

Bayes' theorem

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

Denominator

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