Introduction to Machine Learning - CS725
Instructor: Prof. Ganesh Ramakrishnan

Overview of Probability Theory[1]

[1] Basic notes at https://www.cse.iitb.ac.in/~cs725/notes/classNotes/misc/BasicProbAndStats.pdf and advanced notes at https://www.cse.iitb.ac.in/~cs725/notes/classNotes/misc/CaseStudyWithProbabilisticModels.pdf

A review of probability theory

- Sample space(S): A sample space is defined as a set of all possible outcomes of an experiment. Example of an experiment would be a coin pair toss. In this case $S = \{HH, HT, TH, TT\}$.
- Event (E) : An event is defined as any subset of the sample space. Total number of distinct events possible is $2^{|S|}$, where $|S|$ is the number of elements in the sample space.
- Random variable (X) : A random variable is a mapping (or function) from set of events to a set of real numbers. Continuous random variable is defined thus

$$X : 2^S \rightarrow \mathbb{R}$$

On the other hand a discrete random variable maps events to a countable set (e.g. Natural Numbers)

$$X : 2^S \rightarrow \mathit{Countable}$$

Axioms of Probability

- For every event E, $0 \leq Pr(E) \leq 1$
- $Pr(S) = 1$
- If $E_1, E_2, \ldots, E_n$ is a set of pairwise disjoint events, then

$$Pr(\bigcup_{i=1}^{n} E_i) = \sum_{i=1}^{n} Pr(E_i)$$

Bayes' Theorem

Let $B_1, B_2, ..., B_n$ be a set of mutually exclusive events that together form the sample space S. Let A be any event from the same sample space, such that $P(A) > 0$. Then,

$$Pr(B_i/A) = \frac{Pr(B_i \cap A)}{Pr(B_1 \cap A) + Pr(B_2 \cap A) + \cdots + Pr(B_n \cap A)} \quad (1)$$

Using the relation $P(B_i \cap A) = P(B_i) \cdot P(A/B_i)$

$$Pr(B_i/A) = \frac{Pr(B_i) \cdot Pr(A/B_i)}{\sum_{j=1}^{n} Pr(B_j) \cdot Pr(A/B_j)} \quad (2)$$

Using Bayes' Theorem

A lab test is 99% effective in detecting a disease when in fact it is present. However, the test also yields a false positive for 0.5% of the healthy patients tested. If 1% of the population has that disease, then what is the probability that a person has the disease given that his/her test is positive?

Independent Events

Two events E_1 and E_2 are called independent iff their probabilities satisfy

$$P(E_1, E_2) = P(E_1) \cdot P(E_2) \tag{3}$$

where $P(E_1, E_2)$ means $P(E_1 \cap E_2)$
In general, events belonging to a set are called as mutually independent iff, for every finite subset, $E_1, \cdots, E_n$, of this set

$$Pr(\bigcap_{i=1}^{n} E_i) = \prod_{i=1}^{n} Pr(E_i) \tag{4}$$

Agenda

- Distribution function
  - Probability Density/Mass Function
  - Cumulative Distribution Function
- Statistical Measures
  - Expectation
  - Variance
  - Covariance
- Random Variables
  - Bernoulli Random Variable
  - Binomial Random Variable
  - Normal Random Variable
- Central Limit Theorem

Uncertainty

- We are trying to build systems that understand and (possibly) interact with the real world
- We often can not prove something is true, but we can still ask how likely different outcomes are or ask for the most likely explanation

Probability theory is nothing but common sense reduced to calculation. — **Pierre Laplace, 1812**

We will restrict ourselves to a relatively informal discussion of probability theory

Notations

- A random variable **X** represents the outcome or the state of the world
- We will write Pr(X) to mean probability of event X, **Probability(X=x)**
- **Sample space:** the space of all possible outcomes (may be discrete, continuous or mixed)
- p(x) is the probability mass (density) function
  - Assigns a number to each point in sample space
  - Non-negative, sums (integrates) to 1
  - Intuitively: how often does x occur, how much do we believe in x.

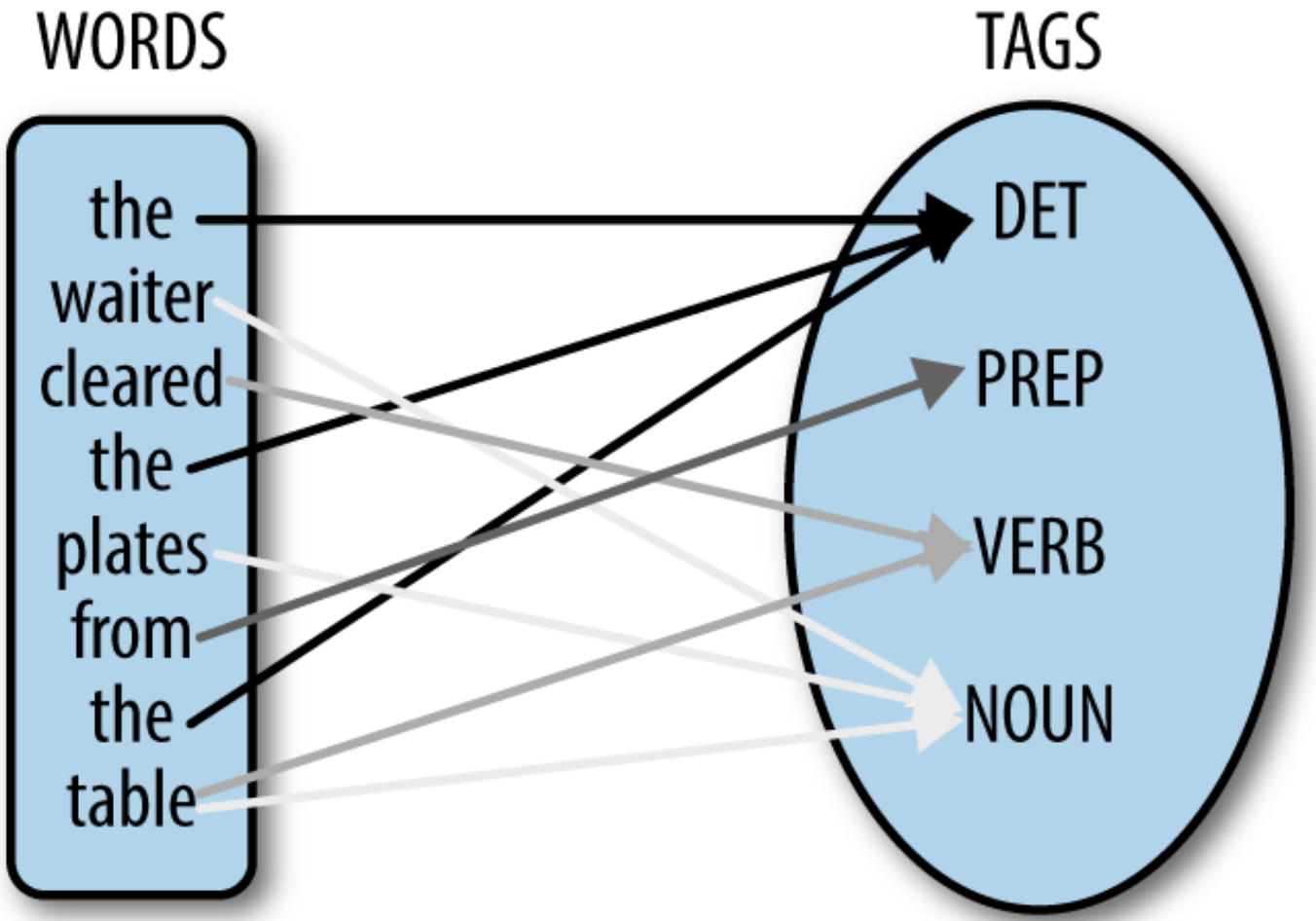
- *Pr* - Probability of an event in general
- *F* - Cumulative distribution function
- *p* - Probability distribution function (pdf) or probability mass function (pmf)
- *pdf* - pdf occurs in case of continuous random variable
- *pmf* - pmf occurs in case of discrete random variable

Example - Part of Speech

POS tagging is a problem of great importance in the field of Natural Language Processing, **NLP**

Input: A set of n-words

Output: POS tag for each word



Assuming the picking of words is done independently, find probability that the set contains a 'noun' given that it contains a 'verb'.

Solution:

- Probability that a word is of part of speech type 'k' is pk
- Let A_k be the probability that the set contains pos type 'k'

$Pr(A_k) = 1 - (1 - p_k)^n$
where $(1 - p_k)^n$ is that all 'n' words are not of pos of type 'k'.

$$Pr(A_{noun}/A_{verb}) = \frac{Pr(A_{noun} \bigcap A_{verb})}{Pr(A_{verb})}$$

$$Pr(A_{k1} \bigcap A_{k2}) = 1 - (1 - p_{k1})^n - (1 - p_{k2})^n + (1 - p_{k1} - p_{k2})^n$$

$$Pr(A_{noun}/A_{verb}) = \frac{1-(1-p_{noun})^n-(1-p_{verb})^n+(1-p_{noun}-p_{verb})^n}{1-(1-p_{verb})^n}$$

Distribution Functions

- **pmf**: It is a function that gives the probability that a discrete random variable is exactly equal to some value (Src: wiki)

$$p_X(a) = Pr(X = a)$$

- **pdf**: A probability density function of a continuous random variable is a function that describes the relative likelihood for this random variable to occur at a given point in the observation space (Src: Wiki)

$$Pr(X \in D) = \int_D p(x)dx$$

Cumulative Distribution Function

Case: Discrete Random Variable

$$F(a) = Pr(X <= a)$$

Case: Continuous Random Variable

$$F(a) = Pr(X <= a) = \int_{-\infty}^{a} p(x)dx$$

Note: pdf for continuous distribution can be obtained by differentiating the cdf of that random variable:

$$f(a) = \frac{dF(x)}{dx}|_{x=a}$$

Joint Distribution Function

- If p(x,y) is a joint pdf i.e. for continuous case:

$$F(a,b) = Pr(X <= a, Y <= b) = \int_{-\infty}^{b} \int_{-\infty}^{a} p(x,y)dxdy$$

$$p(a,b) = \frac{\partial^2 F(x,y)}{\partial x \partial y}|_{a,b}$$

- For discrete case i.e. p(x,y) is a joint pmf:

$$F(a,b) = \sum_{x<=a} \sum_{y<=b} p(x,y)$$

Marginalization

- Marginal probability is the unconditional probability P(A) of the event A; that is, the probability of A, regardless of whether event B did or did not occur.
- If B can be thought of as the event of a random variable X having a given outcome, the marginal probability of A can be obtained by summing (or integrating, more generally) the joint probabilities over all outcomes for X.
- For example, if there are two possible outcomes for X with corresponding events B and B', this means that

$$P(A) = P(A \bigcap B) + P(A \bigcap B')$$

Discrete case: $P(X = a) = \sum_y p(a, y)$

Continuous case: $P_x(a) = \int_{-\infty}^{\infty} p(a, y)dy$

Example

Let X and Y are *independent continuous* random variables with same density functions

$$p(x) = \begin{cases} e^{-x} & \text{if } x > 0; \\ 0 & \text{otherwise.} \end{cases}$$

Find density $\frac{X}{Y}$.

Solution

$$F_{\frac{X}{Y}}(a) = Pr(\frac{X}{Y} <= a)$$
$$= \int_0^\infty \int_0^{ya} p(x,y)dxdy$$
$$= \int_0^\infty \int_0^{ya} e^{-x}e^{-y}dxdy$$
$$= 1 - \frac{1}{a+1}$$
$$= \frac{a}{a+1}$$

$$f_{\frac{X}{Y}}(a) = \text{derivative of } F_{\frac{X}{Y}}(a) \text{ w.r.t a}$$
$$= \frac{1}{(a+1)^2} > 0$$

Conditional Density

Suppose X and Y are two random variable then we can define the conditional probability density of X given Y, denoted as $X|Y$

Discrete Case

$$p_X(\frac{x}{Y=y}) = P(\frac{X=x}{Y=y}) = \frac{P(X=x, Y=y)}{P(Y=y)}$$

Continuous case

$$p_X(\frac{x}{Y=y}) = \frac{p_{X,Y}(\frac{X}{Y})}{p_Y(y)} = \frac{p_{X,Y}(\frac{X}{Y})}{\int_{-\infty}^{\infty} p(x,y)dx}$$

Joint Probability Distribution

- $Prob(X = x, Y = y)$
  - "Probability of X=x and Y=y"
  - $p(x, y)$

Conditional Probability Distribution

- $Prob(X = x|Y = y)$
  - "Probability of X=x given Y=y"
  - $p(x|y) = p(x, y)/p(y)$

Rules of Probability

- Sum Rule (marginalization/ summing out)

$$p(x) = \Sigma_y p(x, y)$$
$$p(x_1) = \Sigma_{x_2} \Sigma_{x_3} ... \Sigma_{x_n} p(x_1, x_2, ..., x_n)$$

- Product/Chain Rule

$$p(x, y) = p(y, x)p(x)$$
$$p(x_1, x_2, ..., x_n) = p(x_1)p(x_2|x_1)...p(x_n|x_1, x_2..., x_n)$$

Bayes Rule

- One of the most important formulas in probability theory

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} = \frac{p(y,x)p(x)}{\Sigma_x p(y|x)p(x)}$$

- This gives away a way of reversing conditional probabilities

Independence of Random Variables

- Two random variables are said to be **independent** iff their joint distribution factors

$$X \perp Y \iff p(x,y) = p(y|x)p(x) = p(x|y)p(y) = p(x)p(y)$$

- Two random variables are **conditionally independent** iff given a third they are independent after conditioning on the third variable

$$X \perp Y|Z \iff p(x,y|z) = p(y|x,z)p(x|z) = p(x|y,z)p(y|z) = p(x|z)p(y|z) \forall Z$$

Expectation

• **Discrete case**: Expectation is equivalent to probability weighted sums of possible values. If X is a discrete random variable

$$E(X) = \Sigma_i x_i Pr(x_i)$$

If the random variable is a function of x, then

$$E(X) = \Sigma_i f(x_i) Pr(x_i)$$

• **Continuous case**: Expectation is equivalent to probability density weighted integral of possible values.

$$E(X) = \int_{-\infty}^{\infty} xp(x)dx$$

If the random variable is a function of x, then

$$E(X) = \int_{-\infty}^{\infty} f(x)p(x)dx$$

Properties of Expectation

1. $E[X + Y] = E[X] + E[Y]$ Proof HW

2. $E[(X - c)^2] \geq E[(X - \mu)^2]$ Proof HW
   where $\mu = E[X]$
   For any constant c and any random variable X

3. $E[cX] = cE[X]$ Proof HW

Variance

For any random variable X, **variance** is defined as follows:

$$Var[X] = E[(X - \mu)^2]$$

$$\Rightarrow Var[X] = E[X^2] - 2\mu E[X] + \mu^2$$
$$\Rightarrow Var[X] = E[X^2] - (E[X])^2$$

$$Var[\alpha X + \beta] = \alpha^2 Var[X]$$

Covariance

For random variables X and Y, **covariance** is defined as:

$$Cov[X, Y] = E[(X - E(X))(Y - E(Y))] = E[XY] - E[X]E[Y]$$

If X and Y are independent then their covariance is 0, since in that case

$$E[XY] = E[X]E[Y]$$

Note: However, covariance being 0 does not necessarily imply that the variables are independent.

Properties:

1. $Cov[X + Z, Y] = Cov[X, Y] + Cov[Z, Y]$
2. $Cov[\Sigma_i X_i, Y] = \Sigma_i Cov[X_i, Y]$
3. $Cov[X, X] = Var[X]$

Chebyshev's Inequality

Chebyshev's inequality states that if X is any random variable with mean μ and variance σ then $\forall k > 0$

$$Pr[|X - \mu| \geq k] \leq \frac{\sigma^2}{k^2}$$

Implications:

• If n tends to infinity, then the data mean tends to converge to μ, giving rise to the *weak law of large numbers*.

• If X_i are independent and identically distributed random variables,

$Pr[|\frac{X_1+X_2+..+X_n}{n} - \mu| \geq k]$ tends to 0 as n tends to ∞

Important Random Variables

Bernoulli Random Variable: It is a *discrete* random variable taking values 0,1

Say, $Pr[X_i = 0] = 1 - q$ where $q \epsilon [0, 1]$
Then $Pr[X_i = 1] = q$

- $E[X] = (1 - q) * 0 + q * 1 = q$
- $Var[X] = q - q^2 = q(1 - q)$

Note: It represents the probability of success in a random event. For example: Coin toss experiment can be modeled as a *Bernoulli random variable* with $Pr[Head] = Pr[X_i = 1] = q$

Binomial Random Variable It is a *discrete* variable where the distribution is of number of 1's in a series of n experiments with $\{0,1\}$ value, with the probability that the outcome of a particular experiment is 1 being q.

A binomial distribution is the distribution of n-times repeated bernoulli trials.

1. $Pr[X = k] = \binom{n}{k} q^k (1-q)^{n-k}$
2. $E[X] = \Sigma_i E[Y_i]$ where Y_i is a bernoulli random variable

$$E[X] = nq$$

3. $Var[X] = \Sigma_i Var[Y_i]$ (since Y_i's are independent)

$$Var[X] = nq(1-q)$$

Example:
An example of Binomial distribution is the distribution of number of heads when a coin is tossed n times.

Normal (Guassian) Distribution

- It is a continuous distribution

$$p(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}$$

  - μ is the mean
  - σ^2 is the variance

- Exercise: Verify there mean and variance. For e.g.

$$E(X) = \mu = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\} dx$$

- Multivariate Guassian

$$p(x|\mu,\Sigma) = |2\pi\Sigma|^{-1/2} \exp\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\}$$

  - x is now a vector
  - μ is the mean vector
  - Σ is the co-variance matrix

Properties of Normal Distribution

- All marginals of a Gaussian are again Gaussian
- Any conditional of a Gaussian is Gaussian
- The product of two Gaussians is again Gaussian
- Even the sum of two independent Gaussian RVs is a Gaussian

Note: Many of the standard distributions belong to the family of **exponential distributions**

- Bernoulli, binomial/multinomial, Poisson, Normal (Gaussian), beta/Dirichlet ...
- Share many important properties - e.g. They have a conjugate prior. (We will discuss this in next lecture)

Central Limit Theorem

If $X_1, X_2, .., X_m$ is a sequence of i.i.d. random variables each having mean μ and variance σ^2
Then for large m, $X_1 + X_2 + .. + X_m$ is approximately normally distributed with mean mμ and variance mσ^2

If $X \sim N(\mu, \sigma^2)$

Then $P[x] = \frac{1}{\sigma \sqrt[2]{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$

It can be shown by CLT

- $\frac{X_1+X_2+..+X_n-n\mu}{\sigma \sqrt[2]{n}} \sim N(0, 1)$
- Sample Mean: $\hat{\mu} \sim N(\mu, \frac{\sigma^2}{m})$