Statistical Computing for Scientists and Engineers

Prof. Nicholas Zabaras
Center for Informatics and Computational Science
https://cics.nd.edu/
University of Notre Dame
Notre Dame, Indiana, USA

Email: nzabaras@gmail.com
URL: https://www.zabaras.com/

August 21, 2018

Course Organization

https://www.zabaras.com/statisticalcomputing

- Two lectures each week TThu 12:30-1:45
- Recitation (not mandatory) F 11:30-12:20.
- Teaching Assistants: Nicholas Geneva, Govinda Anantha-Padmanabha, Navid Shervani-Tabar
- Office Hours: (NZ, Cushing 311I) M 1-2 pm, F 1-2 pm; (Tas, dDeBartolo, 125) M 5-7 pm
- Occasionally lectures on Friday.
- All information regarding the course will be posted on the webpage.
- This includes video lectures, slides, references, homework, etc.
- Grades based on Homeworks (60%), Final Project (40%)

Books

- C Bishop, Pattern Recognition and Machine Learning
- K Murphy, Machine Learning: A Probabilistic Perspective
- JS Liu, Monte Carlo Strategies in Scientific Computing.
- CP Robert, Monte Carlo Statistical Methods.
- A Gelman, JB Carlin, HS Stern and DB Rubin, Bayesian Data Analysis.
- CP Robert, The Bayesian Choice: from Decision-Theoretic Motivations to Computational Implementation.
- ET Jaynes, Probability Theory: The Logic of Science.

Additional references to journal publications will be provided (with html links) in each lecture.

Course Objectives

- To introduce statistical methods used for the stochastic simulation of complex physical systems in the context of Bayesian analysis.

- Introduce Monte Carlo and approximate inference techniques in the context of Bayesian models, and parametric and non-parametric models for supervised and unsupervised learning.

- Acquaint students with a set of powerful tools and theories that can be directly transitioned to their research independently of their field.

- The course is appropriate for graduate students in Engineering, Chemical/Physical and Biological Sciences, Mathematics/Statistics/ Computer Science.

Motivation

- Why Monte Carlo and Approximate Inference

  - To deal with *uncertainties* that are omnipresent in physical systems
  - To perform computational tasks (e.g. integration) in *high dimensional spaces*.

- Why Bayesian?

  - To draw inferences from data. This data is collected experimentally or produced computationally.
  - To quantify uncertainties associated with these inferences.
  - To quantify *predictive uncertainties*.

Bayesian Statistics

- Why is it relevant to Science & Engr?
  - It is highly suitable to many problems from diverse areas: From physics and chemistry to genetics, from econometrics to machine learning.
  - Allows *principled* incorporation of prior knowledge/ information/beliefs.
  - This is a vice and a virtue.
  - Readily handle missing or corrupted data, outliers.
- Why now?
  - Bayesian Statistics have enjoyed a surge of popularity in the last 25 years.
  - This has coincided with advances in Scientific Computation.
  - Bayesian models, powerful as they may be, are analytically intractable.
  - Most often one must resort to Approximate Inference or Monte Carlo.

Topics to Cover

1. Basics of Probability, Statistics, and Information Theory.
2. Foundamentals of Bayesian Statistics: Prior, Likelihood, Posterior, Predictive Distributions
3. Bayesian Inference: Applications to Regresssion, Classification, Model Reduction, etc.
4. Monte Carlo Methods: Applications to Dynamical Systems, Time Series Models, HMM, Probabilistic Robotics, etc.
5. State Space Models
6. Sparse Bayesian Models
7. Bayesian Model Selection
8. Expectation-Maximization, Mixture Models
9. Variational Methods, Approximate Inference (an introduction)
10. Other (if time allows)

Introduction to Probability and Statistics

Prof. Nicholas Zabaras
Center for Informatics and Computational Science
https://cics.nd.edu/
University of Notre Dame
Notre Dame, Indiana, USA

Email: nzabaras@gmail.com
URL: https://www.zabaras.com/

August 21, 2018

Contents

References

- Following closely Chris Bishops' PRML book, Chapter 2.

- Kevin Murphy's, Machine Learning: A probablistic perspective, Chapter 2.

- Jaynes, E. T. (2003). Probability Theory: The Logic of Science. Cambridge University Press.

- Bertsekas, D. and J. Tsitsiklis (2008). Introduction to Probability. Athena Scientific. 2nd Edition.

- Wasserman, L. (2004). All of statistics. A Concise Course in Statistical Inference. Springer.

Frequentist Vs Bayesian

- Frequentisit Probability: Long run frequencies of `events'
- Bayesian Probability: Quantifying our uncertainty about something
  - It can be used to model our uncertainty about events that do not have long term frequencies
    - ✓ E.g. the event that the polar ice cap will melt by 2020
- The rules of probability are the same for both approaches.

For more details: S. Ross, Introduction to Probability Models

Sample Space

Sample space: Set of all possible outcomes of an experiment.

- $\Omega = \{1,2,3,4,5,6\}$ is the sample space for the numbers that appear on a die rolled once.

Event: A subset of the sample space.

- If $E = \{5\}$ then E is the event of rolling a 5.

- If $E = \{J,Q,K\}$ the E is the event of getting a face card.

For more details: S. Ross, Introduction to Probability Models

Union and Intersection

$\cup$ union operator: For any two events E and F of a sample space Ω, we define the new event $E \cup F$ to consist of all outcomes that are either in E or in F or in both E and F.

- If $E = \{1, 5\}$ and $F = \{3\}$ then $E \cup F = \{1, 3, 5\}$ is the event of rolling an odd number.

$\cap$ intersection operator: For any two events E and F of a sample space Ω, we define the new event $E \cap F$ to consist of all outcomes that are in both E and F.

- If $E = \{1, 3, 5\}$ and $F = \{3\}$ then $E \cap F = \{3\}$ is the event of rolling a three.
- If $E = \{J, Q, K\}$ and $F = \{10, K\}$ then $E \cap F = \{K\}$ is the event of getting a King.
- If $E = \{H\}$ and $F = \{T\}$ then $E \cap F = \varnothing$ would not consist of any outcomes and would thus not occur. If $E \cap F = \varnothing$, then E and F are said to be mutually exclusive ($\varnothing$ is called the empty set).

More than two Events

Likewise, $\bigcup_{i=1}^{\infty} E_i$ describes the union of events $E_1, E_2, \ldots$ and corresponds to outcomes that are in E_i for at least one value of $i = 1, 2, \ldots$.

$\bigcap_{i=1}^{\infty} E_i$ describes the intersection of events $E_1, E_2, \ldots$ and corresponds to outcomes that are in all events E_i, $i = 1, 2, \ldots$.

Laws of Probability

$\Pr(E)$*:* The probability of event E. It is a number satisfying the following three conditions:

1) $0 \leq \Pr(E) \leq 1.$

2) $\Pr(\Omega) = 1$, and $\Pr(\varnothing) = 0.$

3) For any sequence of events $E_1, E_2, \ldots$ that are mutually exclusive (i.e., $E_i \cap E_j = \varnothing$) when *i≠j*, the following holds:

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \Pr(E_i)$$

Properties of Probability

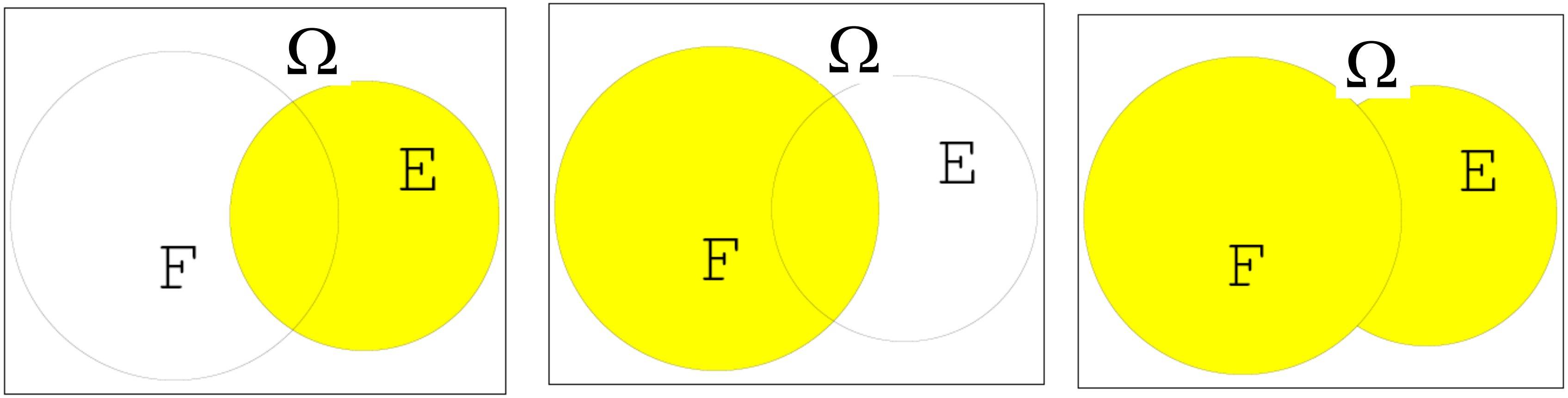
E^c: The complement of E (i.e., all of the outcomes that are not in event E).

$$E \cup E^c = \Omega$$

Example: If E is the event ``rolling two dice three times and getting three sevens'', then $E^c$ is ``the set of outcomes where someone when rolling the dice three times would get anything but three sevens''.

- $\Pr(E)+\Pr(F)=\Pr(E\cup F)+\Pr(E\cap F)$ $\quad \Pr(E\cup F)=$

$\Pr(E)$ $\quad \Pr(F)$ $\quad \Pr(E)+\Pr(F)-\Pr(E\cap F)$



Discrete Random Variables

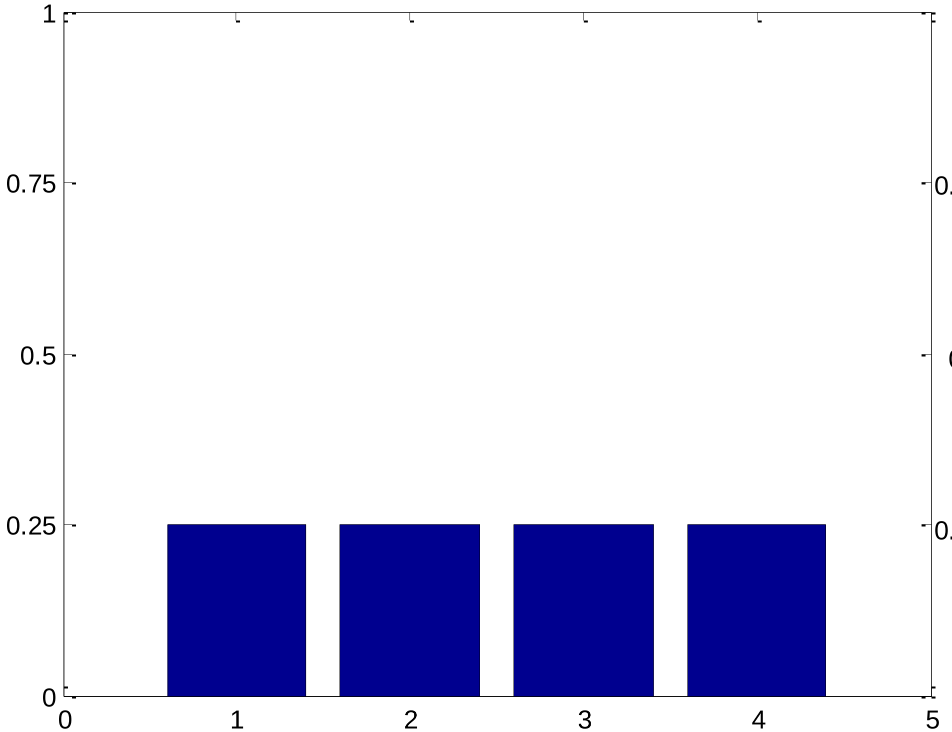
- Discrete random variable X, which can take on any value from a finite or countably infinite set $\mathcal{X}$.

- Denote the probability of the event that $X = x$ by $P(X = x)$, or just $P(x)$ for short. Here $P(\cdot)$ is called a probability mass function or pmf. It satisfies the properties

$$0 \leq P(x) \leq 1, \sum_{x \in \mathcal{X}} P(x) = 1$$

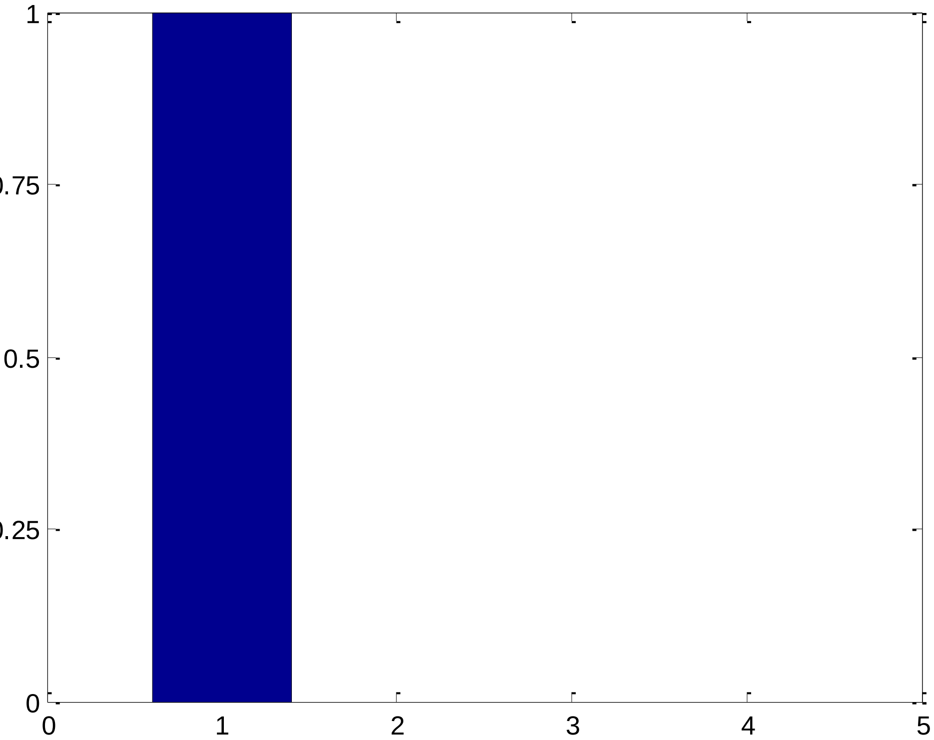
- Let us plot next the pmf for $\mathcal{X} = \{1,2,3,4\}$ for (a) a uniform random variable $P(x = k) = 1/4$, and for a degenerate distribution defined as $P(x) = 1$ if $x = 1$, otherwise zero. This last distribution can be written in terms of the indicator function as: $P(x) = \mathbb{I}(x = 1)$

Discrete Random Variables: Example



$$\mathcal{X} = \{1,2,3,4\}$$

$$P(x) = 1/4$$



$$\mathcal{X} = \{1,2,3,4\}$$

$$P(x) = \mathbb{I}(x=1)$$

- Run MatLab function *discreteProbDistFig* from Kevin Murphys' PMTK

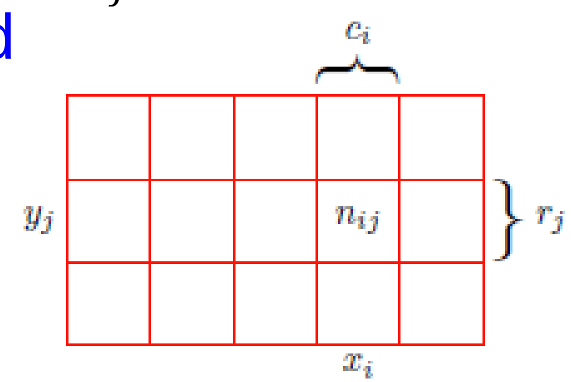
Joint Probability

- Probability theory provides a consistent framework for the quantification and manipulation of uncertainty.

- It forms one of the central foundations for pattern recognition and machine learning.

- The probability that X will take the value x_i and Y will take the value y_j is written $P(X = x_i, Y = y_j)$ and is called the joint probability of $X = x_i$ and $Y = y_j$.

$$P(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$



- Here n_{ij} is the number of times (in N trials) that the event $X = x_i, Y = y_j$ occurs. Similarly c_i is the number of times that $X = x_i$ occurs.

The Sum and Product Rules

- Even complex calculations in probability are simply derived from the sum and product rules of probability.

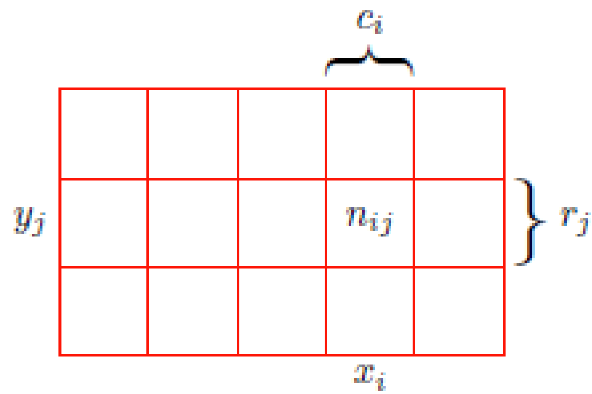
- Sum Rule:

$$P(X = x_i) = \sum_{j=1}^{L} P(X = x_i, Y = y_j)$$

- Product Rule:

$$P(X = x_i, Y = y_j) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i}\frac{c_i}{N}$$

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



- The product rule leads to the Chain Rule:

$$P(X_{1:D}) = P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_1, X_2)...P(X_D \mid X_{1:D-1})$$

The Sum and Product Rules

- Even complex calculations in probability are simply derived from the sum and product rules of probability.

- Sum Rule:

$$P(x) = \int P(x, y)dy$$

- Product Rule:

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

Conditional Probability and Bayes' Rule

$$P(X=x|Y=y)=\frac{P(X=x,Y=y)}{P(Y=y)}=\frac{P(X=x)P(Y=y|X=x)}{\sum_{x'}P(X=x')P(Y=y|X=x')}$$

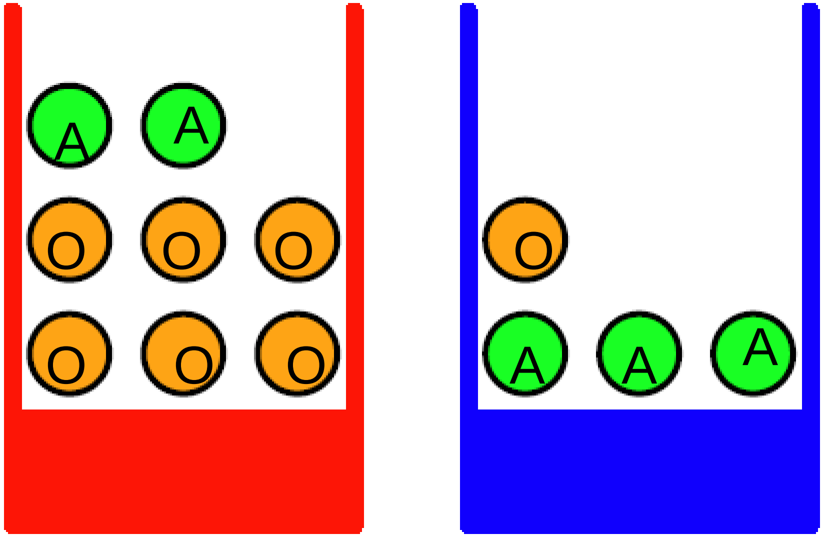
- *Bayes'* theorem plays a central role in pattern recognition and machine learning

- The normalizing factor $P(Y)$ is given as:

$$P(Y=y)=\sum_{X}P(X,Y=y)=\sum_{x'}P(X=x')P(Y=y|X=x')$$

Example of Bayes' Theorem

- Suppose a red and a blue box with probabilities of selecting each of them being $P(B=r)=\frac{4}{10}<\frac{1}{2}$, $P(B=b)=6/10$.



- We select an orange. What is the probability that we chose from the red box $P(B=r\,|\,F=o)$?

$$P(F=o)=P(F=o\,|\,B=r)P(B=r)+P(F=o\,|\,B=b)P(B=b)$$

$$=\frac{6}{8}\frac{4}{10}+\frac{1}{4}\frac{6}{10}=\frac{9}{20}$$

- Then: $$P(B=r\,|\,F=o)=\frac{P(F=o\,|\,B=r)P(B=r)}{P(F=o)}=\frac{\frac{6}{8}\frac{4}{10}}{\frac{9}{20}}=\frac{2}{3}>\frac{1}{2}$$

Example: Medical Diagnosis

- Coming back from a trip, you feel sick and your doctor thinks you might have contracted tuberculosis (TB) (0.4% of the population has the disease): P(TB)= 0.004.
- A test is available but not perfect.
  - If a tested patient has the disease, 80% of the time the test will be positive: P(Positive|TB)= 0.80
  - If a tested patient does not have the disease, 90% of the the time the test will be negative (10% false positive):

$$P(Positive \mid \overline{TB}) = 0.1$$

- Your test is positive, should you really care? What is P(TB|Positive)?
- Base Rate Fallacy: People will assume that there are 80% likely to have the disease – that's wrong as it does not account for the prior probability.

Example: Medical Diagnosis

- We use Bayes' rule as follows:

$$P(TB \mid Positive) = \frac{P(Positive \mid TB)P(TB)}{P(Positive \mid TB)P(TB) + P(Positive \mid \overline{TB})P(\overline{TB})}$$

$$= \frac{0.8 \times 0.004}{0.8 \times 0.004 + 0.1 \times 0.996} \approx 0.031$$

- If you test positive, you only have a 3.1% chance to have the disease.

- Such a test would be a complete waste of money.

Example: Generative Classifier

$$P(y=c\,|\,\boldsymbol{x},\theta)=\frac{P(y=c\,|\,\theta)P(\boldsymbol{x}\,|\,y=c,\theta)}{\sum_{x'}P(y=c'\,|\,\theta)P(\boldsymbol{x}\,|\,y=c',\theta)}$$

- Here we classify feature vectors $\boldsymbol{x}$ to classes using the above posterior.

- It is a generative classifier as it specifies how to generate data $\boldsymbol{x}$ using *the class-conditional probabilities* $P(\boldsymbol{x}\,|\,y=c,\theta)$ *and class priors* $P(y=c\,|\,\theta)$.

- *In a discriminative setting,* the posterior $P(y=c\,|\,\boldsymbol{x})$ is directly fitted.

Independency, Conditional Probability

- Two events A *and* B are independent (written as $A \perp B$) if

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

- Conditional probability: Probability that A happens provided that B happens,

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \quad \left(Note: P(A \mid B) \geq P(A \cap B)\right)$$

- Using the above Eqs, we see that for independent events,

$$P(A \mid B) = P(A)$$

Independence, Conditional Independence

- X and Y are *unconditionally independent or marginally independent*, denoted $X \perp Y$, if we can represent the joint as the product of the two marginals

$$X \perp Y \Leftrightarrow P(X,Y) = P(X)P(Y)$$

- X and Y are *conditionally independent (CI)* given Z iff the conditional joint can be written as a product of conditional marginals:

$$X \perp Y \mid Z \Leftrightarrow P(X,Y \mid Z) = P(X \mid Z)P(Y \mid Z)$$

Pairwise Vs. Mutual Independence

- Pairwise independence does not imply mutual independence.

  - ✓ Consider 4 balls (numbered 1,2,3,4) in a box. You draw one at random. Define the following events:

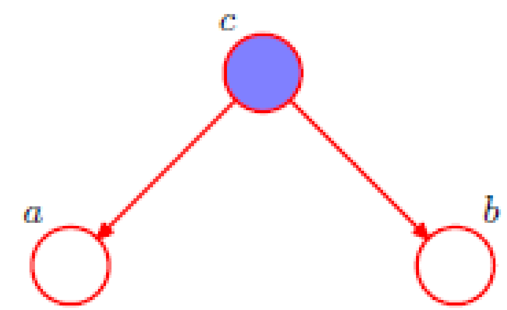
$$X_1 : ball\ 1\ or\ 2\ is\ drawn$$
$$X_2 : ball\ 2\ or\ 3\ is\ drawn$$
$$X_3 : ball\ 1\ or\ 3\ is\ drawn$$

  - ✓ Note that the three events are pairwise independent, e.g.

$$X_1 \perp X_2 \Leftrightarrow P(X_1, X_2) = \underbrace{P(X_1)}_{1/2}\underbrace{P(X_2)}_{1/2} = \frac{1}{4}$$

  - ✓ However: $P(X_1, X_2, X_3) = 0,\ \underbrace{P(X_1)}_{1/2}\underbrace{P(X_2)}_{1/2}\underbrace{P(X_3)}_{1/2} = \frac{1}{8}$

Independence, Conditional Independence



- Consider the following example. Define:
- Event a = `it will rain tomorrow'
- Event b =`the ground is wet today' and
- Event c =`raining today'.
- c causes both a and b – thus given c we don't need to know about b to predict a

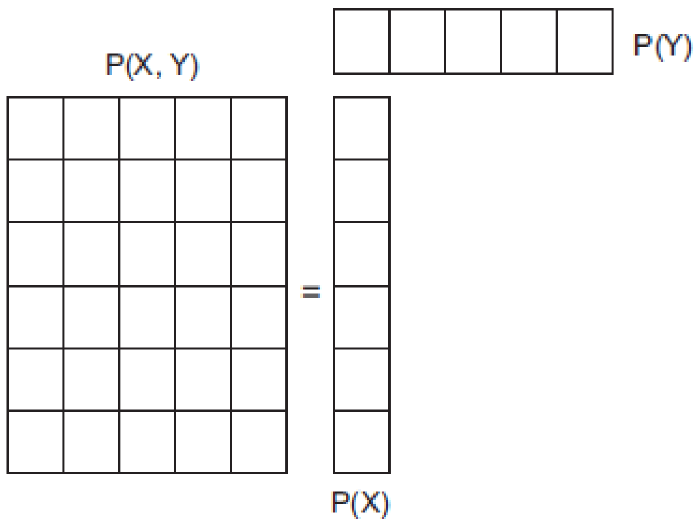
$$a \perp b \mid c \Leftrightarrow P(a,b \mid c) = P(a \mid c)P(b \mid c)$$

$$a \perp b \mid c \Leftrightarrow P(a \mid b,c) = P(a \mid c)$$

- Observing a "root node", separates "the children"!

Independence, Conditional Independence

- Assume unconditional independence $X \perp Y$. Let X take 6 values and Y takes 5 values. The cost for defining $p(X,Y)$ is drastically reduced if $X \perp Y$.



- The parameters are reduced from $29\ (= 30 - 1)$ to $9 = 5 + 4 = (6-1) + (5-1)$.[a] *Independence is key to efficient probabilistic modeling* (naïve Bayes classifiers, Markov Models, graphical models, etc.).

[a] We subtract one on the counts to account for the sum-to-one probability constraint rule.