Mathematical Background
Probability Theory

Palacode Narayana Iyer Anantharaman

5 Aug 2016

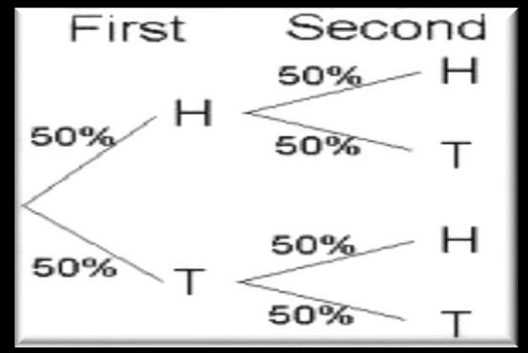
References

- Deep Learning book, Chapter 3, Probability and Information Theory – Ian Goodfellow, Yoshua Bengio, Aaron Courville

Why Probability for Machine Learning?

- Consider the following example scenarios:
  - You are travelling in an autorikshaw on a busy road in Bangalore and are a on a call with your friend.
  - We are watching an Hollywood English film. We may not understand exactly every word that is spoken either due to the accent of the speaker or the word is a slang that not everyone outside the context can relate to.
  - We are reading tweets that are cryptic with several misspelled words, emoticons, hashtags and so on.

- Commonality in all the above cases is the presence of noise along with the signal



- The noise or ambiguities result in uncertainty of interpretation

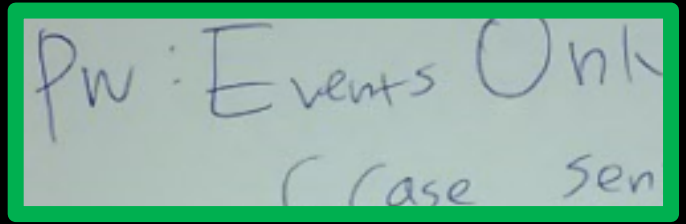
- To process such text, we need an appropriate mathematical machinery.

- Probability theory is our tool to handle such cases.

Sources of Uncertainty

- Inherent stochasticity
  - Quantum mechanics – the Heisenberg's **uncertainty** principle states that one can't exactly determine the position and momentum of a particle simultaneously
  - Will all phones of a given model, say, iPhone 6, have exactly the same weight, even if they are produced using the same process?

- Incomplete Observability
  - What are the words you see in the image shown?



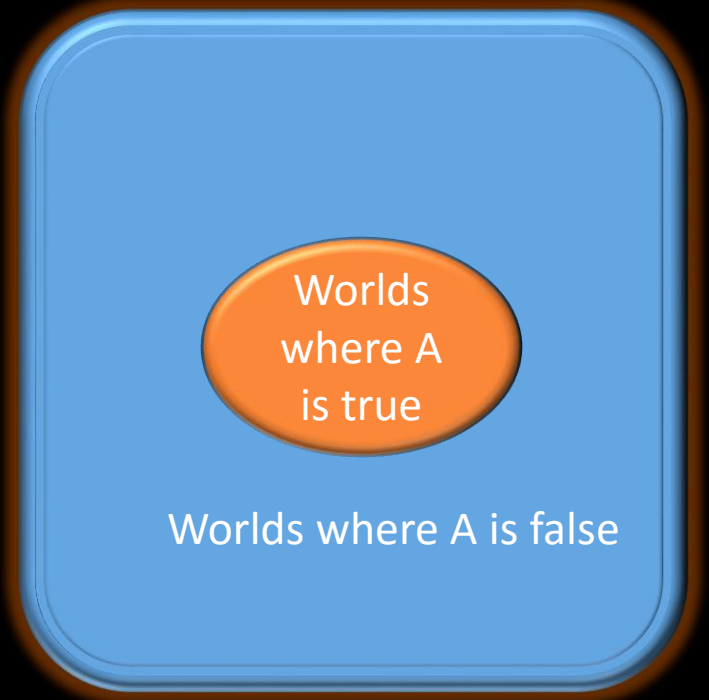
- Incomplete Modelling
  - Sub sampling a high resolution image to a lower resolution loses some information that leads to uncertainty

Notion of Frequentist and Bayesian Probabilities

- Some real world events are repeatable and some or not
  - We can toss a coin or roll a dice or shuffle cards any number of times and observe the outcomes. We can repeat such experiments many times
  - If a nuclear scientist states that the probability of a nuclear accident is only once every million years, will we be able to verify the claim with some repeatable experiments?
- Frequentist notion of probability applies to situations where we can repeat events and observe the frequencies of occurrence of outcomes
- When the probability is related to qualitative beliefs, we are dealing with priors and Bayesian probability
- Both Bayesian and frequentist models of probability obey the same rules!

What is a Random Variable?

- **A** is a Boolean valued RV if **A** denotes an event and there is some degree of uncertainty to whether **A** occurs.
  - Example: It will rain in Manchester during the 4th Cricket test match between India and England
- Probability of A is the fraction of possible worlds in which A is true
- The area of blue rectangle = 1
- Random Variable is not a variable in the traditional sense. It is rather a function mapping.



Types of Random Variables

- Random Variables can be:
  - Boolean
    - Side of a coin that can take values: Head, Tails
  - Discrete, multivalued
    - The red pixel value of a pixel in an RGB image
  - Continuous
    - The screen size of a mobile phone
  - A “feature” vector
    - Weather record: (minimum_temperature, maximum_temperature, humidity, chance_of_rain)

Axioms of Probability

The following axioms always hold good:

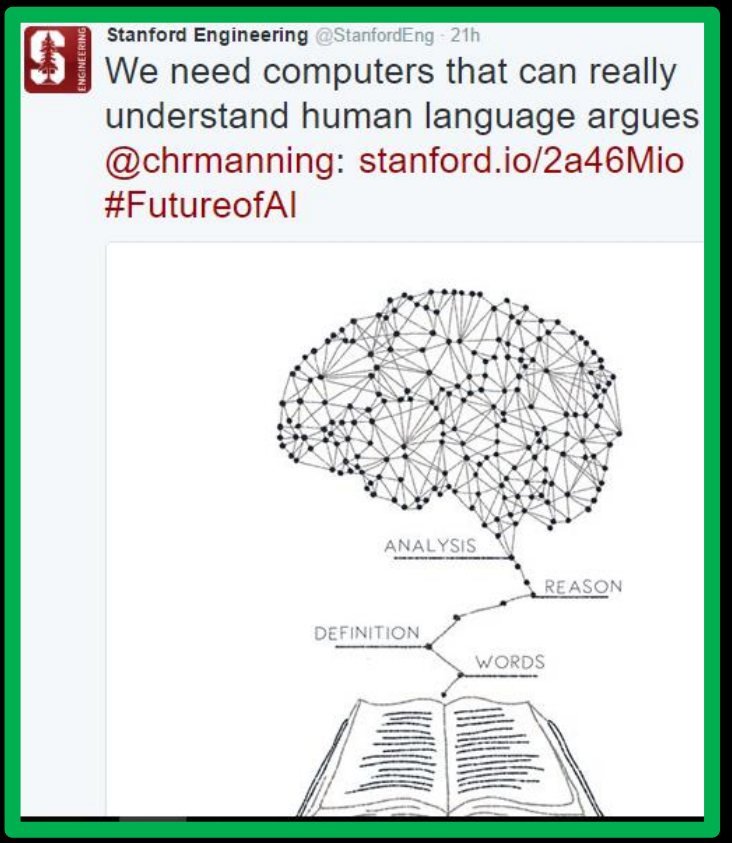
- $0 <= P(A) <= 1$
- $P(True) = 1$
- $P(False) = 0$
- $P(A \text{ or } B) = P(A) + P(B) – P(A \text{ and } B)$

Note: We can diagrammatically represent the above and verify these

Multivalued Discrete Random Variables

Examples of multivalued RVs

- Number of hashtags in a Tweet
- Number of URLs in a tweet
- Number of tweets that mention the term iPhone in a given day
- Number of tweets sent by Times Now channel per day



Example : Discrete Random Variable

- We can treat the number of hashtags contained in a given tweet (h) as a RV
- A tweet can have a number of hashtags: 0 to n, where n = 46
  - If a tweet has 140 characters and if the minimum string length required to specify a hashtag is 3 (one # character, one character for the tag, one delimiter), then n = floor(140 / 3) = 46
- As we observe, h is a discrete RV taking values: $0 <= h <= 46$
- The probability of finding h hashtags in a tweet satisfies:

  $0 <= P(h \text{ hashtags in a tweet}) <= 1$

  $P(h = 0 \text{ V } h = 1 \text{ V } h = 2 \ldots . \text{V } h = 46) = 1$ where h is the number of hashtags in the tweet

  $P(h = hi, h = hj) = 0$ when i and j are not equal – hi and hj are 2 different values of hashtag

Probability Distributions for discrete variables

- Suppose a random variable X can take on one of the several values (or states) from a finite set, we can describe a probability distribution over them.
- The probability distribution describes how likely the random variable takes a given state.
- The function that assigns probabilities to the different states of a random variable is termed a probability mass function (pmf)
- The pmf function is denoted by P. But P(X) and P(Y) refer to different distributions, though the function name is P
  - Suppose x is a value X can take, we may write: $x \sim P(X)$ to indicate that x distributes as P(X)

Properties of a PMF

- The domain of P is all possible states the random variable can take

$$\forall_x \in X, 0 \leq P(X) \leq 1$$

- Normalization:

$$\sum_{x \in X} P(X) = 1$$

- Example: Uniform Distribution
  - Consider a discrete random variable X that can hold any one of the 6 values of a fair die. The PMF for this: $P(X = x) = \frac{1}{6}$ and $\sum_{x \in X} P(X) = 1$
  - We can generalize this in to k states as $P(X = x) = \frac{1}{k}$

Probability Distributions for continuous variables

- When the random variable X can take real numbered values we describe a probability distribution over them using probability density function (PDF)

- The probability density function for the state x, that is, p(x) refers to the area under the curve of the infinitesimally small region between x and x+δx .

- The probability mass for the interval a, b is obtained by integrating p(x) over this interval:

$$probability\ of\ x\ lying\ in\ the\ interval\ (a,b): \int_a^b p(x)dx$$

Properties of a PDF

- The domain of P is all possible states the random variable can take

$$\forall_x \in x, \qquad p(x) \geq 0$$

- Normalization:

$$\int p(x)dx = 1$$

- Example: Uniform Distribution u(x; a, b) where [a, b] is the interval and b > a

$$u(x; a, b) = 0, when\ x \notin [a, b]$$

$$Within\ [a, b], u(x; a, b) = \frac{1}{b - a}$$

Example: Continuous Random Variables

- A sentiment polarity as a real number predicted by a sentiment analyzer is an example of a continuous RV

- While the sentiment polarity can be a scalar variable, it can also be a vector of continuous random variables. For example, some systems model emotions as a multi dimensional vector of real.

- Likewise a vector whose elements are the average values of hashtag, URL, Screen Names, Retweets per tweet, averaged over a corpus constitutes a vector of continuous Random Variables

Joint Distribution of Discrete Variables

- We described the notion of probability distribution for a discrete random variable X
- We can generalize this for multiple random variables, say: X, Y, Z
- Such a distribution that describes the probability of many discrete random variables taking on specific values is termed a joint probability distribution.
  - $P(X = x, Y = y, Z = z)$ where X, Y, Z are discrete RVs and x, y, z are the values (or states) that the respective RVs can take. For brevity we may refer this as P(X, Y, Z)
- To be a valid probability distribution the PMF needs to satisfy the axioms of probability

Joint Distribution of Discrete Random Variables

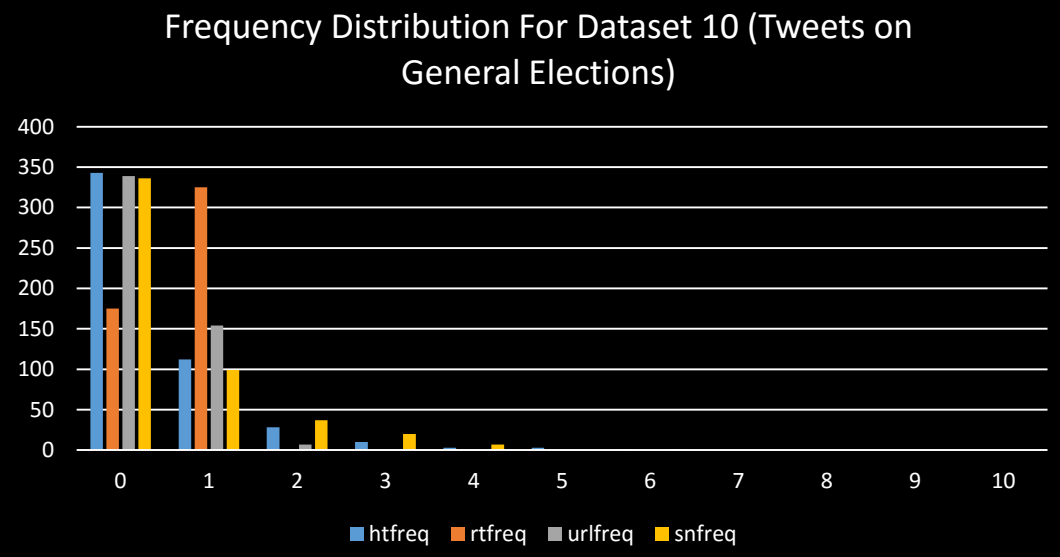
- Consider 2 RVs X and Y, where X and Y can take discrete values. The joint distribution is given by: $P(X = x, Y = y)$

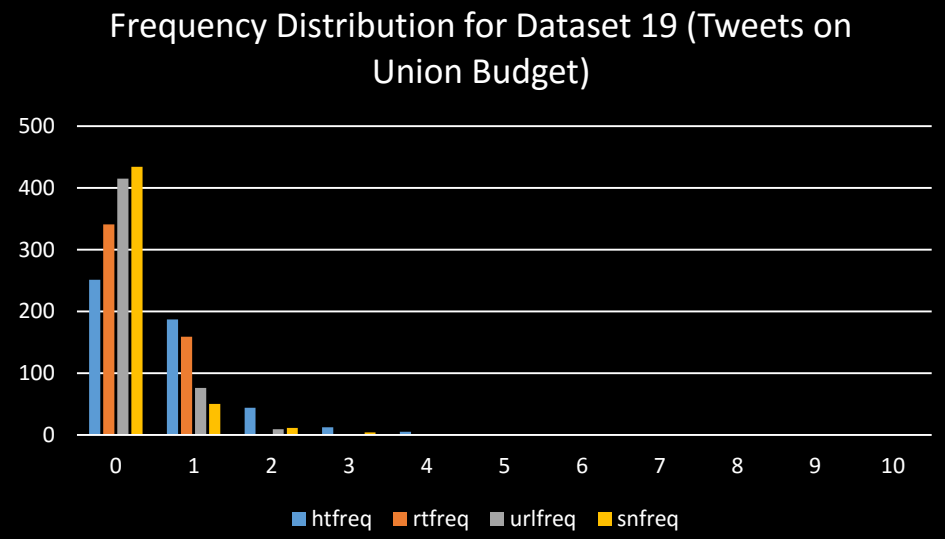
- The above satisfies:
  1. $P(X, Y) >= 0$
  2. $\Sigma \Sigma P(X = xi, Y = yj) = 1$ where the summation is done for all i and all j

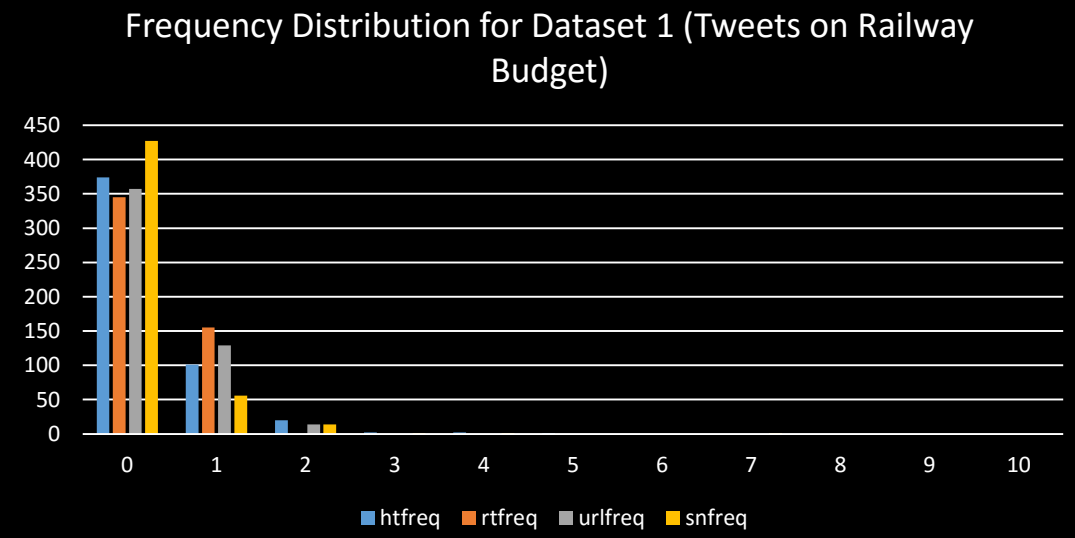
Table shows an example of joint distribution over number of hashtags, retweets, URLs and screen names of a tweet corpus.

HASHTAG	RT	URL	SNAME
0	0	1	0
3	0	1	1
1	0	0	0
0	0	0	0
0	0	1	0
0	0	0	0
1	0	0	1
0	0	0	0
0	0	0	0
0	0	0	0
1	0	1	0
1	1	0	0
1	0	1	0
1	0	0	0
0	0	1	0
3	0	0	0
0	1	0	0

Frequency Distributions for our Tweet Corpora







Some Observations

- Frequency distribution shows distinguishable patterns across the different corpora. For e.g, in the case of Union budget, the number of tweets that have hashtags one or more is roughly 50% of total tweets while that for Railway budget is roughly 25%
- The expectation value of hashtags/Retweets/URLs/Screen names per tweet can be treated as a vector of real valued Random Variables and used for classification.

Random Variables: Illustration

- Suppose we follow tweets from 2 media channels, say, Times News (x1) and Business News (x2). x1 is a general news channel and x2 focuses on business.

- Suppose we measured the tweets generated by either of these 2 channels over a 1 month duration and observed the following:
  - Total tweets generated (union of tweets from both channels) = 1200
  - Break up of the tweets received as given in the table below:

	x1	x2
y1	600	40
y2	200	360

Random Variables

- We can model the example we stated using 2 RVs:
  - Source of tweets (X), Type of tweet (Y)

	x1	x2
y1	600	40
y2	200	360

- X and Y are 2 RVs that can take values:
  - X = x where x ε {x1, x2}, Y = y where y ε {y1, y2}

- We can compute prior probabilities and likelihood:

  P(X = x1) = (600 + 200)/1200 = 2/3 = 0.67

  P(X = x2) = 1 – 2/3 = 1/3 = 0.33

  P(Y = y1 | X = x1) = 600 / (600 + 200) = 3/4 = 0.75

  P(Y = y2 | X = x1) = 0.25

  P(Y = y1 | X = x2) = 0.1

  P(Y = y2 | X = x2) = 0.9

Conditional Probability

- Conditional probability is the probability of an event, given the other event has occurred.

$$Conditional\ Probability\ of\ Y = y_j\ given\ X = x_i$$

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

$$n_{ij} = number\ of\ instances\ where:\ Y = y_j\ and\ X = x_i$$

$$c_i = number\ of\ instances\ where\ X = x_i$$

- Example:
  - Assume that we know the probability of finding a hashtag in a tweet. Suppose we have a tweet corpus C on a domain, where there is a increased probability of finding a hashtag. In this example, we have a prior idea about the probability of finding a hashtag in a tweet. When given an additional fact that the corpus from where the tweet was drawn was C, we now can revise our probability estimate on hashtag, which is: P(hashtag|C). This is called posterior probability

Sum Rule

In our example:

P(X = x1) =

P(X = x1, Y = y1) + P(X = x1, Y = y2)

Note:

P(X = x1) + P(X = x2) = 1

	x1	x2
y1	**600**	**40**
y2	**200**	**360**

The sum rule allows us to obtain marginal probability

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

Product Rule and Generalization

	x1	x2
y1	600	40
y2	200	360

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

$$= P(Y = y_j | X = x_i) P(X = x_i)$$

The above is the product rule

From product rule, we have: P(X, Y) = P(Y|X) P(X)

We can generalize this in to: P(An, ..,A1)=P(An|An-1..A1)P(An-1,..,A1)

For n = 4: P(A4, A3, A2, A1) = P(A4|A3, A2, A1) P(A3|A2, A1) P(A2|A1) P(A1)

Bayes Theorem

From product rule, we have:

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

We know: $P(X, Y) = P(Y, X)$, hence:

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

From the above, we derive:

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

The above is the Bayes Theorem

Independence

- Independent Variables: Knowing Y does not alter our belief on X

  From product rule, we know:

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

  If X and Y are independent random variables:

  $P(X|Y) = P(X)$, hence: $P(X, Y) = P(X)\,P(Y)$

  We write: $X \perp Y$ to denote X, Y are independent

- Conditional Independence
  - Informally, suppose X, Y are not independent taken together alone, but are independent on observing another variable Z. This is denoted by: $X \perp Y \mid Z$
  - Definition: Let X, Y, Z be discrete random variables. X is conditionally independent of Y given Z if the probability distribution governing X is independent of the value of Y given a value of Z.

  $P(X|Y, Z) = P(X|Z)$, Also: $P(X, Y \mid Z) = P(X|Y, Z)\,P(Y|Z) = P(X|Z)\,P(Y|Z)$

Expectation Value

- For discrete variables:
  - Expectation value: $E[x] = \sum f(x)p(x)dx$
  - If a random sample is picked from the distribution, the expectation is simply the average value of f(x)
  - Thus, the values of hashtag, screen name etc shown in the previous slide correspond to their respective expectation values

- For continuous variables:
  - $E[x] = \int f(x)p(x)dx$

Variance

- Let X be a RV and $x_1, x_2, \dots, x_n$ are samples from its probability distribution.
- If the variance of the distribution is small, the sampled values $x_1, x_2, \dots, x_n$ would be relatively close to each other as if they cluster together around their expected value
- A large variance distribution would make the $x_i's$ farther apart
- Thus, the variance gives a measure of how much the values of a function of a random variable vary from the samples of the distribution

$$Variance\colon\ Var\big(f(x)\big) = E[(f(x) - E[f(x)])^2]$$

- Standard deviation is the square root of variance

Covariance Definition

- The covariance mathematically captures the notion of how much two random variables vary together.

- $Covariance\ Cov\big(f(x), g(y)\big) = E[(f(x) - E[f(x)])(g(y) - E[g(y)])]$

- Covariance Matrix of a random vector $x \in \mathbb{R}^n\ is\ an\ n \times n\ matrix\ such\ that$:
$$Cov(x)_{i,j} = Cov(x_i, x_j)$$

- The diagonal elements of covariance gives the variance
$$Cov(x_i, x_i) = Var(x_i)$$

Covariance Example

- Covariance is positive when higher values of one variable corresponds to higher values in the other. Likewise, if higher values of one corresponds to lower values of another, the covariance is negative
  - Suppose we have 2 random variables: Monsoon M = {Good, Bad}, Stock Index S = {Growth, Decline}
  - A good monsoon improves stock index and a bad one causes decline in the stocks. The covariance here between M and S is positive.

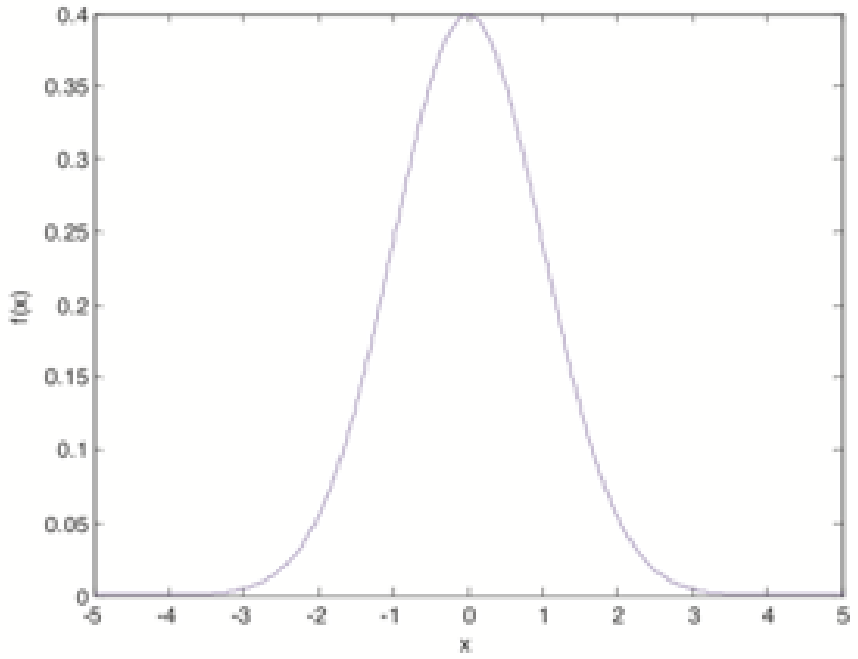
- Covariance and Correlation are related but different
  - Correlation normalizes the contribution of each variable in order to measure only how much the variables are related, rather than also being affected by the scale of the separate variables

- Covariance is zero is the 2 random variables are independent. But independence is a stronger requirement as one can look at non linear relations

Gaussian Distribution

- Normal X $N(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



$$\sigma = \sqrt{\frac{\sum (x-\mu)^2}{N}}$$

- Examples
  - Heights of people
  - Measurement errors in devices

Bernoulli Distribution

- Bernoulli distribution is the probability distribution of a binary random variable
- The binary random variable x takes the value x = 1 with a success probability μ and the value x = 0 with the failure probability of $1 - \mu$

$$Bern(x|\mu) = \mu^x(1-\mu)^{1-x}$$

- If we have a dataset D of N observations of x, the likelihood function is:

$$p(D|\mu) = \prod_{n=1}^{N} p(x_n|\mu)$$

- In several situations where we determine the model parameters of a machine learning classifier, we maximize the log likelihood. For a Bernoulli distributed variable, we can write:

$$\log P(D|\mu) = \sum_{n=1}^{n} \log P(x_n|\mu) = \sum_{n=1}^{n} [x_n \log \mu + (1-x_n)\log(1-\mu)]$$

Mixture Distributions

- Distributions such as Gaussian, Bernoulli, Poisson etc. are commonly used.
- It is also possible to define probability distributions by combining other probability distributions in order to create a richer distribution.
- One common way of combining distributions is to construct a *mixture distribution*.
- A mixture distribution is made up of several component distributions.
  - For instance, Gaussian Mixture Models (GMMs) are a linear combination of several Gaussians and these are used to model speech recognition tasks

Estimating probabilities

- Joint Distributions are awesome for estimating probabilities!
  - We can determine the probability of any logical expression involving the random variables once we have a joint distribution over them

- Unfortunately, it is not practical to get a complete joint distribution table when the number of variables are large and they can take many values.
  - E.g: A feature vector having 20 Boolean elements results in 1 million entries to the joint distribution table.
  - Many real world applications might use hundreds of features

Estimating Probabilities

- How to estimate probabilities from a finite sized data?
  - Maximum Likelihood Estimation
  - Maximum a posteriori estimation

Head or Tail

- Suppose we flip the coin shown here and observe that:
  - α_1 times heads turn up
  - α_2 times tails turn up

- What is the probability estimate of finding a head P(X=1)?

Head or Tail

- Case 1: We tossed 100 times, observed Heads in 51 trials and Tails showed up in 49 trials. What is P(X = 1)?

- Case 2: We tossed 5 times, observed Heads in 4 trials and Tails showed up in 1 trial. What is P(X = 1)?

Head or Tail

- Assume that we keep flipping and make our model estimation every step. We need an algorithm that can give us good estimates after each flip.

$$\alpha_1 = number\ of\ observed\ heads, X = 1$$
$$\alpha_0 = number\ of\ observed\ tails, X = 0$$

- Let us hallucinate that there are some heads turning up in addition to the trials we performed.

$$\beta_1 = number\ of\ hallucinated\ heads$$
$$\beta_0 = number\ of\ hallucinated\ tails, X = 0$$

- With the above "smoothing", we have:

$$P(X = 1) = \frac{\alpha_1 + \beta_1}{(\alpha_1 + \beta_1) + (\alpha_0 + \beta_0)}$$

Estimation Principles: MLE and MAP

- MLE: Choose parameters that maximize likelihood $P(data|\theta)$

$$P(X = 1) = \frac{\alpha_1}{(\alpha_1 + \alpha_0)}$$

- MAP: Choose parameters that maximize $P(\theta|data)$

$$P(X = 1) = \frac{\alpha_1 + \beta_1}{(\alpha_1 + \beta_1) + (\alpha_0 + \beta_0)}$$

Information Theory

- The basic intuition behind information theory is that a common event that has a high probability of occurrence has less information content.

- Rare events have greater information content

- Independent events carry additive information

Intuition behind entropy

- Consider a discrete random variable x. How much information we receive when we observe a specific value of x?

- The amount of information can be viewed as the degree of surprise on learning the value of x. A highly improbable event provides more information compared to a more likely event.
  - When an event is certain we receive zero information. E.g. Sun raises in the east.

- Hence the information h(x) is a monotonic function of probability p(x)

Information and Independent Events

- If there are 2 independent events, the information we receive on both the events is the sum of information we gained from each of them separately.
  - Hence: $h(x, y) = h(x) + h(y)$
- Two unrelated events will be statistically independent if: $p(x, y) = p(x)\ p(y)$
- From the above 2 relationships we deduce that $h(x)$ should be related to $\log p(x)$.
  - Specifically: $h(x) = - \log_2 p(x)$.

Entropy

- Suppose a sender transmits the value of a random variable to a receiver.

- The average amount of information transmitted is obtained by taking the expectation of h(x) with respect to the distribution p(x) and is given by:

  $H[x] = -\sum_x p(x)\, log_2\, p(x)$, $H[x]$ is called the Entropy of the random variable x

- If p(x) and q(x) are two probability distributions occurring over the same set of events, we define the cross entropy to be:

$$H(p, q) = -\sum_x p(x)\, log_2\, q(x)$$

Entropy and Uniform Distribution

- Consider a random variable that can assume one of 8 possible values, where each value is equally likely.

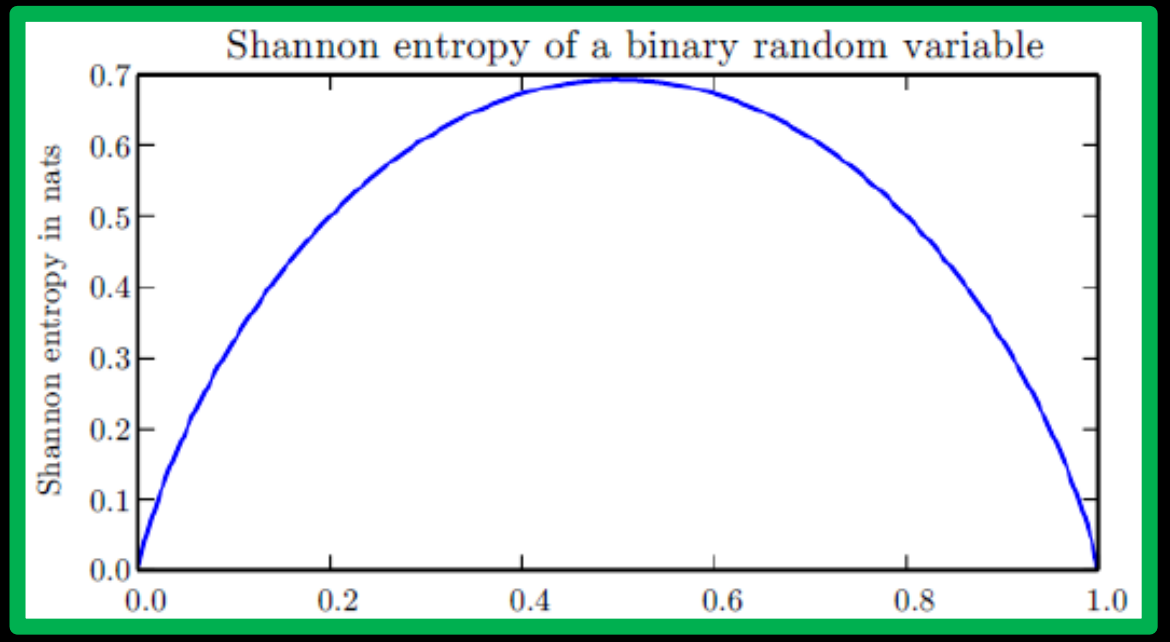
  $H[x]$ is given by $-8 \times \frac{1}{8} \times log_2 \frac{1}{8}$ = 3 bits

- Now consider the distribution to be:

  p(x) = $\{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\}$.

  The entropy turns out to be 2 bits

- If a RV has a uniform probability distribution, the entropy is maximum



Kullback-Leibler (KL) Divergence

- Suppose P(x) and Q(x) be two probability distributions over the same RV

- KL Divergence is a means of measuring the difference between 2 distributions:

$$D_{KL}(P||Q) = E_{x\sim P}\left[\log\frac{P(x)}{Q(x)}\right] = E_{x\sim P}[\log P(x) - \log Q(x)]$$

- KL Divergence is non negative and is zero iff P and Q are same distributions

- KL Divergence is not symmetric: $D_{KL}(P||Q) \neq D_{KL}(Q||P)$. Hence KL Divergence is not strictly a distance metric.

- KL Divergence is related to cross entropy:

$$H(P,Q) = H(P) + D_{KL}(P||Q) = E_{x\sim P}\log Q(x)$$

Graph Models: Bayesian Networks

Graph models: Bayesian networks, belief networks and probabilistic networks

- Each node corresponds to a random variable X and the value of the node is the probability of X
- If there is a direct edge between two vertices X to Y, it means there is a influence of X on Y
- This influence is specified by the conditional probability P(Y|X)
- This is a DAG
- Nodes and edges define the structure of the network and the conditional probabilities are the parameters given the structure

Examples

- Preparation for the exam R, and the marks obtained in the exam M
- Marketing budget B and the advertisements A
- Nationality of Team N and chance of qualifying for quarter final of world cup, Q
- In all cases, the **Probability distribution P respects the graph G**

Representing the joint distributions

- Consider P(A, B, C) = P(A) P(B|A) P(C|A, B). This can be represented as a graph (fig a)
- Key Concept: Factorization
- The joint probability distribution with conditional probability assumptions respects the associated graph.
- The graph of the distributions useful for: Visualization of conditional dependencies and Inferencing
- Determining Conditional Independence of a distribution is vital for tractable inference

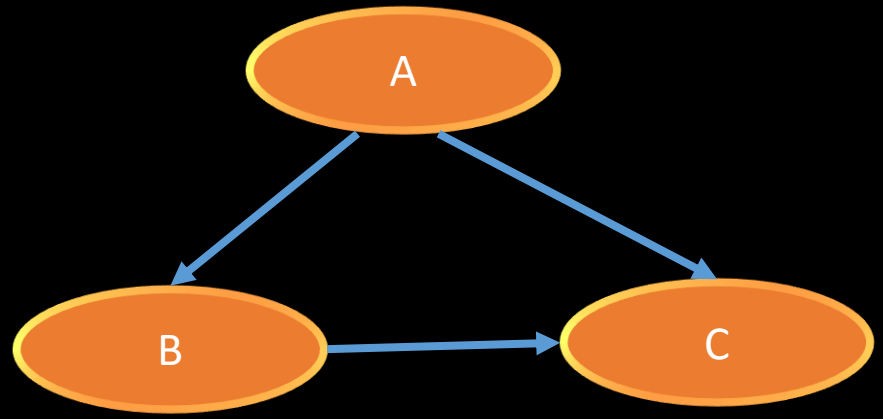


Fig (a)

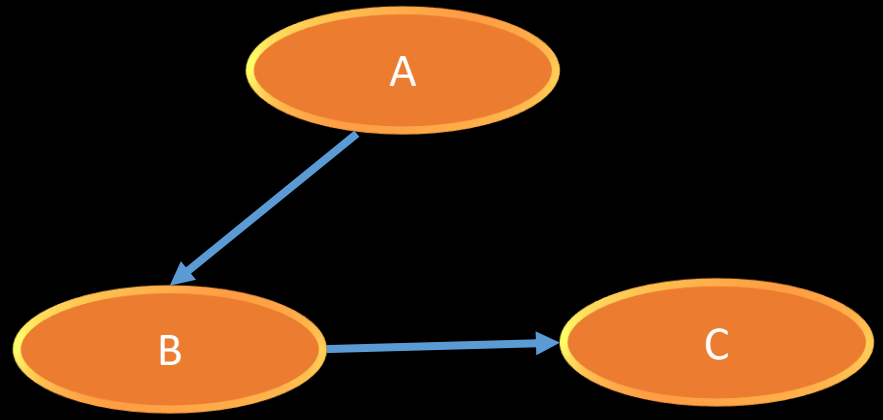


Fig (b)