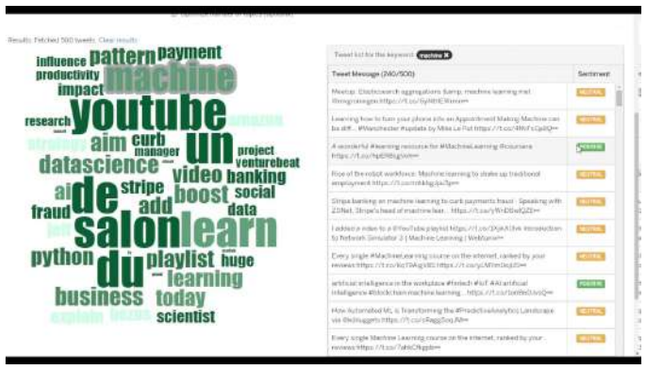
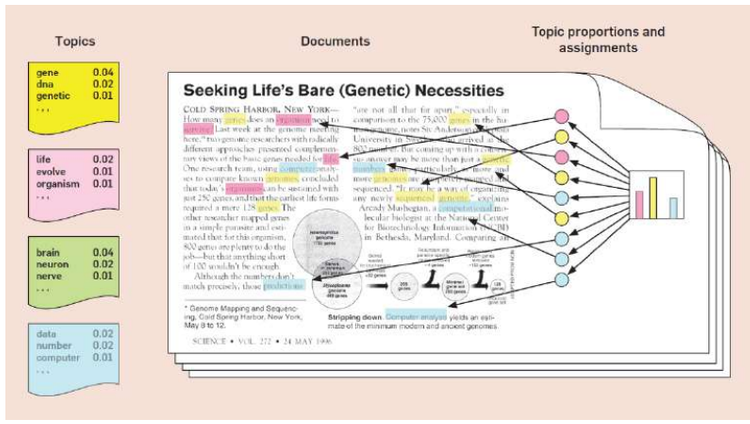
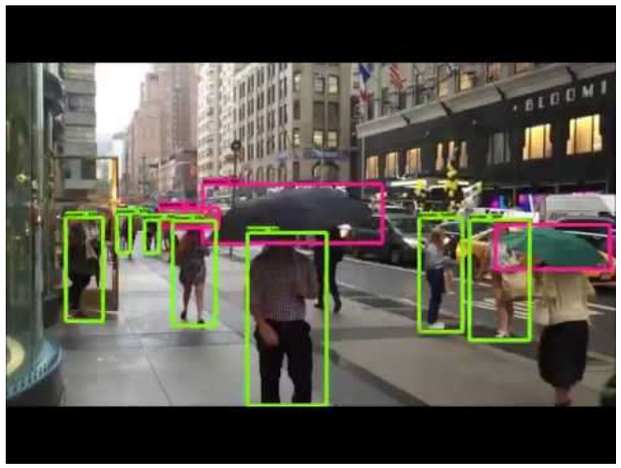
Probability in data science

Dr. Srijith P K

data Science







Big data era

"We are drowning in information and starving for knowledge." – John Naisbitt.

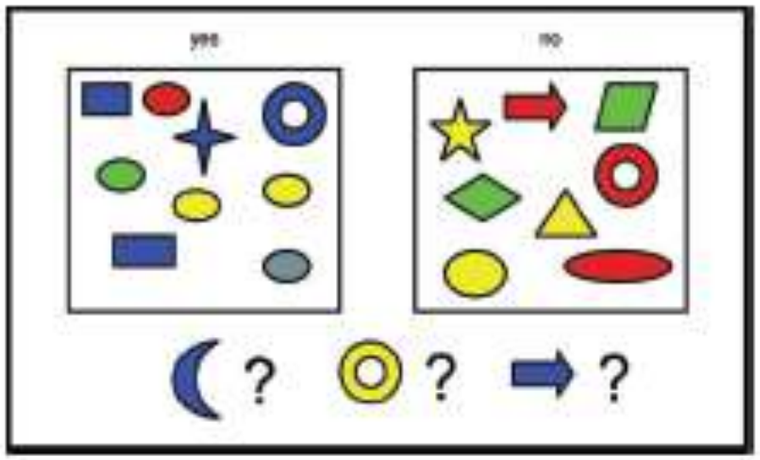
- How many web pages in the web ?
- rate (per second) at which video is uploaded to YouTube ?
- Number of transactions per hour handled by Walmart/Amazon and their data base sizes ?

This deluge of data calls for automated methods of data analysis, which is what machine learning provides.

 "set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty"
- Kevin Murphy (Machine Learning : A Probabilistic Perspective)

Need for Probabilistic machine Learning

- Probability theory provides a mathematical framework to handle uncertainty



(a)

D features (attributes)

N cases

Color	Shape	Size (cm)	Label
Blue	Square	10	1
Red	Ellipse	2.4	1
Red	Ellipse	20.7	0

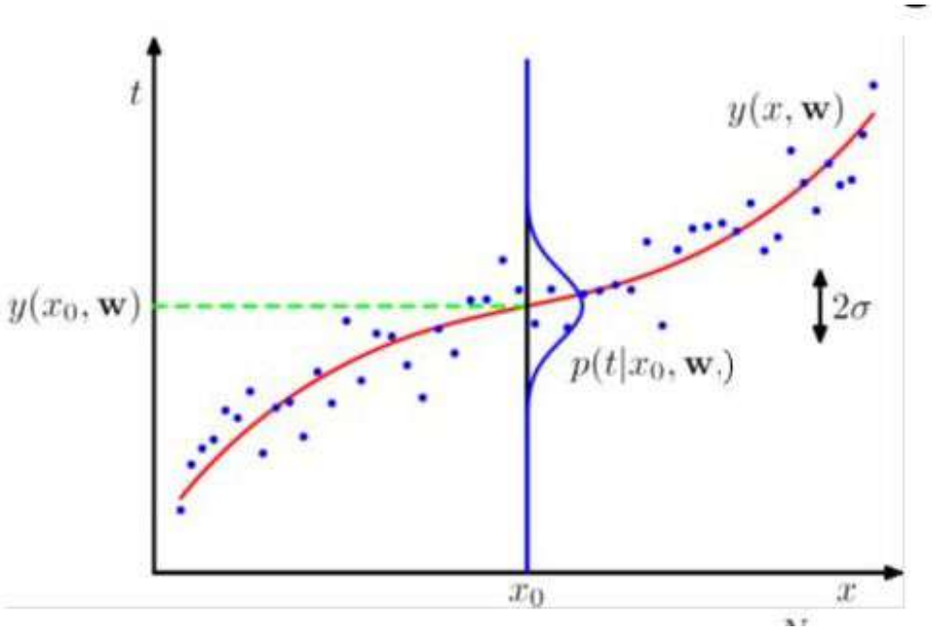
(b)

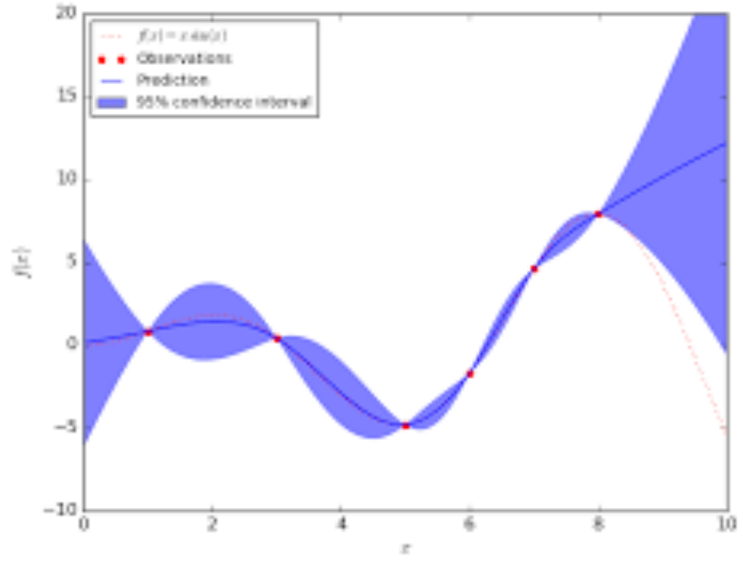
Probability in machine learning

- Probability theory can be applied to any problem involving uncertainty.
- In machine learning, uncertainty comes in many forms:
  - what is the best prediction about the future given some past data?
  - what is the best model to explain some data
  - what measurement should I perform next?

Probabilistic machine learning models - regression

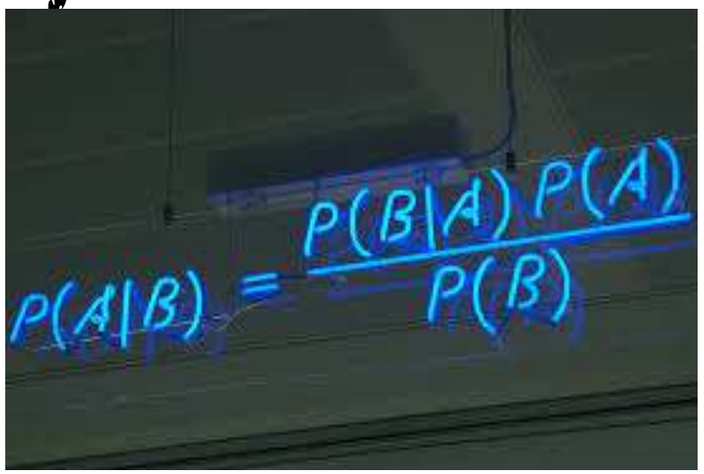
- Probabilistic linear regression
- Bayesian logistic regression
- Gaussian process regression

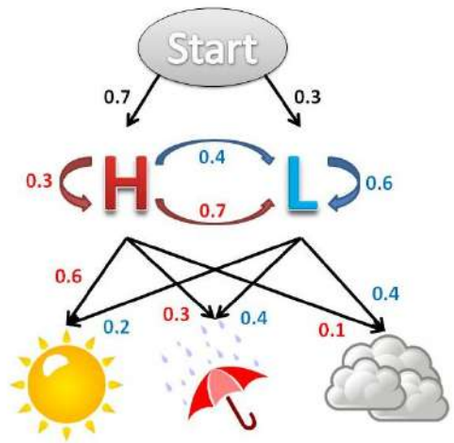


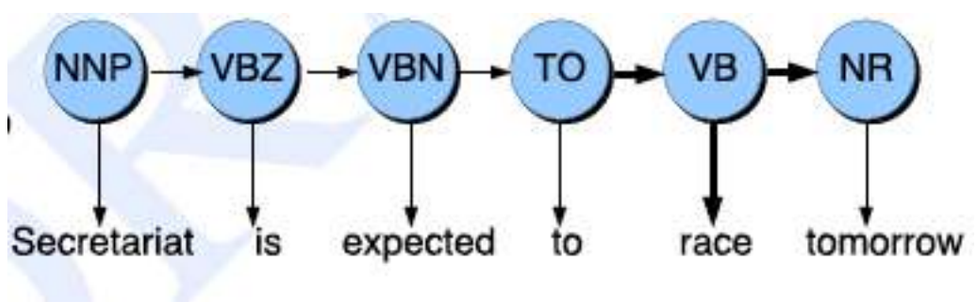


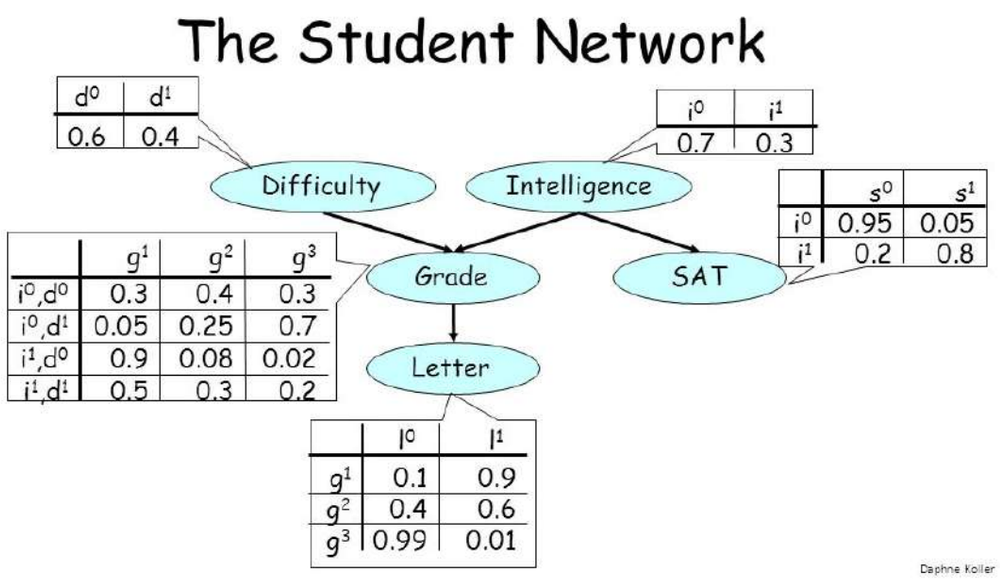
Probabilistic machine learning models - Classification

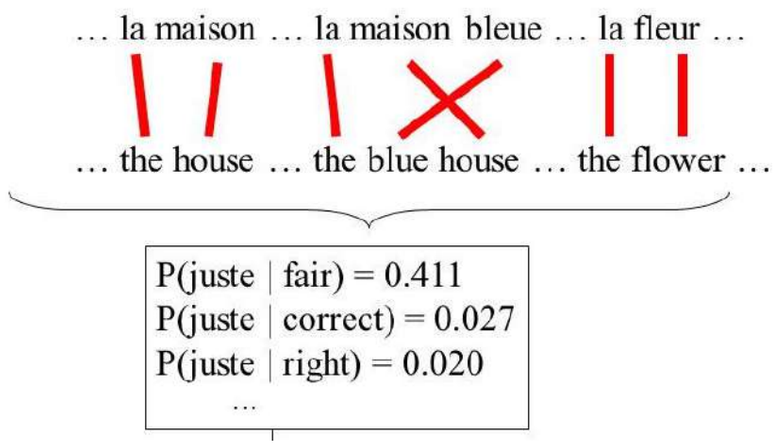
- Naive bayes classifier
- Logistic regression
- Sequence classification
  - Hidden markov models
- Conditional Random Fields





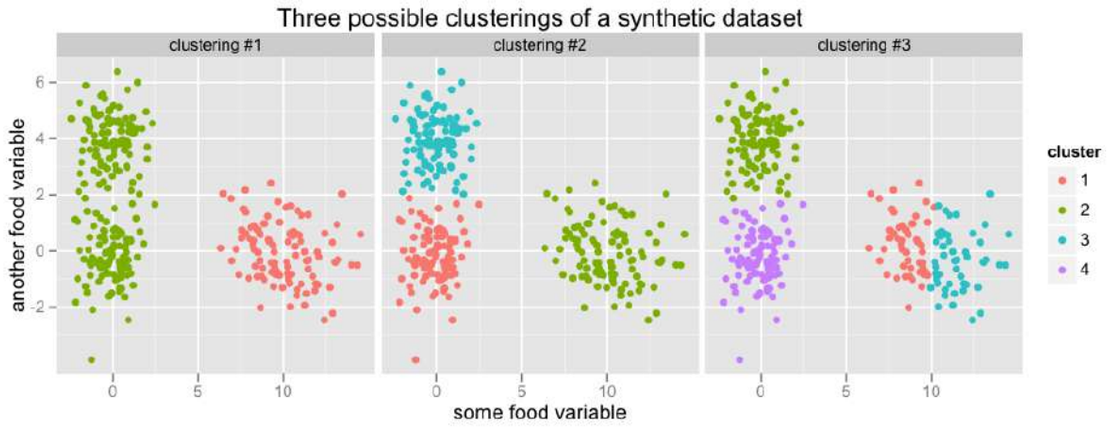


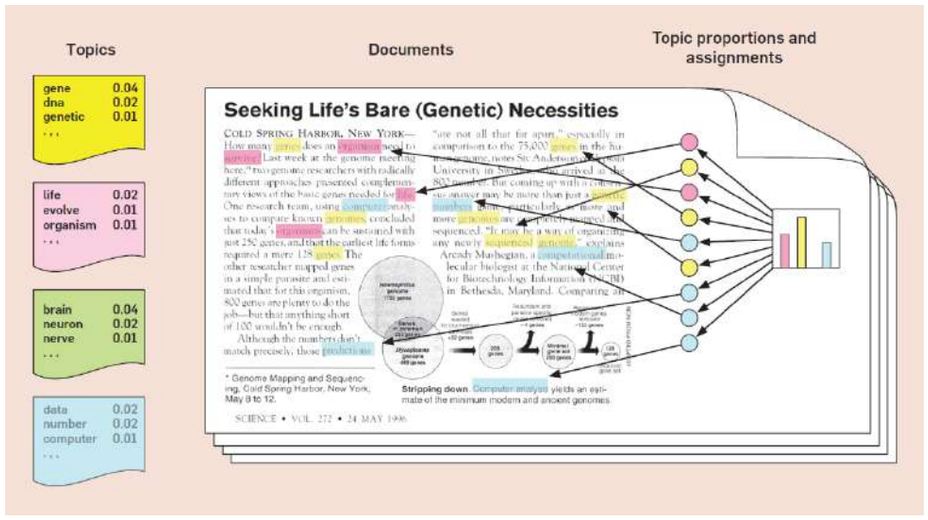


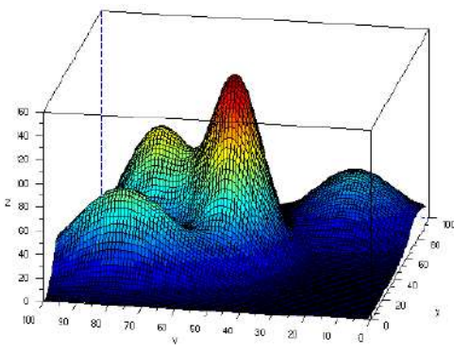


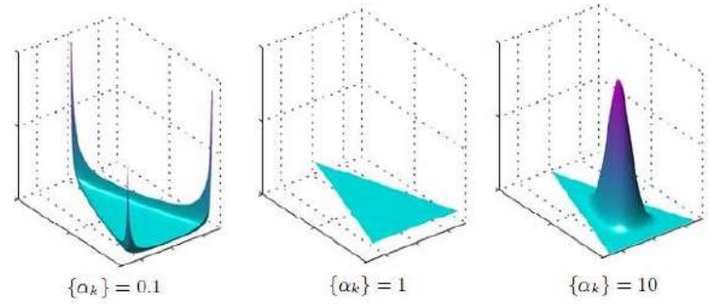
Probabilistic machine learning models - Clustering

- Gaussian Mixture Model
- Latent Dirichlet Allocation
- Dirichlet Process Mixture









Probability

“Probability theory is nothing but common sense reduced to calculation” — Pierre Laplace, 1812

Whats Probability ?

Coin toss : “the probability that a coin will land heads is 0.5”.
“probability that the polar ice cap will melt by 2020 CE”

- Frequentist interpretation
- Bayesian interpretation

Compute the probability that a specific email message is spam.

Games of chance provided the impetus for the mathematical study of probability

Probabilistic interpretation

- geologist quotes “there is a 60 percent chance of oil in a certain region,”
  - the geologist feels that, over the long run, in 60 percent of the regions whose environmental conditions are very similar to that of the region under consideration, there will be oil;
  - the geologist believes that it is more likely that the region will contain oil than it is that it will not; and in fact .6 is a measure of the geologist’s belief in the hypothesis that the region will contain oil.

Probability Theory

- The theory of probability is a representation of its concepts in formal terms—that is, in terms that can be considered separately from their meaning. These formal terms are manipulated by the rules of mathematics and logic, and any results are interpreted or translated back into the problem domain.

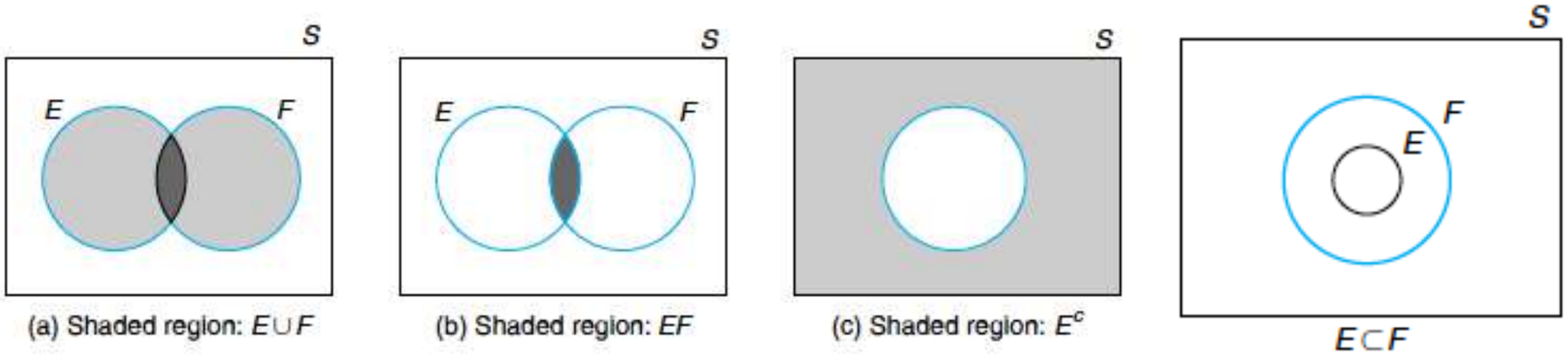
- There have been at least two successful attempts to formalize probability, namely the Kolmogorov formulation and the Cox formulation. In Kolmogorov's formulation, sets are interpreted as events and probability itself as a measure on a class of sets. In Cox's theorem, probability is taken as a primitive and the emphasis is on constructing a consistent assignment of probability values to propositions. In both cases, the laws of probability are the same.

SAMPLE SPACE AND EVENTS

- set of all possible outcomes of an experiment is known as the sample space (S)
- If the outcome of an experiment consists in the determination of the sex of a newborn child, then S = {g , b}
- Suppose we are interested in determining the amount of dosage that must be given to a patient until that patient reacts positively let S = $(0,\infty)$
- Any subset E of the sample space is known as an event. That is, an event is a set consisting of possible outcomes of the experiment. E = {g }, then E is the event that the child is a girl
- Dice experiment : whats S and E

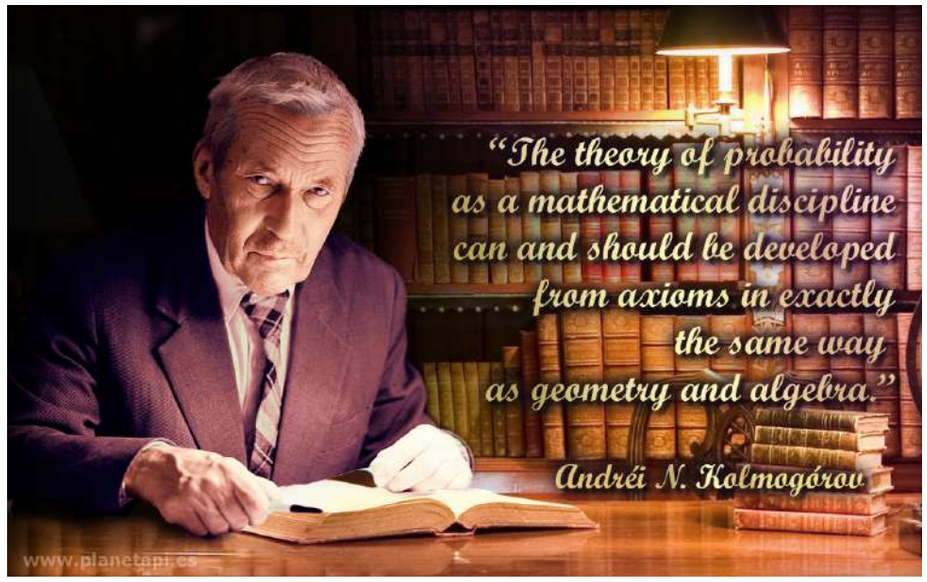
Sample Space and Events

- union of the events E and F, consist of all outcomes that are either in E or in F or in both E and F.E = {g } and F = {b}, then E ∪ F = {g , b}.
- EF, called the intersection of E and F, to consist of all outcomes that are in both E and F. E = (0, 5), F =(2, 10), EF =(2, 5)
- EF = ∅, implying that E and F cannot both occur, then E and F are said to be mutually exclusive.
- Ec referred to as the complement of E, to consist of all outcomes in the sample space S that are not in E. if E = {b}Ec = {g }
- E is contained in F and write E ⊂ F, if E ⊂ F and F ⊂ E, then E=F



Axioms of Probability

- Probability : the proportion of time that the outcome is contained in E (frequentist)/strength of belief(bayesian)
- Kolmogorov Axioms/Cox theorem



AXIOM 1

$$0 \leq P(E) \leq 1$$

AXIOM 2

$$P(S) = 1$$

AXIOM 3

For any sequence of mutually exclusive events $E_1, E_2, \ldots$ (that is, events for which $E_i E_j = \emptyset$ when $i \neq j$),

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} P(E_i), \qquad n = 1, 2, \ldots, \infty$$

We call $P(E)$ the probability of the event E.

suppose the experiment consists of the rolling of a pair of dice and suppose that E is the event that the sum is 2, 3, or 12 and Fis the event that the sum is 7 or 11. Then if outcome E occurs 11 percent of the time and outcome F 22 percent of the time, then 33 percent of the time the outcome will be either 2, 3, 12, 7, or 11.

Probability

- P(Ec) = 1–P(E); P(E ∪ F) = P(E)+P(F)–P(EF)
- Each point in S={1,…,N} is equally likely

$$P(E) = \frac{\text{Number of points in } E}{N}$$

- P({1}) = P({2}) = …. = P({N})= p

- A total of 28 percent of Indian males smoke cigarettes, 7 percent smoke cigars, and 5 percent smoke both cigars and cigarettes. What percentage of males smoke neither cigars nor cigarettes?

Probability Space

- A probability space is a mathematical triplet (S, F, P) consisting of

1. A sample space, S , which is the set of all possible outcomes.
2. A set of events F , where each event is a set containing zero or more outcomes.
3. The assignment of probabilities to the events; that is, a function P from events to probabilities.

If the experiment consists of just one flip of a fair coin, then the outcome is either heads or tails:S ={H,T} , F={{},{H},{T},{H,T}}. There is a fifty percent chance of tossing heads and fifty percent for tails, so the probability measure in this example is P({})=0,P({H})=0.5,P({T})=0.5,P({H,T})=1

Conditional Probability

- Calculating probabilities when some partial information concerning the result of the experiment is available
- Rolls a pair of dice ; S = {(i, j), i = 1, 2, 3, 4, 5, 6, j = 1, 2, 3, 4, 5, 6}, each possible outcomes is equally likely, first die lands on side 3 (F), sum of the two dice equals 8 (E)?

$$P(E|F) = \frac{P(EF)}{P(F)}$$

- Naive Bayes classifier uses conditional probability of generating observation give a class

Conditional Probability

- Kannan figures that there is a 30 percent chance that his company will set up a branch office in Mumbai. If it does, he is 60 percent certain that he will be made manager of this new operation. What is the probability that Kannan will be a Mumbai branch office manager?

Conditional Probability

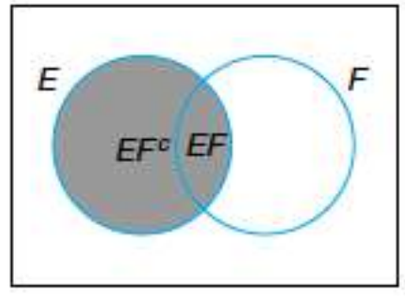
- The organization that Praveen works for is running a father–son dinner for those employees having at least one son. Each of these employees is invited to attend along with his youngest son. If Praveen is known to have two children, what is the conditional probability that they are both boys given that he is invited to the dinner?

Conditional Probability

- $E = EF \cup EF^c$
- $P(E) = P(EF)+P(EF^c)$

  $= P(E|F)P(F)+P(E|F^c)P(F^c)$

  $= P(E|F)P(F)+P(E|F^c)[1-P(F)]$



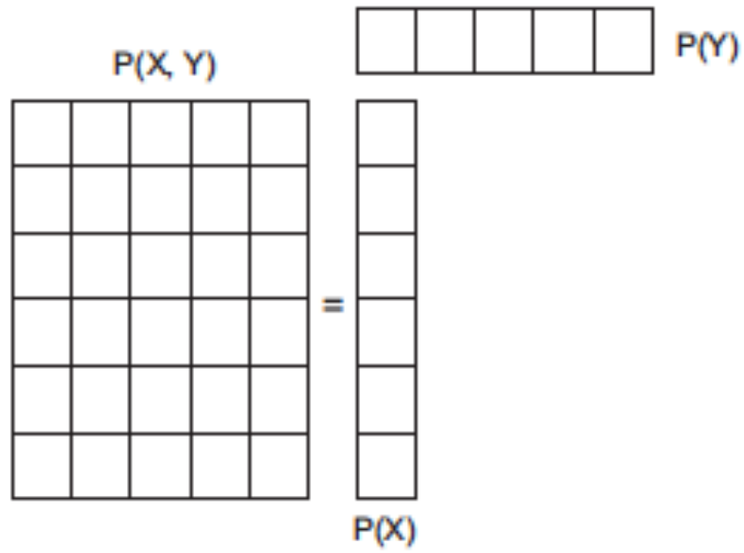
- Helpful when it is difficult to compute the probability of an event directly
- An insurance company believes that people can be divided into two classes — those that are accident prone and those that are not. Their statistics show that an accident-prone person will have an accident at some time within a fixed 1-year period with probability .4, whereas this probability decreases to .2 for a non-accident-prone person. If we assume that 30 percent of the population is accident prone, what is the probability that a new policy holder will have an accident within a year of purchasing a policy?

Independence

- Two events E and F are said to be independent iff P(EF) = P(E)P(F)
- If E and F are independent, then so are E and F c
- Data points are assumed to independent and identically distributed in most of the machine learning models
- A card is selected at random from an ordinary deck of 52 playing cards. If A is the event that the selected card is an ace and H is the event that it is a heart, then A and H are independent
- Conditionally independent : P(EF|G) = P(E|G)P(F|G) e.g Naive Bayes classifier, probabilistic graphical models.
- Probability it will rain tomorrow (event E) is independent of whether the ground is wet today (event F), given knowledge of whether it is raining today (event G)

Unconditionally or marginally independent

$$X \perp Y \iff p(X,Y) = p(X)p(Y)$$



- Computing p(x, y) = p(x)p(y), where $X \perp Y$. Here X takes 6 possible values and Y takes 5. A joint distribution on two such variables would require (6*5)–1 = 29 parameters to define it. By assuming (unconditional) independence, we only need (6 – 1) + (5 – 1) = 9 parameters to define p(x, y).

Independence

- A pair of fair dice is rolled. Let E denote the event that the sum of the dice is equal to 7.
- (a) Show that E is independent of the event that the first die lands on 4.
- (b) Show that E is independent of the event that the second die lands on 3.
-

Sum and Product rules in Probability

- Probability of the joint event A and B, $p(A,B) = p(A \wedge B) = p(A|B)\,p(B)$
- This is called the product rule .
- Given a joint distribution on two events p(A,B), we define the marginal distribution as follows:

$$p(A) = \sum_b p(A, B) = \sum_b p(A|B = b)p(B = b)$$

- This is called the sum rule or rule of total probability
- Product rule can be applied multiple times to yield the chain rule of probability, used to model sequences for e.g language modelling. P("I am Maunika") = P(I)P(am|I)P(Maunika|I, am)

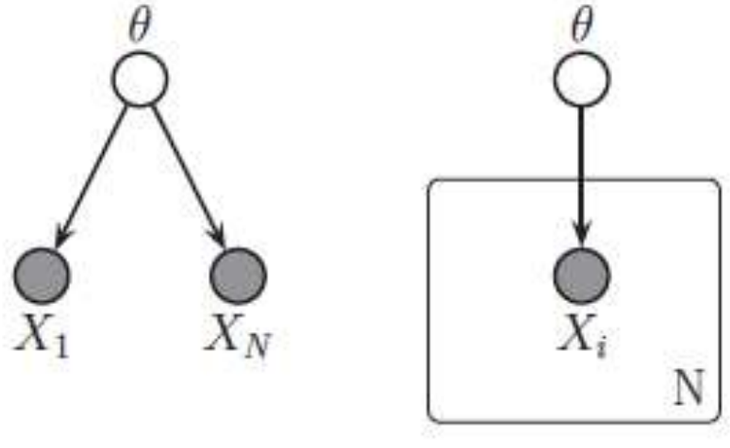
$$p(X_{1:D}) = p(X_1)p(X_2|X_1)p(X_3|X_2, X_1)p(X_4|X_1, X_2, X_3)\ldots p(X_D|X_{1:D-1})$$

Markov models

- Markov assumption : “the future is independent of the past given the present” xt+1 ⊥ x1:t–1|xt

$$p(\mathbf{x}_{1:V}) = p(x_1) \prod_{t=1}^{V} p(x_t|x_{t-1})$$

- This is called a (first-order) Markov chain .
- P(“I am Maunika”) = P(I)P(am|I)P(Maunika| am)
- Plate notation



Belief update

- Predict the personality
- Predict the concept

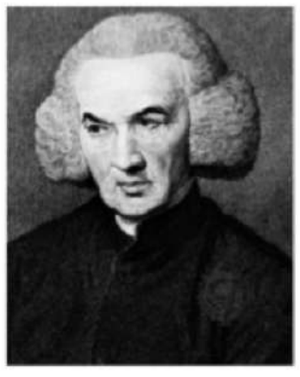
Bayes Theorem

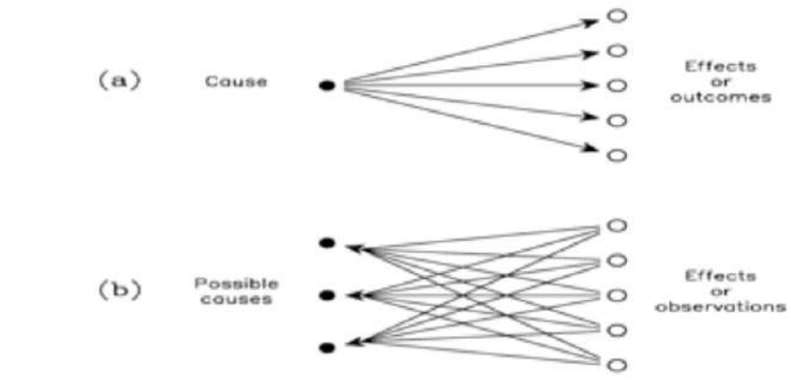
- Bayes formulae (Inverse Probability)
- Hypotheses held before the experiment P(H) to be modified by the evidence of the experiment P(H|E)
- It provides a formulae to calculate inverse probability P(H|E) from P(E|H)

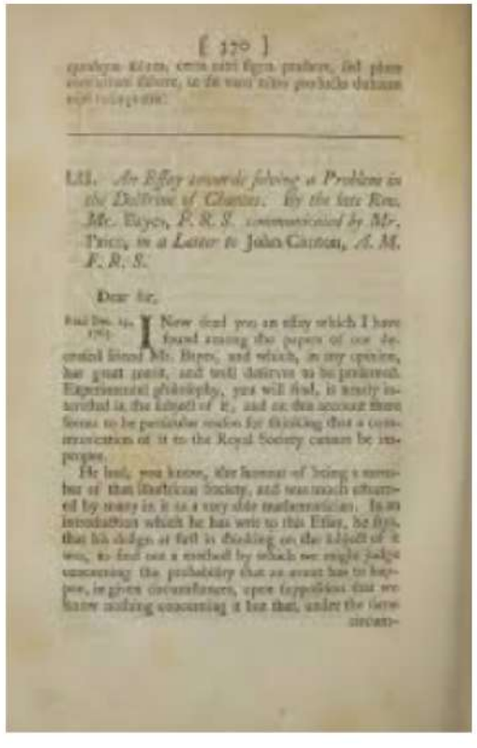
$$p(H|E) = \frac{p(E|H)p(H)}{p(E)}$$

Bayes Theorem

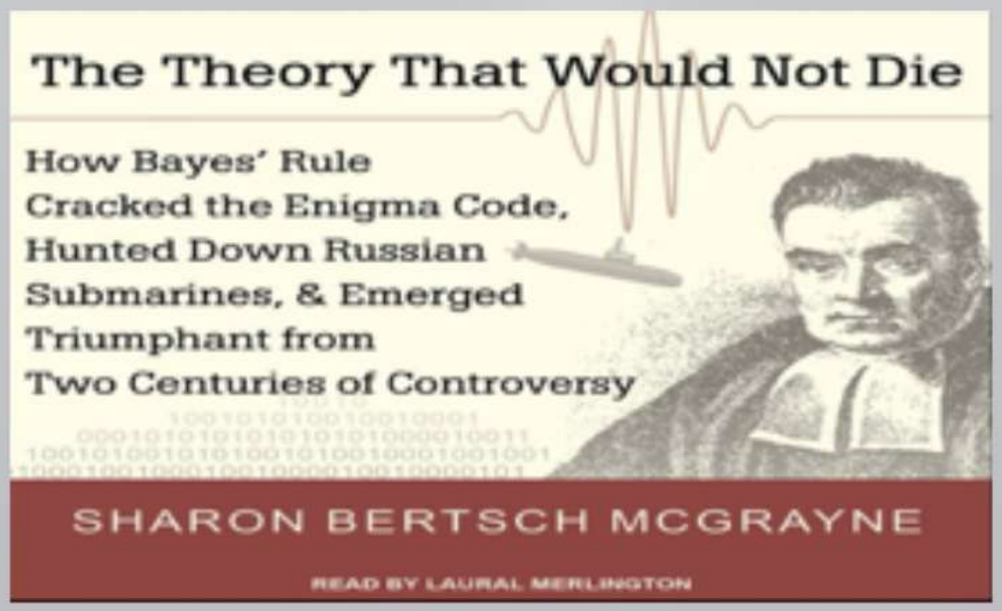
Publication on Dec. 23, 1763 of *"An Essay towards solving a Problem in the Doctrine of Chances"* by the late Rev. Mr. Bayes, communicated by Mr. Price in the *Philosophical Transactions of the Royal Society of London.*

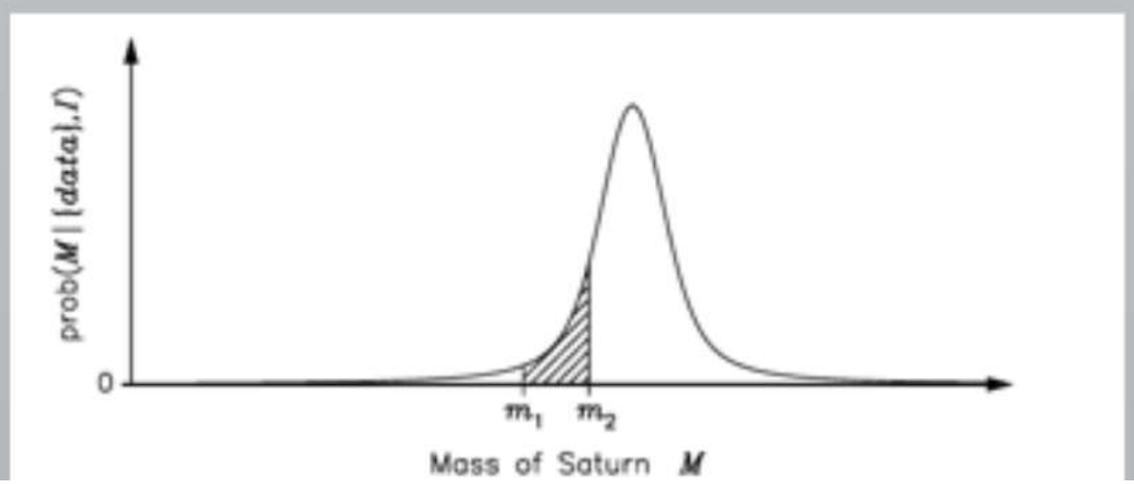




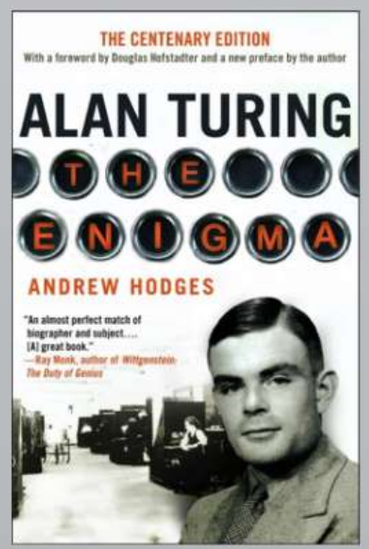


Bayes Theorem





- Laplace independently re-discovered Bayes' Theorem in 1774 as: the probability of a cause (given an event) is proportional to the probability of the event (given its cause).
- Alan Turing used it to decode the German Enigma cipher and arguably save the Allies from losing the Second World War
  - Banburismus : 'a highly intensive, Bayesian system' that allowed Turing and colleagues to guess a stretch of letters in an Enigma message



Bayes Theorem

- An insurance company believes that people can be divided into two classes — those that are accident prone and those that are not. Their statistics show that an accident-prone person will have an accident at some time within a fixed 1-year period with probability .4, whereas this probability decreases to .2 for a non-accident-prone person. If we assume that 30 percent of the population is accident prone, what is the probability that a new policy holder will have an accident within a year of purchasing a policy? Suppose that a new policy holder has an accident within a year of purchasing his policy. What is the probability that he is accident prone?

Bayes Theorem

- In a community of 100,000 people, 1,000 people will have cancer and 200 people will be 65 years old. Of the 1000 people with cancer, only 5 people will be 65 years old. Thus, of the 200 people who are 65 years old, how many is expected to have cancer?

- A laboratory blood test is 99 percent effective in detecting a certain disease when it is, in fact, present. However, the test also yields a “false positive” result for 1 percent of the healthy persons tested. (That is, if a healthy person is tested, then, with probability .01, the test result will imply he or she has the disease.) If .5 percent of the population actually has the disease, what is the probability a person has the disease given that his test result is positive?

Bayes theorem

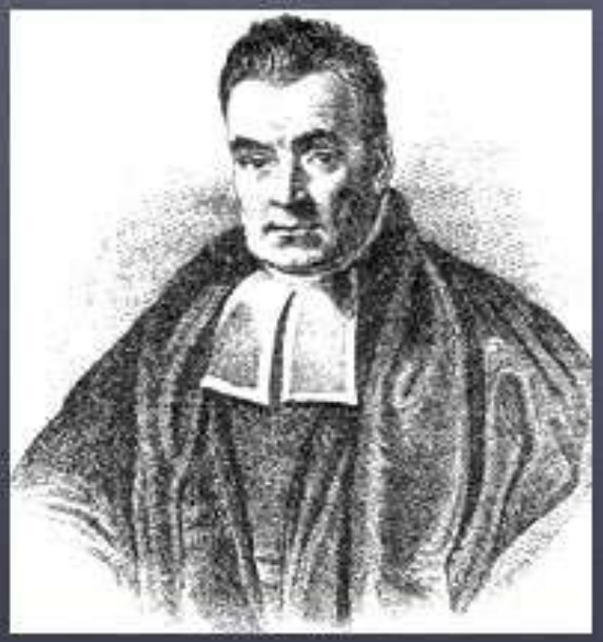
- Generative classifier, it specifies how to generate the data using the class conditional density p(x|y = c) and the class prior p(y = c).

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

- Discriminative classifier : directly fit the class posterior, p(y = c|x,theta)

Bayesian Modelling

Bayes' Theorem



Reverend Thomas Bayes
(1702-1761)

$$P(model|data, I) = P(model, I)\frac{P(data|model, I)}{P(data,I)}$$

Posterior Probability

represents the degree to which we believe a given **model** accurately describes the situation given the available **data** and all of our prior information **I**

Prior Probability

describes the degree to which we believe the model accurately describes reality based on all of our prior information.

Likelihood

describes how well the model predicts the data

Normalizing constant

Bayesian Modelling

$$P(\text{hypothesis}|\text{data}) = \frac{P(\text{data}|\text{hypothesis})P(\text{hypothesis})}{P(\text{data})}$$

Rev'd Thomas Bayes (1702–1761)

- Bayes rule tells us how to do inference about hypotheses from data.
- Learning and prediction can be seen as forms of inference.

Everything follows from two simple rules:

Sum rule: $P(x) = \sum_y P(x, y)$

Product rule: $P(x, y) = P(x)P(y|x)$

$$P(\theta|\mathcal{D}, m) = \frac{P(\mathcal{D}|\theta, m)P(\theta|m)}{P(\mathcal{D}|m)}$$

$P(\mathcal{D}\|\theta, m)$	likelihood of parameters θ in model m
$P(\theta\|m)$	prior probability of θ
$P(\theta\|\mathcal{D}, m)$	posterior of θ given data $\mathcal{D}$