Friday, April 9, 2021

ECS-132

Notes By: Shuxian Zhang

Lecture 1~2 4

Probability & Random Variables 4

The basics of counting 5

Sampling Ways 5

Lecture 3 6

Probability 6

Inclusion-Exclusion Principle 6

Lecture 4 7

Conditional Probability 7

The Law of Total Probability 7

Independence of Events 7

Lecture 5 8

Discrete Random Variables 8

PMF 8

CDF 8

Lecture 6 8

Review independent, non-independent vs. disjoint events 8

Expectation and Variance 9

Bernoulli Distribution 9

Binomial Distribution 10

Geometric Distribution 10

Lecture 7 11

Poisson Process and the Poisson distribution 11

Lecture 8 12

Poisson distribution’s relationship with other distributions 12

Lecture 9 12

Indicator Random Variable 12

Expectation 13

Lecture 11 14

boxplots, median, quantiles, quartiles 14

Lecture 12 14

Continuous Random Variables 14

PDF (Probability Density function) 14

CDF (Cumulative Distribution Function) 15

Mean and Variance 15

Lecture 13 15

Uniform Distribution 15

Exponential Distribution 16

Normal Distribution 17

*Power-law Distribution 17

*Entropy 18

Lecture 14 18

Binary Classification 18

Prevalence, Sensitivity, and Specificity 18

Confusion matrix 19

How to determine the quality of the classifier? 19

Lecture 15 20

Stochastic Process 20

Markov Chain 20

2-state discrete time Markov chains 21

Lecture 18 22

Bounding Probability 22

Sample Mean and Sample Variance 23

Error bars 23

Central Limit Theorem 23

Lecture 19 24

Maximum Likelihood Estimation 24

Linear Model - a model of Linear Correlation 24

Measure of Correlation 24

Measure of Covariance 25

The best-fit line 25

Appendix - R 26

Vector 26

Useful function 26

Matrix 26

Distributions 27

Linear Regression 27

Lecture 1~2

Probability & Random Variables

- **Non-determinism**: While we know the sample space, i.e., the set of all possible outcomes, for a specific experiment we do not know what will be the outcome. When we measure the protein level in the blood, we know the range within which the measured value could lie but we do not know what will be specific value for a given blood sample.
- One **Random Variable** (R.V.): an observation of a stochastic process, i.e. an outcome
  - Roll of a dice
  - Poker hand
- The R.V. can be discrete (e.g. an integer) or continuous (e.g. an real number)
- The **event space / sample space**: the set of all possible values the R.V. can assume
  - Finite sample space: $S_{\text{dice}} = \{1,2,3,4,5,6\}$, discrete
  - Countable infinite sample space: $S_{\text{The Drowsy Programmer}} = \{C, EC, EEC\ldots\}$ (E means error, C means success, the programmer will keep compiling until C)
  - Uncountable infinite sample space: $S_{\text{velocity}} = \{-\infty < v < +\infty\}$, continuous
- An **event** E is a subset of outcomes $E \subset S$ to which we assign a probability $P(E)$
- **The axioms of probability**
  - Possibility: $P(E) \geq 0$
  - Normalization: $P(S) = 1$ (The R.V. must have an outcome in S)
  - Additivity: $P(B \text{ or } C) = P(B) + P(C)$, if B and C are mutually exclusive events.
  - Complementarity: $P(\bar{E}) = 1 - P(E)$, where $\bar{E}$ or E^c is the event not E.
- Naive Definition of **Probability**: $\hat{P}(E) = \dfrac{\text{\# of times event E is observed}}{\text{Total \# of trials}}$
  - When #trail >> 1, $\hat{P}(E) = P(E)$
  - Only applicable for
    - Finite sample space and

- Equally likely outcomes

The basics of counting

- **Permutation** (How many ways we can order the objects): $P(N, r) = \dfrac{N!}{(N-r)!}$, also use notation ${}^{N}P_{r}$
  - If repetition is allowed, $P^{\text{rep}}(N, r) = N^r$
- **Combination** (Order doesn't matter): $C(N, r) = \dfrac{N!}{(N-r)!r!}$, also use notation ${}^{N}C_{r}$, $\binom{N}{r}$
- **Product rule**: when combining independent events, the number of possibilities multiply.
  - Choose r1 from set N1, and r2 from set N2. # of possibilities=$P(N_1, r_1) * P(N_2, r_2)$
- **Sum rule**: when combining mutually exclusive events, the number of possibilities add together.
  - Choose r1 from set N1 or r2 from set N2. # of possibilities=$P(N_1, r_1) + P(N_2, r_2)$
- **Multinomial Coefficients**
  - A set of n distinct objects is divided into r distinct groups of size $n_1, n_2, \cdots, n_r$, such that $\sum_{i=1}^{r} n_i = n$. The total number of ways is given by

$$T = \underbrace{\binom{n}{n_1}}_{\text{no. ways of choosing 1st group}} \underbrace{\binom{n-n_1}{n_2}}_{\text{no. ways of choosing 2nd group}} \cdots \underbrace{\binom{n-n_1-\ldots-n_{r-1}}{n_r}}_{\text{no. ways of choosing rth group}} = \frac{n!}{n_1!n_2!\ldots n_r!}$$

Sampling Ways

- Suppose we have n distinguishable objects and we want to find out the number of ways we can pick k objects out of n.

	Order matters	**Order doesn't matter**
With replacement	n^k	$\binom{k+n-1}{k}$
Without replacement	$n(n-1)\ldots(n-k+1) = \dfrac{n!}{(n-k)!}$	$\binom{n}{k}$

- Simulation
  - Estimate probabilities from a simulation (the frequency perspective)

Friday, April 9, 2021

Lecture 3

Probability

- Formal Definition: Consider the probability space consisting of P and S where S is the sample space and P is a function that takes as an input an event A and returns $P(A) \in [0,1]$ such that
  - Axiom1: $P(\Phi) = 0$
  - Axiom2: $P(S) = 1$
  - Axiom3: If $A_1, A_2, \ldots,$ are disjoint non-overlapping events then $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$
- Properties
  - Properties1: If $A^{\complement}$ denotes the complement of event A, then $P(A^{\complement}) = 1 - P(A)$
  - Properties2: If event A is contained in event B, i.e. $A \subseteq B$, then $P(A) \leq P(B)$
  - Properties3: Consider two events A, B, then $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- Set theory (*deMorgan's Law*)
  - $\overline{(A \cup B)} = \overline{A} \cap \overline{B}$
  - $\overline{(A \cap B)} = \overline{A} \cup \overline{B}$

Inclusion-Exclusion Principle

- Now we consider the general case. If $A_1, A_2, \ldots, A_n$ are n events, then

$$\begin{aligned}P(A_1 \cup A_2 \cup A_3 \cup \ldots \cup A_n) = &\sum_{i=1}^{i=n} P(A_i) \\ &- \sum_{i_1<i_2} P(A_{i_1}A_{i_2}) \\ &+(-1)^{r+1} \sum_{i_1<i_2<\ldots<i_r} P(A_{i_1}A_{i_2}\ldots A_{i_r}) + \ldots \qquad r = 3,\ldots,n-1 \\ &+(-1)^{n+1}P(A_1A_2A_3\ldots A_n)\end{aligned}$$

- What the above equation states is that the probability of the union of n events is equal to the sum of the probabilities of the of events taken one at a time, minus the probability of the intersection of events taken two at a time, plus the probability of the events taken three at a time, and so on until the probability of the intersection of events all taken together.

Friday, April 9, 2021

- In exam, $n \leq 4$,

$$\begin{aligned}P(A_1 \cup A_2 \cup A_3 \cup A_4) = {} & P(A_1) + P(A_2) + P(A_3) + P(A_4) \\ & -P(A_1 \cap A_2) - P(A_1 \cap A_3) - P(A_1 \cap A_4) - P(A_2 \cap A_3) - P(A_2 \cap A_4) - P(A_3 \cap A_4) \\ & +P(A_1 \cap A_2 \cap A_3) + P(A_1 \cap A_2 \cap A_4) + P(A_1 \cap A_3 \cap A_4) + P(A_2 \cap A_3 \cap A_4) \\ & -P(A_1 \cap A_2 \cap A_3 \cap A_4)\end{aligned}$$

Lecture 4

Conditional Probability

- Definition: $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)} \qquad P(B) > 0$
- Theorems
  - Theorems1: $P(A \cap B) = P(B)P(A \mid B) = P(A)P(B \mid A)$, note that $P(A \mid B) \neq P(B \mid A)$
  - **Multiplication rule**: $P(A \cap B \cap C) = P(A) \times P(B \mid A) \times P(C \mid A \cap B)$
  - **Bayes' Rule**: $P(A \mid B) = \dfrac{P(A) \times P(B \mid A)}{P(B)}$
    - $P(A \mid B)$ is called the posterior probability of A
    - $P(A)$ is called the prior probability
  - **The Law of Total Probability**: if $A_1, A_2, \ldots, A_n$ are mutually exclusive and exhaustive events, then $P(B) = P(B \mid A_1) \times P(A_1) + P(B \mid A_2) \times P(A_2) + \ldots + P(B \mid A_n) \times P(A_n)$
  - Normalization: $P(A \mid B) = 1 - P(A^c \mid B)$

The Law of Total Probability

$$\begin{aligned}P(B) &= P(B \cap A_1) \cup P(B \cap A_2) \cup \ldots \cup P(B \cap A_n) \\ &= P(B \cap A_1) + P(B \cap A_2) + \ldots + P(B \cap A_n) \\ &= P(B \mid A_1) \times P(A_1) + P(B \mid A_2) \times P(A_2) + \ldots + P(B \mid A_n) \times P(A_n)\end{aligned}$$

Independence of Events

We consider two events A and B are independent if $P(A \cap B) = P(A) \times P(B)$

Friday, April 9, 2021

Lecture 5

Discrete Random Variables

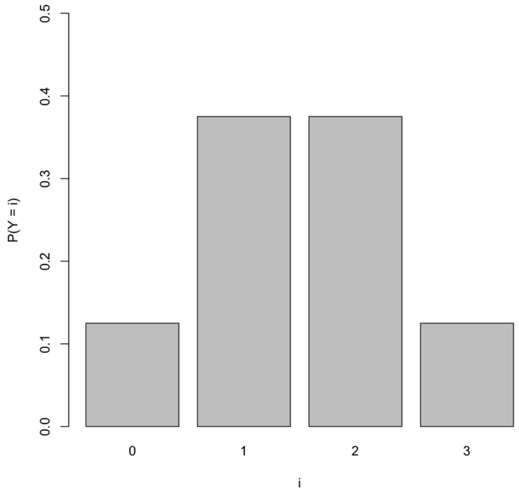
- A **random variable** is a function mapping the sample space S to the set of real numbers $\mathbb{R}$

PMF

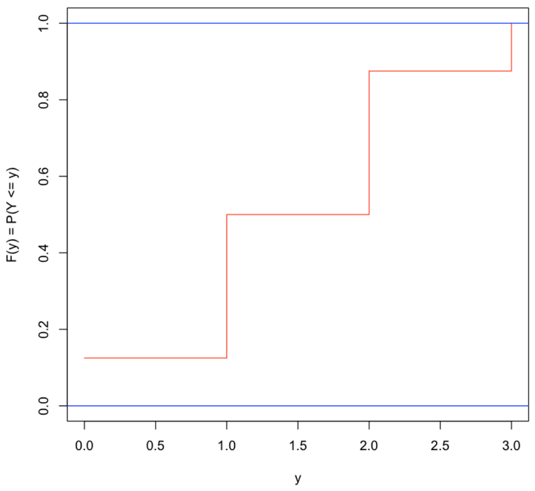
- Probability Mass Function, $P_j = P(X = X_j)$, $\sum_j P_j = 1$

CDF

- Cumulative Distribution Function $F_X(x) = P(X \leq x)$
  - Properties: $P(a < X \leq b) = F_X(b) - F_X(a)$



PMF



CDF

Lecture 6

Review independent, non-independent vs. disjoint events

- $P(A \cap B)$
  - $P(A \cap B) = P(A) \times P(B)$ if independent
  - $P(A \cap B) = P(A) \times P(B \mid A) = P(B) \times P(A \mid B)$ if non-independent (Multiplication Rule)
  - $P(A \cap B) = 0$, $P(A \mid B) = 0$, $P(B \mid A) = 0$) if disjoint / non-overlapping / mutually exclusive

- $P(A \cup B)$
  - $P(A \cup B) = P(A) + P(B)$ if disjoint
  - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ if non-disjoint (Inclusion-Exclusion Principle)
- Independent means they are not disjoint
- Disjoint means extremely dependent/mutually exclusive ($P(A \mid B) = 0$)

Expectation and Variance

- **Expectation**
  - Def: If X is a discrete random variable and takes values $x_1, x_2, x_3, \dots, x_n$ then expected value of X denoted by E(X) is given by: $E(X) = \sum_{i=1}^{n} p_i x_i$, where $p_i = P(X = x_i)$
    - This is also called the mean or the average or the first moment.
    - This would also apply to a countably infinite sample space in which case n=∞
- **Variance**
  - Def: If X is a discrete random variable and takes values $x_1, x_2, x_3, \dots, x_n$ then the variance of X denoted by Var[X] is given by:
    $Var(X) = E(X^2) - (E(X))^2 = \sum_{i=1}^{n} p_i x_i^2 - (\sum_{i=1}^{n} p_i x_i)^2$
    - $E(X^2)$ is called the second moment
    - The standard deviation typically denoted by σ is the square root of the variance, i.e, $\sigma = \sqrt{Var(X)}$
    - The standard deviation measures how the probability mass is distributed around the mean. The large the value of σ the wider the spread.
    - Equivalent to $E((X - E(X))^2)$

Bernoulli Distribution

- Def: A random variable X is said to have a Bernoulli distribution if X has two possible values 1 (success) and 0 (failure) where $P(X$=$1)$=p and $P(X$=$0)$=q=$1-p$.
- Notion: $X \sim \text{Bern}(p)$
- Expectation: $E(X) = p$
- Variance: $Var(X) = p - p^2 = p(1-p)$

Binomial Distribution

- Def: The number of successes X in n independent Bern(p) trials where p is the probability of success. Its distribution is given by
  - $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \qquad 0 \le k \le n$
- Notion: $X \sim \text{Binom}(n, p)$
- Expectation: $E(X) = np$
- Variance: $Var(X) = np(1-p)$
- Finite space, order doesn't matter.

Geometric Distribution

- Def: Consider a sequence of independent Bern(p) trials from two different but equivalent perspectives.
  - Let the random variable X denote **the number of trial until first success**.
    - Notation: X follows a geometric distribution: $X \sim \text{Geom}(p)$
    - Here, we want to know the number of trials **up to and including** the first success. The pmf is given by $P(X = k) = (1-p)^{k-1}p \qquad k \in \{1,2,3,\ldots,\infty\}$
    - Expectation: $E(X) = \dfrac{1}{p}$
    - Variance: $Var(X) = \dfrac{1-p}{p^2}$
  - Let the random variable Y denote **the number of failures before the first success**
    - Notation: Y follows a geometric distribution: $Y \sim \text{Geom}^*(p)$
    - The pmf is given by $P(Y = k) = (1-p)^k p \qquad k \in \{0,1,2,3,\ldots,\infty\}$
    - Expectation: $E(Y) = \dfrac{1-p}{p}$
    - Variance: $Var(X) = \dfrac{1-p}{p^2}$
- The Geometric distribution has only one true parameter: p

- The values that the random variable takes is countably infinite, i.e., the set of positive integers.
- Show that the above pmf is valid. This is done by showing two properties. First, we need to show that $P(X{=}i)\geq 0$ for all $i{=}0,1,2,\ldots$ Second, we need to show that $\sum_{i=0}^{\infty} P(X=i)=1$. These are easily shown.
- This distribution has an interesting property called the **memoryless property**. More about this after when we study the Poisson random variable.

Lecture 7

Poisson Process and the Poisson distribution

- Poisson Process: Consider that events occur in time. Let $N(t)$ be a counting process that counts the number of events that occur in some interval of time $t\geq 0$.
- λ given as events per unit time (i.e., a rate)
  - $N(t)$ is a **Poisson process** with **rate** λ if the following properties are satisfied:
    - ▹ $N(0)=0$
    - ▹ The number of events that occur in disjoint time intervals are independent;
    - ▹ The distribution of the number of events in a given interval depends only on the length of the interval and not on the location;
    - ▹ For small values of h, $P(N(h)=1)=\lambda h+o(h)$. Where $o(h)$ is any function $f(h)$ for which $\lim_{h\to 0}\frac{f(h)}{h}=0$
    - ▹ $P(N(h)=2)=o(h)$
  - Notation: $N(t)\sim \text{Pois}(\lambda)$
  - PMF: $P(N(t)=k)=\dfrac{e^{-\lambda t}(\lambda t)^k}{k!} \quad k\in\{0,1,2,\ldots\}$
- λ given as the expected number (integer) in the time window of interest (**official**)
  - Let X denote the random variable corresponding to how many events occur in that given time window.
  - Notation: $X\sim \text{Pois}(\lambda)$
  - $P(X=k)=\dfrac{e^{-\lambda}(\lambda)^k}{k!} \quad k\in\{0,1,2,\ldots\}$
- Countable infinite

- Expectation & Variance: λ
  - Stem from the fact that $e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$ by the Taylor Series Expansion of the Exponential
- Superposition of Poisson Process
  - $X_1 \sim \text{Pois}(\lambda_1)$, $X_2 \sim \text{Pois}(\lambda_2)$; then $X_1 + X_2 \sim \text{Pois}(\lambda_1 + \lambda_2)$

Lecture 8

Poisson distribution's relationship with other distributions

- With Binomial distribution
  - Let $A_1, A_2, \ldots, A_n$ denote events with $P(A_j) = p_j$. If n is large and the p_j's are small and if the events are independent or "weakly dependent," then the number of events that occur is approximately Pois(λ) where $\lambda = \sum_{j=1}^{n} p_j$
  - If the events are independent and all the probabilities are equal, meaning $p_j = p \, \forall j$, then we have a sequence of Bernoulli trails with large n and small p. The number of "successful" events out of the n total events is a random variable X~**Binom**(n,p). Furthermore, when $n \rightarrow \infty$ and $p \rightarrow 0$ such that np is constant, then X can be well approximated by Pois(λ) where $\lambda = np$.
- With Exponential distribution
  - Consider events that follow the Poisson process with parameter λ. Let Y be a random variable that denotes the waiting time between events; i.e., starting from an event what is the time until the next event. This is also called the **inter-arrival time**.

$$\begin{aligned} P(Y \leq t) &= 1 - P(Y > t) \\ &= 1 - \text{Probability there no arrivals in time t} \\ &= 1 - P(N(t) = 0) \\ &= 1 - e^{-\lambda t} \end{aligned}$$

  - which means that Y~Expo(λ).

Lecture 9

Indicator Random Variable

- Def: Given any event A in the sample space S with P(A) denoting the probability of event A. We define indicator random variable as follows, $X = \begin{cases} 1 & \text{if A occurs} \\ 0 & \text{otherwise} \end{cases}$

- $P(A = 1) = P(A), P(A = 0) = P(A^c) = 1 - P(A)$
- The Bernoulli random variable is an Indicator random variable with success being defined as event A occurring.

Expectation

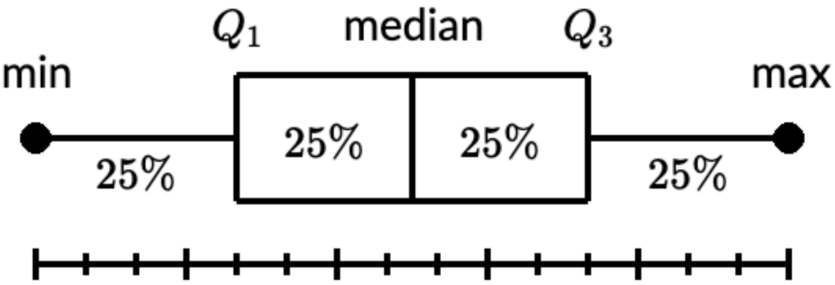
- Def: The mean/expectation/average value of a random variable X is denoted as $E(X)$. It is simply the result of multiplying each possible value of X by its probability and summing.
  - $E(X) = \sum_{x \in X} x \times P(X = x)$
  - This is called the weighted average of the different values the random variable takes, where the weights are the probabilities.
  - In the summation we need only consider non-zero values that the random variable takes.
  - The mean value need not be a value that the random variable takes.
- Properties
  - Expectation of a Function of a Random Variable: $E(g(X)) = \sum_{x \in X} g(x)P(X = x)$
  - Scaling and translation: suppose $g(X) = a + bX$, then $E(g(X)) = a + bE(X)$
  - Linearity: If $X_1, X_2, \dots, X_n$ are random variables (independent or dependent), then $E(X_1 + X_2 + \dots + X_n) = E(X_1) + E(X_2) + \dots + E(X_n)$
- Moments and Variance
  - First moment: $E(X)$
  - Second moment: $E(X^2)$
  - Variance: $Var(X) = E((X - E(X))^2) = E(X^2) - E(X)^2$
- Standard Deviation
  - The standard deviation of a random variable (denoted σ) is a measure of how "spread out" a function is, or the typical deviation from the mean.
  - $\sigma(X) = \sqrt{Var(X)}$
  - $\sigma(aX) = |a|\sigma(X)$

Friday, April 9, 2021

Lecture 11

boxplots, median, quantiles, quartiles

- Median: The median is the data value such that 50% of the data values are less than it (consequently, 50% of the data values are greater than that value).
- Quantile: The 25th quantile (called the "lower quartile") and denoted Q_{25} is the data value such that 25% of the data is below that value (and hence 75%) of the data values are above it.
- Quartile: if the quantiles are in increments of 25% they are called quartiles.
- Boxplot:
    - Interquartile range: $IQR = Q_{75} - Q_{25}$
    - The whiskers extend only to the most extreme data point which is no more than range×IQR from the box (out of range data will be marked as outliers)



Lecture 12

Continuous Random Variables

- Def: continuous random variables are functions that map the outcomes of a sample space to the $\mathbb{R}$.
    - Difference with discrete RV: the sample space now is uncountably infinite and they map to real values
    - Let X denotes the random variable and takes value $x \in \mathbb{R}$. Sample space: $S = \{-\infty < x < \infty\}$

PDF (Probability Density function)

- If X is a random variable, $f_X(x)$ will denote its density function.

- $f_X(x)$ is a function of x for which the area under the curve $f_X(x)$ between x and $x+\Delta x$ gives the probability that X lies between x and $x+\Delta x$.
- $P(x \leq Y \leq x + dx) = f_Y(x)dx$x
- Properties
  - Property1: $f_X(x) \geq 0$
  - Property2: $\int_{-\infty}^{\infty} f_X(x)dx = 1$
  - Quantifying $P(X = a)$: $P(X = a) = \int_{a}^{a} f(x)dx = 0$
  - This is because, even in an arbitrary small neighborhood of x there are infinite number of values so the probability that X exactly takes the value x will be 0. (Otherwise their sum will exceed 1)

CDF (Cumulative Distribution Function)

- $F_Y(k) = P(Y \leq k) = \int_{-\infty}^{k} f_Y(x)dx$
- $P(a < X \leq b) = F_X(b) - F_X(a) = \int_{b}^{a} f_X(x)dx$
  - In continuous random variables < (or >) and ≤ (or ≥) are the same since the probability that the random variable takes a specific values is equal to 0.

Mean and Variance

- Mean / Expectation: $E(x) = \int_{-\infty}^{\infty} x f_X(x)dx$
- Variance: $Var(x) = \int_{-\infty}^{\infty} (x - E(X))^2 f_X(x)dx = E(X^2) - (E(X))^2$

Lecture 13

Uniform Distribution

- Def: A random variable X follows a Uniform Distribution, $X \sim \text{Unif}(a, b)$.

- **PDF**: $f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$
- **CDF**: $F_X(y) = P(X \leq y) = \dfrac{y-a}{b-a}$
- **Expectation**: $E(X) = \dfrac{b+a}{2}$
- **Variance**: $\text{Var}(X) = \dfrac{(b-a)^2}{12}$

Exponential Distribution

- Def: A random variable X is exponentially distributed, $X \sim \text{Expo}(\lambda)$ with $\lambda > 0$ where λ is called the rate parameter.
  - Specifically, λ is given in units of: #-of-events/unit-time.
  - The exponential distribution is typically **concerned with the amount of time until some specific event occurs**. So for X~Expo(λ) the random variable X corresponds to the amount of time one has to wait for the occurrence of an event (e.g., an earthquake to happen, a phone call to end, the battery to die).
- **PDF**: $f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$
- **CDF**: $F_X(y) = \displaystyle\int_0^y \lambda e^{-\lambda x} dx = -e^{-\lambda x}\Big|_0^y = 1 - e^{-\lambda y}$
- **Expectation**: $E(X) = \dfrac{1}{\lambda}$
- **Variance**: $\text{Var}(X) = \dfrac{1}{\lambda^2}$
- **Memory-less Property**
  - If $X \sim \text{Expo}(\lambda)$, Suppose, we have observed that time s has passed and the next event has not yet occurred. Let Y denote the remaining time until next event. We can find $Y \sim \text{Expo}(\lambda)$.
  - This means that the expected waiting time is Expo(λ) irrespective of the moment that we first started counting.
- Relation to Poisson process
  - Consider events that follows a Poisson process with rate r events per unit time ($\lambda^{\text{Poiss}} = r \cdot t$), $X \sim \text{Pois}(\lambda)$, X is the number of events that happened in a specific time window. Let Y be a random variable that denotes the waiting time between events (also called inter-arrival time). We have $P(Y \leq t) = 1 - e^{-rt}$, so $Y \sim \text{Expo}(\lambda)$

- For Poisson, λ is the number of events in a specific time window.
- For Exponential, λ is the number of events per unit time.
- $\lambda^{\text{Exp}} = \lambda^{\text{Poiss}}/(\text{time window})$

Normal Distribution

- Def: If X has a Normal distribution (a.k.a Gaussian Distribution), then we write $X \sim \text{Norm}(\mu, \sigma^2) = \text{N}(\mu, \sigma^2)$
- **PDF**: $f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad -\infty < x < \infty$
  - The pdf is symmetrical around μ. If we consider two values $x_1 = \mu + t$ and $x_2 = \mu - t$ then you can show that $f_X(x_1) = f_X(x_2)$
  - The maximum value of $f_X(x)$ occurs at $x = \mu$ and the maximal value is $\frac{1}{\sqrt{2\pi}\sigma}$
  - From the maximum value the function decays as e^{-x^2} (very fast).
- **Expectation**: $E(X) = \mu$
- **Variance**: $\text{Var}(X) = \sigma^2$
- **Properties**:
  - If $X \sim \text{Norm}(\mu, \sigma^2)$ and suppose $Y = b + aX$, then $Y \sim \text{Norm}(a\mu + b, a^2\sigma^2)$
- The Standard Normal Distribution $Z \sim N(0,1)$
  - Let $\Phi(x)$ denote $P(Z \le x)$, we have $\Phi(x) = 1 - \Phi(-x)$
  - If $X \sim N(\mu, \sigma^2)$, we have $X = \mu + \sigma Z$. Thus, $\frac{X-\mu}{\sigma} = Z \sim N(0,1)$
  - $F_X(a) = P(X \le a) = P(\frac{X-\mu}{\sigma} \le \frac{a-\mu}{\sigma}) = P(Z \le \frac{X-\mu}{\sigma}) = \Phi(\frac{X-\mu}{\sigma})$

*Power-law Distribution

- PDF: $f_X(x) = ax^{-k}$, $k \ge 1$, a is a normalization constant that depends on k.
- “Fat tailed” distribution

*Entropy

- Def: Suppose a random variable X takes on values $x_1, x_2, \dots, x_k$ with probability $p_1, p_2, \dots, p_k$. $H(X) = -\sum_{i=1}^{n} p_i \log_2(p_i)$
- Intuition: The entropy is essentially the uncertainty that we have about the outcome of a random variable.

Lecture 14

Binary Classification

- The binary classifier returns *T=1* for a sample if $X > x*$ (threshold) for that sample, otherwise it returns *T=0*.
- Example:

$$D = \begin{cases} 1 & \text{sample is infected with the disease} \\ 0 & \text{sample is not infected with the disease} \end{cases}$$

$$T = \begin{cases} 1 & \text{test indicates a positive result} \\ 0 & \text{test indicates a negative result} \end{cases}$$

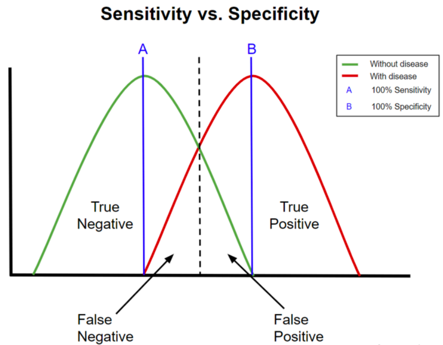


Image from Wikipedia

- Two population have two different distributions
  - The population with D=1 contributes to
    - TPR/Sensitivity: $\eta = \int_{x*}^{\infty} f_{D=1}(x)dx = 1 - \text{CDF}_{D=1}(x*)$
    - FNR: $1 - \eta = \int_{-\infty}^{x*} f_{D=1}(x)dx$
  - The population with D=0 contributes to
    - TNR/Specificity: $\theta = \int_{-\infty}^{x*} f_{D=0}(x)dx = \text{CDF}_{D=0}(x*)$
    - FPR: $1 - \theta = \int_{x*}^{\infty} f_{D=0}(x)dx = 1 - \text{CDF}_{D=0}(x*)$

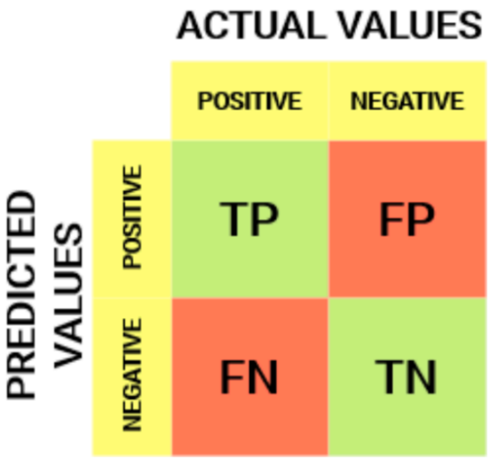
Prevalence, Sensitivity, and Specificity

- Prevalence: Prevalence $(\pi) = P(D = 1)$, How common is the disease?

- Sensitivity / True positive rate (TPR): Sensitivity $(\eta) = P(T = 1 \mid D = 1)$
- Specificity / True negative rate (TNR): Specificity $(\theta) = P(T = 0 \mid D = 0)$
- False positive rate (FPR)=1 - Specificity(θ) / TNR=$P(T = 1 \mid D = 0)$
- False negative rate (FNR)=1 - Sensitivity(η) / TPR=$P(T = 0 \mid D = 1)$

Confusion matrix

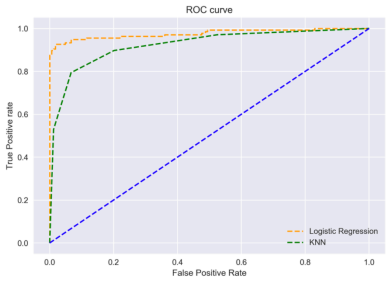
- The entries of the matrix elements are the **counts** (the number of times such a sample was observed).
- $TP = |T = 1 \cap D = 1|$
- $TN = |T = 0 \cap D = 0|$
- $FP = |T = 1 \cap D = 0|$
- $FN = |T = 0 \cap D = 1|$
- Sensitivity/ Recall / TPR $\quad \eta = \dfrac{TP}{TP + FN}$
- Specificity/ TNR $\quad \theta = \dfrac{TN}{TN + FP}$
- In information retrieval,
  - Precision $= \dfrac{TP}{TP + FP}$ (how many of the actual positive cases were correctly predicted by the classifier)
  - Recall $= \dfrac{TP}{TP + FN}$ (how many of the actual positive cases did the classifier predict correctly)



		ACTUAL VALUES	
		POSITIVE	NEGATIVE
PREDICTED VALUES	POSITIVE	TP	FP
	NEGATIVE	FN	TN

How to determine the quality of the classifier?

- Predictive value of a positive test, $\gamma = P(D = 1 \mid T = 1) = \dfrac{\pi\eta}{\pi\eta + (1 - \pi)(1 - \theta)}$
- Predictive value of a negative test, $\delta = P(D = 0 \mid T = 0) = \dfrac{\theta \times (1 - \pi)}{\theta \times (1 - \pi) + \pi \times (1 - \eta)}$
- ROC (Receiver Operating Characteristic) is plot of TPR η versus FPR $1 - \theta$.
  - Used to evaluate the performance of a classifier.
  - AUC (Area Under the Curve): The bigger the better.



Friday, April 9, 2021

Lecture 15

Stochastic Process

- Important features
  - **Time series**: observation of the random variable evolving in time.
  - **Autocorrelation**:
    - Consider a window of a fixed length denoted as the lag and calculate the sum of least squares in that window. Now slide the window over one and do it again. This is called a sliding time window.
    - The length of the window is called the lag.
    - Math formulation: Suppose $X_1, X_2, \ldots, X_n$ are observations at times $t_1, t_2, \ldots, t_n$, then the k lag auto-correlation is given by $r_k = \dfrac{\sum_{i=1}^{n-k} (X_i - \bar{X})(X_{i+k} - \bar{X})}{\sum_i^N (X_i - \bar{X})^2}$. Where $\bar{X}$ is the average value of observations. Note that when $k = 0, r_k = 1$.
- Def:
  - A stochastic process is a collection of random variables indexed by time.
  - $X_1, X_2, \ldots$ are random variables and the subscripts $1, 2, \ldots$ are steps in time.
  - **State**: The values assumed by the random variables $X_1, X_2, \ldots$. And the set of all states is called the **state space,** $S = \{X_1, X_2, \ldots\}$
  - **Evolution of a stochastic process**
    - If $X_1 = i$ and $X_2 = j$ then we say that the (stochastic) process made a transition from state i to state j in one time step (i.e. timestep 1 to timestep 2)
    - Typically, we are interested in the long-run behavior. That is after many transitions what are the probabilities of finding the process in each of the different states.

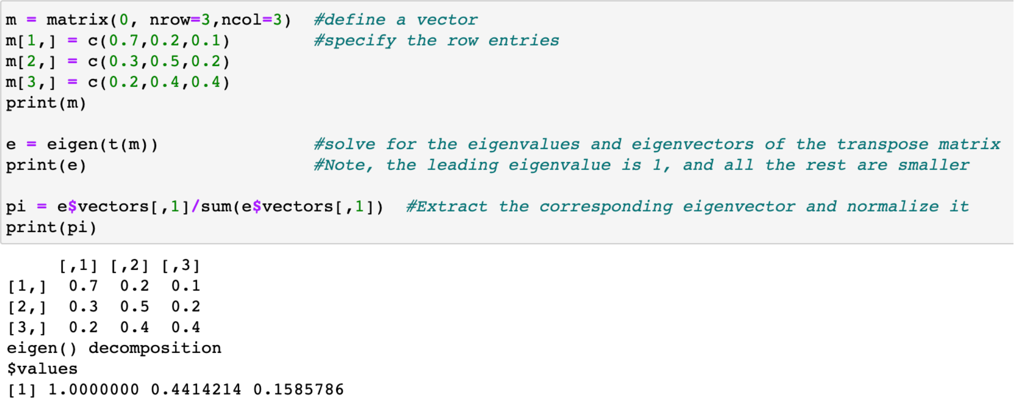
Markov Chain

- A Markov Chain is a mathematical model to capture **one-step dependence** in a stochastic process with a finite size state-space.
- Typically, we are interested in the long time behavior once the process no longer "remembers" what state it started in.
- **Markov assumption**: the dependency is only one-step. Explicitly, $P(X_n = 1 \mid X_{n-1} = 0, X_{n-2} = i_{n-2}, \ldots, X_1 = i_1) = P(X_n = 1 X_{n-1} = 0) \equiv p_{01}$.
  - This implies that that next state only depends on the current state.

2-state discrete time Markov chains

- A 2-state Markov Chain is a sequence of random variables $X_n, n = 1,2,\ldots$ where X_n can take on only two values 0 and 1.
- **State transition matrix**, P
  - **One-step transition probability matrix**: $P = \begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix} = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix}$
  - Matrix element p_{ij} is the probability to transit from state i to state j in one time step.
  - To track the evolution of the system we have a vector encoding our initial condition and multiply it by P to get the probabilities in the next step. For instance, if we started in state 0, the initial vector $\lambda_0 = [1,0]$ and the probabilities of being at any state at time one $\lambda_1 = \lambda_0 P$.
  - r-step transition matrix: Elements of P_r call $P_{i,j}(r) = P(X_r = j \mid X_0 = i)$
  - $P^r = \dfrac{1}{\alpha+\beta}\begin{bmatrix} \beta & \alpha \\ \beta & \alpha \end{bmatrix} = \dfrac{(1-\alpha-\beta)^r}{\alpha+\beta}\begin{bmatrix} \alpha & -\alpha \\ -\beta & \beta \end{bmatrix}$
  - 2 important properties
    - For a stochastic process with *n* states, the transition matrix will be an *n*×*n* matrix.
    - Since *P*(*S*)=1, we must end up in some state. Therefore, each row of a valid transition matrix will sum to 1.
  - Right stochastic matrix (if each row must sum to 1)
    - the eigenvectors are row-vectors which are multiplied with the matrix on the right
    - If Λ is an eigenvector with eigenvalue l then $\Lambda P = l\Lambda$
  - *Left stochastic matrix (if each column must sum to 1)
    - the eigenvectors are column-vectors which are multiplied with the matrix on the left
    - If Λ is an eigenvector with eigenvalue l then $P\Lambda = l\Lambda$
- **State occupancy probability** λ: The elements of the vector λ_r are the **state occupancy probabilities** at r-steps into the process
  - After r steps, $\lambda_r = \lambda_0 P^r$
  - Occupancy probability after r steps: Elements of $\lambda_r = [p_0(r), p_1(r), \ldots, p_n(r)]$
- **Stationary / Steady-State probability vector** π (as $r \to \infty, \lambda \to \pi$)
  - Def: $\pi P = \pi$, in other words π is an eigenvector with eigenvalue of 1.
  - Solving the equation, we get the stationary distribution of P:

    - As $r \to \infty$, $P^{\infty} = \begin{bmatrix}\pi \\ \pi\end{bmatrix} = \frac{1}{\alpha+\beta}\begin{bmatrix}\beta & \alpha \\ \beta & \alpha\end{bmatrix}$, $\lambda_r = \pi = [p_0, p_1] = [\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta}]$
    - In other words, p_0 is the long-run probability (steady-state) that, regardless of what state you began in, you will end up in state 0. p_1 is the same for state 1.
  - Solve steady-state probability with R



```
m = matrix(0, nrow=3,ncol=3)  #define a vector
m[1,] = c(0.7,0.2,0.1)        #specify the row entries
m[2,] = c(0.3,0.5,0.2)
m[3,] = c(0.2,0.4,0.4)
print(m)

e = eigen(t(m))               #solve for the eigenvalues and eigenvectors of the transpose matrix
print(e)                      #Note, the leading eigenvalue is 1, and all the rest are smaller

pi = e$vectors[,1]/sum(e$vectors[,1])  #Extract the corresponding eigenvector and normalize it
print(pi)
```

```
     [,1] [,2] [,3]
[1,]  0.7  0.2  0.1
[2,]  0.3  0.5  0.2
[3,]  0.2  0.4  0.4
eigen() decomposition
$values
[1] 1.0000000 0.4414214 0.1585786
```

Lecture 18

Bounding Probability

- Markov Inequality (Given the **mean** of the distribution)
  - If X is a random variable that takes only nonnegative values, then for any a>0, $P(X \geq a) \leq \dfrac{E(X)}{a}$
- Chebyshev's Inequality (Given the **mean** and **variance** of the distribution)
  - If X is a random variable with finite mean μ and variance σ^2, then for any k>0, $P(|X-\mu| \geq k) \leq \dfrac{\sigma^2}{k^2}$
- Weak law of large number
  - Let $X_1, X_2, \ldots$ be a sequence of **independent and identically distributed** (IID) random variables with finite mean $E(X_i) = \mu$. Then for any $\epsilon > 0$,
$$P\left(\left|\frac{X_1 + X_2 + \ldots + X_n}{n} - \mu\right| \geq \varepsilon\right) \to 0 \quad \text{as} \quad n \to \infty$$

Sample Mean and Sample Variance

- The goal of **statistical inference** (estimation) is to infer (estimate) characteristics (parameters) when we cannot observe the entire population or to infer (estimate) the underlying distribution that generated the observed data.
- **Observations** are values of a random variable X sampled from the density function $f_X(x)$.
- Let X_i, *i=1, 2, 3, …, n* be the random variables that represent the i^{th} sample each having the same probability density function $f_X(x)$. If X_i's are random (independent) samples then $f(x_1 \cap x_2 \cap \ldots \cap x_n) = f_X(x_1)f_X(x_2)\ldots f_X(x_n)$
- Sample Mean: $\bar{X} = \dfrac{X_1 + X_2 + \ldots + X_n}{n}$
- Sample Variance: $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

Error bars

- Error bars: how much uncertainty there is in your estimate obtained from the **sample mean**.
- Error bars are drawn above and below a data point, which extend the length of the **standard error** in each direction.
- Standard error: $\sqrt{\dfrac{S^2}{n}} = \dfrac{S}{\sqrt{n}}$, where S^2 denotes the sample variance and n denotes the number of observations.

Central Limit Theorem

- Let $X_1, X_2, \ldots, X_n$ be a sequence of **independent and identically distributed** (IID) random variables with mean μ and variance σ^2.
- The true distribution doesn't matter.
- Then the distribution of Sample Mean $\bar{X} \sim N(\mu, \dfrac{\sigma^2}{n})$, when n is large. Thus, $\dfrac{X_1 + X_2 + \ldots + X_n}{n} \sim N(\mu, \dfrac{\sigma^2}{n})$
- With the linear transformation, we have $\dfrac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$
- $P(\mu - a \leq \bar{X} \leq \mu + a) = P(\dfrac{\mu - a - \mu}{\frac{\sigma}{\sqrt{n}}} \leq Z \leq \dfrac{\mu + a - \mu}{\frac{\sigma}{\sqrt{n}}}) = 2\Phi(\dfrac{\sqrt{n}}{\sigma}a) - 1$

Friday, April 9, 2021

Lecture 19

Maximum Likelihood Estimation

- Let $X_1, X_2, \ldots, X_n$ be independent random variables that are taken from a distribution with density function $f_X(x;\theta)$ where θ is the parameter of the distribution. We define likelihood function $L(x_1, x_2, \ldots, x_n; \theta) = f(x_1 \cap x_2 \cap \ldots \cap x_n; \theta) = f(x_1;\theta)f(x_2;\theta)\ldots f(x_n;\theta)$
- The maximum likelihood estimator is the value of the parameter θ that results in the maximum value of the joint probability.
- The maximum is at the point where the first derivative $\dfrac{dL}{d\theta} = 0$ equals 0 and the second derivative $\dfrac{d^2L}{d\theta^2} = 0$ is negative.
- **Log likelihood**: In most cases it is easier to work with log of the likelihood function and we find the value of the parameter that maximizes the log likelihood function.

Linear Model - a model of Linear Correlation

- We consider two independent, random variables X and Z with $X \sim N(0, \sigma_x)$ and $Z \sim N(0, \sigma_z)$. We define $Y = mX + Z$.
- We want a measure of correlation between X and Y.
  - Perfect correlation ($\sigma_z = 0$) (Z has no fluctuation, Z=0)
  - Some correlation ($\sigma_z < \sigma_x$) (Z has limited fluctuation that infects Y)
  - Zero correlation ($m = 0$)

Measure of Correlation

- Def: Correlation Coefficient $r = \dfrac{m\sigma_x}{\sigma_y} = \dfrac{m\sigma_x}{\sqrt{m^2\sigma_x^2 + \sigma_z^2}}$
- The correlation coefficient r measures the proportion of the standard deviation of Y that is contributed by mX. If the proportion is large, then the correlation is large. (In the extreme case, $\sigma_z = 0 \rightarrow r = 1$)
- If m<0 then X and Y will be negative correlated. The value of $-1 \leq r \leq 1$.
- $Var(aX) = a^2\sigma_x^2, Var(aX + bY) = a^2\sigma_x^2 + b^2\sigma_y^2$

Measure of Covariance

- Def: The Covariance of two random variables X and Y is defined as $Cov(X,Y) = E[(X-\mu_x)(Y-\mu_y)]$, where μ_x, μ_y are the mean of X and Y.
- If X=Y, then $Cov(X,Y) = E[(X-\mu_x)^2] = Var(X)$
- If $\mu_x = \mu_y = 0$, then $Cov(X,Y) = E(XY)$.
- Correlation Coefficient: $r = \dfrac{Cov(X,Y)}{\sigma_x\sigma_y} = \dfrac{E[(X-\mu_x)(Y-\mu_y)]}{\sigma_x\sigma_y}$
- Two correlation coefficient definitions are the same: $r = \dfrac{E[(X-\mu_x)(Y-\mu_y)]}{\sigma_x\sigma_y} = \dfrac{m\sigma_x}{\sigma_y}$ where $Y = mX + c$.
- Correlation Coefficient on given sampled data
  - $Cov(X,Y) = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$
  - $S_x^2 = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ and $S_y^2 = \dfrac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$
  - $r = \dfrac{Cov(X,Y)}{\sigma_x\sigma_y} = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$
  - $m = r\dfrac{S_y}{S_x}$

The best-fit line

- Def: Give a set of points, we define the best fit line to be the one that minimizes the sum of the squares of the vertical distances between the line and the points $\Gamma = \sum_{i=1}^{n}(y_i - (Ax_i + B))^2$.
- When $A = r\dfrac{S_y}{S_x}$, Γ is the minimum.

Appendix - R

Vector

- numeric(3) => [0, 0, 0]
- c(1:5) => [1, 2, 3, 4, 5]
- seq(1:5) => [1, 2, 3, 4, 5], seq(from, to, by, length)
- c(1, 2, 3) => [1, 2, 3]
- Index starts at 1

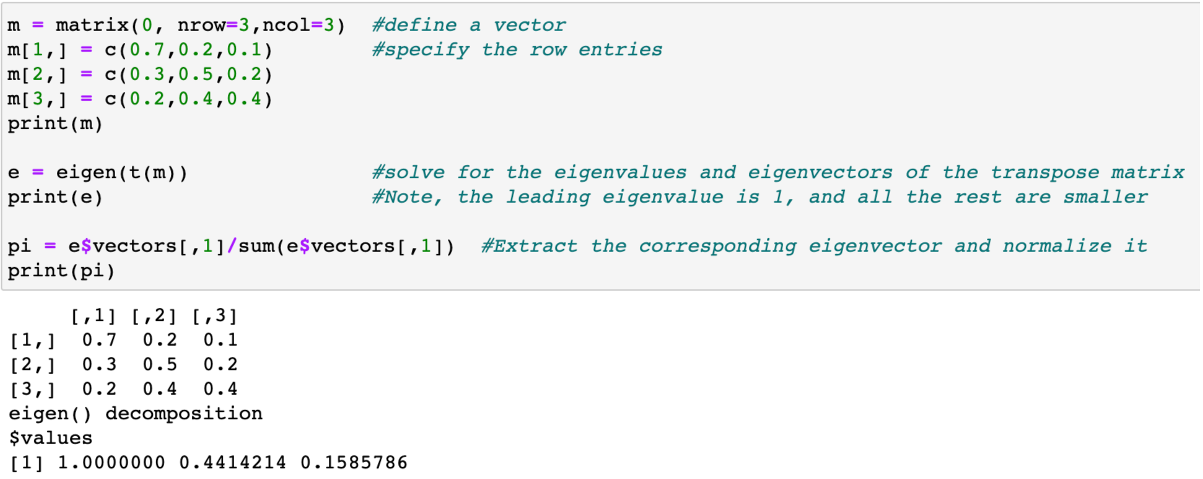
Useful function

- Random sample: sample(x, size, replace=FALSE, prob=NULL): sample size times from x.
- Arithmetic mean: mean(x==1)
- Format print: sprintf('x is %.3f', x)

Matrix

```
1.  M1 + M2              # addition
2.  M1 * M2              # element-wise multiplication
3.  M1 %*% M2            # matrix multiplication
4.  nrow(M), ncol(M)     # number of rows and columns respectively
5.  t(M)                 # transpose
6.  solve(M)             # inverse
7.  diag(c(1, 2, 4))     # Construct diagonal matrix
8.  eigen(M)             # Eigen decomposition
9.  det(M)               # Computes determinant
```

- Eigenvector: eigen(matrix)



```r
m = matrix(0, nrow=3,ncol=3)  #define a vector
m[1,] = c(0.7,0.2,0.1)        #specify the row entries
m[2,] = c(0.3,0.5,0.2)
m[3,] = c(0.2,0.4,0.4)
print(m)

e = eigen(t(m))               #solve for the eigenvalues and eigenvectors of the transpose matrix
print(e)                      #Note, the leading eigenvalue is 1, and all the rest are smaller

pi = e$vectors[,1]/sum(e$vectors[,1])  #Extract the corresponding eigenvector and normalize it
print(pi)
```

```
     [,1] [,2] [,3]
[1,]  0.7  0.2  0.1
[2,]  0.3  0.5  0.2
[3,]  0.2  0.4  0.4
eigen() decomposition
$values
[1] 1.0000000 0.4414214 0.1585786
```

Distributions

- Binomial Distribution
  - dbinom takes arguments j, n, p and returns the PMF, $p_j = P(X = j)$
  - pbinom takes arguments x, n, p and returns the CDF, $F_X(x) = P(X \leq x)$
  - qbinom takes arguments x, n, p and returns the first value of j with $F_X(j) \geq x$
  - rbinom samples from the binomial distribution
- Geometric Distribution
  - dgeom takes arguments j, p and returns the PMF, $p_j = P(X = j)$
  - pgeom takes arguments x, p and returns the CDF, $F_X(x) = P(X \leq x)$
  - qgeom takes arguments x, p and returns the first value of j with $F_X(j) \geq x$
  - rgeom samples from the geometric distribution
- Poisson Distribution
  - dpois, ppois, qpois, rpois
- Normal Distribution (μ, σ)
  - dnorm, pnorm, qnorm, rnorm. The second argument is standard deviation, not variance!!!
- Graph
  - plot(x, y, xlab, ylab, type)
  - hist(data, breaks): breaks usually seq(low, high, increment)
  - lines(x, y): draw another line
  - abline(h, v): draw a straight line
  - boxplot(data): plot 0th, 25th, 50th, 75th, 100th quantiles.
    - ▹ the whiskers extend only to the most extreme data point which is no more than range×IQR ($IQR = Q_{75} - Q_{25}$) from the box.

Linear Regression

- cor(x, y) computes the correlation of x and y if these are vectors
- cov(x, y) computes the covariance of x and y if these are vectors
- lm(y ~ x) used to fit linear models