Georgian

Statistics and Probability part 1

Garima Malik

January 10, 2022

Outline

Central Tendencies

Probability and its axioms

Random Variables

Probability Distributions

Central Tendencies

Central Tendencies

- A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. As such, measures of central tendency are sometimes called measures of central location. They are also classed as summary statistics.
- The mean (often called the average) is most likely the measure of central tendency that you are most familiar with, but there are others, such as the median and the mode.
- The mean, median and mode are all valid measures of central tendency, but under different conditions, some measures of central tendency become more appropriate to use than others.

Mean

- The mean (or average) is the most popular and well known measure of central tendency. It can be used with both discrete and continuous data, although its use is most often with continuous data.
- Sample Mean :

$$\bar{x} = \frac{x_1 + x_2 ..... + x_n}{n} \tag{1}$$

- To acknowledge that we are calculating the population mean and not the sample mean, we use the Greek lower case letter "mu", denoted as
- population Mean :

$$\mu = \frac{\sum x}{n} \tag{2}$$

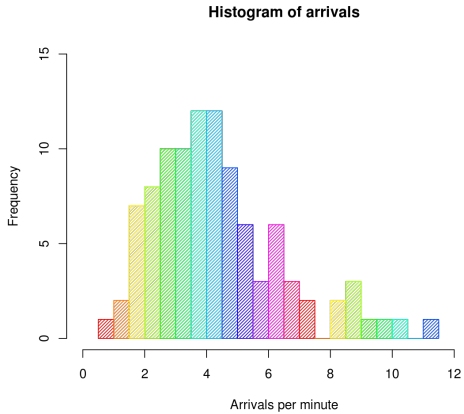
- The mean has one main disadvantage: it is particularly susceptible to the influence of outliers.

Median

- The median is the middle score for a set of data that has been arranged in order of magnitude. The median is less affected by outliers and skewed data.

Mode

- The mode is the most frequent score in our data set. On a histogram it represents the highest bar in a bar chart or histogram. You can, therefore, sometimes consider the mode as being the most popular option.



Summary - Central tendencies

Type of Variable	Best measure of central tendency
Nominal	Mode
Ordinal	Median
Interval/Ratio (not skewed)	Mean
Interval/Ratio (skewed)	Median

Probability

- Empirical Definition : Given an event A,

$$Pr(A) = \lim_{n \to \infty} \frac{n_A}{n} \tag{3}$$

where n_A is no of times event A is observed and n is the number of trials.

- Classical Definition : Given an event A,

$$Pr(A) = \frac{N_A}{N} \tag{4}$$

where N_A Total number of outcomes that are favorable to A, N is the total number of all possible outcomes that are equally likely

Probability

- Example : Estimate the prob. of rolling 2 dice where the sum = 7 using definitions of probability.

Axioms of Probability

- For any event A, P(A) $\geq$ 0.
- Probability of the sample space S is P(S)=1.
- if $A_1, A_2, A_3 ..... A_n$ are disjoint events then

$$p(A_1 \cup A_2 \cup A_3 \cup A_4 ...) = P(A_1) + P(A_2) + P(A_3) + ... + P(A_n) \quad (5)$$

- P(A $\cap$ B) = P(A and B) = P(A,B)
- P(A $\cup$ B) = P(A or B)

Example

- In a presidential election, there are four candidates. Call them A, B, C, and D. Based on our polling analysis, we estimate that A has a 20 percent chance of winning the election, while B has a 40 percent chance of winning. What is the probability that A or B win the election?

Conditional Probability

- If A and B are two events in a sample space S, then the conditional probability of A given B is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{6}$$

where P(B) should be greater than 0.

Example

- I roll a fair die twice and obtain two numbers X1= result of the first roll and X2= result of the second roll. Given that I know X1+X2=7, what is the probability that X1=4 or X2=4?

Random Variables

- A random variable is a real-valued variable whose value is determined by an underlying random experiment.
- A random variable X is a function from the sample space to the real numbers

$$X : S \rightarrow R \tag{7}$$

Example

- I toss a coin five times. This is a random experiment and the sample space can be written as S=TTTTT,TTTTH,...,HHHHH.
- Note that here the sample space S has 32 elements. Suppose that in this experiment, we are interested in the number of heads. We can define a random variable X whose value is the number of observed heads. The value of X will be one of 0,1,2,3,4 or 5 depending on the outcome of the random experiment.

Random Variables

- Discrete Random Variable : Range is countable
- Continuous Random Variable : Range is not countable

Discrete Random Variables

- Concept of PMF (Probability Mass Function):

Let X be a discrete random variable with range $R_X = \{x_1, x_2, x_3, \dots\}$ (finite or countably infinite). The function

$$P_X(x_k) = P(X = x_k), \text{ for } k = 1, 2, 3, \dots,$$

is called the *probability mass function (PMF)* of X.

- Thus, the PMF is a probability measure that gives us probabilities of the possible values for a random variable. While the above notation is the standard notation for the PMF of X, it might look confusing at first. The subscript X here indicates that this is the PMF of the random variable X. Thus, for example, $P_X(1)$ shows the probability that X=1.

PMF

- For discrete random variables, the PMF is also called the probability distribution.
- I toss a fair coin twice, and let X be defined as the number of heads I observe. Find the range of X, R_X, as well as its probability mass function P_X.

Bernoulli RV

A random variable X is said to be a *Bernoulli* random variable with *parameter* p, shown as $X \sim Bernoulli(p)$, if its PMF is given by

$$P_X(x) = \begin{cases} p & \text{for } x = 1 \\ 1-p & \text{for } x = 0 \\ 0 & \text{otherwise} \end{cases}$$

where $0 < p < 1$.

Geometric RV

A random variable X is said to be a *geometric* random variable with *parameter* p, shown as $X \sim Geometric(p)$, if its PMF is given by

$$P_X(k) = \begin{cases} p(1-p)^{k-1} & \text{for } k = 1, 2, 3, \dots \\ 0 & \text{otherwise} \end{cases}$$

where $0 < p < 1$.

Example

- I roll a fair die repeatedly until a number larger than 4 is observed. If N is the total number of times that I roll the die, find P(N=k), for k=1,2,3,....

Binomial RV

A random variable X is said to be a *binomial* random variable with parameters n and p, shown as $X \sim Binomial(n, p)$, if its PMF is given by

$$P_X(k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & \text{for } k = 0, 1, 2, \cdots, n \\ 0 & \text{otherwise} \end{cases}$$

where $0 < p < 1$.

Example

- What is the probability of finding at most one defective part in picking 3 parts at a time from a box containing 500 parts. The probability of finding the defective part is 0.25?

Poisson RV

A random variable X is said to be a *Poisson* random variable with parameter λ, shown as $X \sim Poisson(\lambda)$, if its range is $R_X = \{0, 1, 2, 3, \dots\}$, and its PMF is given by

$$P_X(k) = \begin{cases} \frac{e^{-\lambda}\lambda^k}{k!} & \text{for } k \in R_X \\ 0 & \text{otherwise} \end{cases}$$

Example

- The number of emails that I get in a weekday can be modeled by a Poisson distribution with an average of 0.2 emails per minute.
  - What is the probability that I get no emails in an interval of length 5 minutes?
  - What is the probability that I get more than 3 emails in an interval of length 10 minutes?

Thank You

Georgian

Statistics and Probability part 2

Garima Malik

January 17, 2022

Outline

Continuous Probability Distributions

Probability Distributions

Statistical Tests

Continuous Probability Distributions

CDF: Cumulative Distribution Function

- The PMF is one way to describe the distribution of a discrete random variable. As we will see later on, PMF cannot be defined for continuous random variables. The cumulative distribution function (CDF) of a random variable is another method to describe the distribution of random variables. The advantage of the CDF is that it can be defined for any kind of random variable (discrete, continuous, and mixed).

Definition 3.10
The cumulative distribution function (CDF) of random variable X is defined as

$$F_X(x) = P(X \leq x), \text{ for all } x \in \mathbb{R}.$$

Example

- I toss a coin twice. Let X be the number of observed heads. Find the CDF of X.

Example

- I choose a real number uniformly at random in the interval [a,b], and call it X. By uniformly at random, we mean all intervals in [a,b] that have the same length must have the same probability. Find the CDF of X.

PDF: Probability Density Function

- We have the following definition for the PDF of continuous random variables:

Definition 4.2

Consider a continuous random variable X with an absolutely continuous CDF $F_X(x)$. The function $f_X(x)$ defined by

$$f_X(x) = \frac{dF_X(x)}{dx} = F'_X(x), \qquad \text{if } F_X(x) \text{ is differentiable at } x$$

is called the probability density function (PDF) of X.

Properties of PDF

- Basic properties of PDF:

Consider a continuous random variable X with PDF $f_X(x)$. We have

1. $f_X(x) \geq 0$ for all $x \in \mathbb{R}$.
2. $\int_{-\infty}^{\infty} f_X(u)du = 1$.
3. $P(a < X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(u)du$.

Example

- Let X be a continuous random variable with the following PDF :

$$f_X(x) = \begin{cases} ce^{-x} & x \geq 0 \\ 0 & otherwise \end{cases} \tag{1}$$

where c is a positive constant.

- Find c.
- Find the CDF of X, $F_X(x)$.
- Find P(1<X<3).

Exponential RV

A continuous random variable X is said to have an *exponential* distribution with parameter $\lambda > 0$, shown as $X \sim Exponential(\lambda)$, if its PDF is given by

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

Example

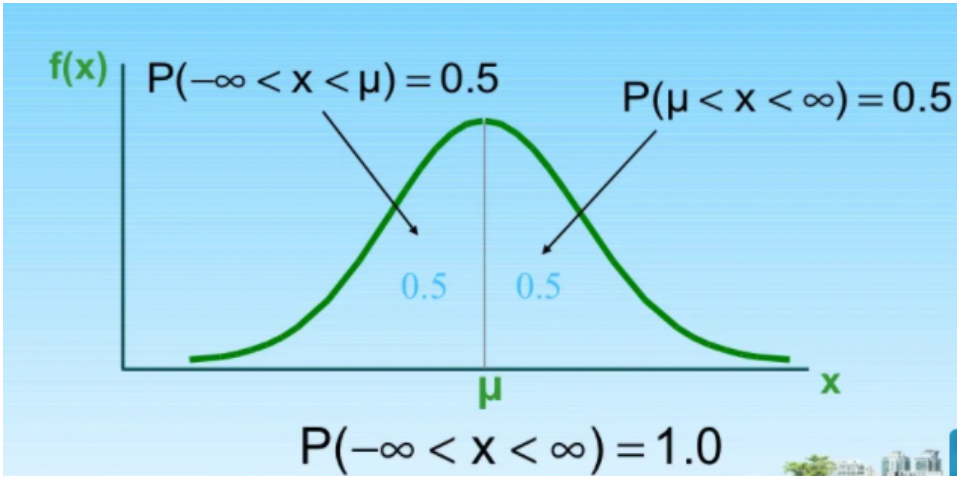
- Let X = amount of time (in minutes) a postal clerk spends with his or her customer. The time is known to have an exponential distribution with the average amount of time equal to four minutes. What will be the probability of when postal clerk takes 5 minutes with the customer?

Standard Normal(Gaussian) RV

A continuous random variable Z is said to be a *standard normal (standard Gaussian)* random variable, shown as $Z \sim N(0,1)$, if its PDF is given by

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2}\right\}, \qquad \text{for all } z \in \mathbb{R}.$$

Empirical Properties



Georgian

Example

- If X is normally distributed normally between mean of 100 and standard deviation of 50 find the probability of $P(0 < z < 2.0)$?
  - General way of finding probability in Normal RV
    - Draw the normal curve
    - Translate x values to z values
    - Read the z table

Statistical Tests

- A statistical test provides a mechanism for making quantitative decisions about a process or processes. The intent is to determine whether there is enough evidence to "reject" a conjecture or hypothesis about the process.
- A classic use of a statistical test occurs in process control studies. For example, suppose that we are interested in ensuring that photomasks in a production process have mean linewidths of 500 micrometers. The null hypothesis, in this case, is that the mean linewidth is 500 micrometers. Implicit in this statement is the need to flag photomasks which have mean linewidths that are either much greater or much less than 500 micrometers. This translates into the alternative hypothesis that the mean linewidths are not equal to 500 micrometers.

Statistical Tests

- The null hypothesis is a statement about a belief. We may doubt that the null hypothesis is true, which might be why we are "testing" it. The alternative hypothesis might, in fact, be what we believe to be true. The test procedure is constructed so that the risk of rejecting the null hypothesis, when it is in fact true, is small. This risk, ‘alpha’, is often referred to as the significance level of the test. By having a test with a small value of alpha, we feel that we have actually "proved" something when we reject the null hypothesis.

Hypothesis Testing

Common format of hypothesis test

- H_0 : Null hypothesis
- H_a : Alternate hypothesis
- The test statistic is based on the specific hypothesis test.
- α : significance level

Examples

- A researcher thinks that if knee surgery patients go to physical therapy twice a week (instead of 3 times), their recovery period will be longer. Average recovery times for knee surgery patients is 8.2 weeks.
- A principal at a certain school claims that the students in his school are above average intelligence. A random sample of thirty students IQ scores have a mean score of 112.5. Is there sufficient evidence to support the principalâs claim? The mean population IQ is 100 with a standard deviation of 15.
- Blood glucose levels for obese patients have a mean of 100 with a standard deviation of 15. A researcher thinks that a diet high in raw cornstarch will have a positive or negative effect on blood glucose levels. A sample of 30 patients who have tried the raw cornstarch diet have a mean glucose level of 140. Test the hypothesis that the raw cornstarch had an effect.

Georgian

Linear Algebra

Garima Malik

January 26, 2022

Outline

Vectors

Matrices

Eigen values and vectors

Functions

Vectors

- Science and Maths are used to describe the world around us. There are many quantities which require only 1 measurement to describe them. e.g Length of a string, or area of any shape or temperature of any surface. Such quantities are called scalars. Any quantity which can be represented as a number (positive or negative) is called scalar. This value is known as magnitude.
- On the other hand, there are quantities which require at least 2 measurements to describe them. Along with the magnitude, they have a “direction" associated e.g velocity or force. These quantities are known as “Vectors".
- When we say that a person ran for 2 Km, its a scalar but when we say that a person ran for 2 Km, North-east from his initial position, its a vector.

Operations on vectors

Consider two vectors $\vec{A} = (a_1, a_2, ...a_n)$ and $\vec{B} = (b_1, b_2, ...b_n)$

- Vector Addition
  - $\vec{C} = (a_1 + b_1, a_2 + b_2, ...a_n + b_n)$
- Vector Subtraction
  - $\vec{C} = (a_1 - b_1, a_2 - b_2, ...a_n - b_n)$
- Vector Multiplication
- Vector Subtraction
  - $\vec{C} = (a_1 * b_1, a_2 * b_2, ...a_n * b_n)$
- Scalar Multiply to vector
  - $K.(\vec{A}) = K * a_1, K * a_2.....K * a_n$

Operations on vectors

Consider two vectors $\vec{A} = (a_1, a_2, ...a_n)$ and $\vec{B} = (b_1, b_2, ...b_n)$

- Magnitude of Vectors
  - $|A| = \sqrt{a_1^2 + a_2^2 + ...a_n^2}$
- A vector of magnitude, or length, 1 is called a unit vector.

For all vectors u, v, and w, and for all scalars b and c:

1. u + v = v + u.
2. u + (v + w) = (u + v) + w.
3. v + O = v.
4. 1.v = v; 0.v = O.
5. v + (- v) = O.
6. b(cv) = (bc)v.
7. (b + c)v = bv + cv.
8. b(u + v) = bu + bv.

Component form of vectors

- unit vectors can have any direction, the unit vectors parallel to the x - and y - axes are particularly useful. They are defined as i = < 1, 0 > and j = < 0, 1 >.
- Any vector can be expressed as a linear combination of unit vectors i and j. For example, let $\vec{v} =< v_1, v_2 >$. Then
$\vec{v} =< v_1, v_2 >=< v_1, 0 > + < 0, v_2 >= v_1 < 1, 0 > +v_2 < 0, 1 >= v_1 i + v_2 j$.

Directions in vectors

- The terminal point P of a unit vector in standard position is a point on the unit circle denoted by ($\cos\theta$, $\sin\theta$). Thus the unit vector can be expressed in component form, $\vec{u} =< cos\theta, sin\theta >$, or as a linear combination of the unit vectors i and j, $\vec{u} = (cos\theta)i + (sin\theta)j$, where the components of u are functions of the direction angle θ measured counterclockwise from the x - axis to the vector. As θ varies from 0 to 2π, the point P traces the circle $x^2 + y^2 = 1$. This takes in all possible directions for unit vectors so the equation $\vec{u} = (cos\theta)i + (sin\theta)j$ describes every possible unit vector in the plane.

Angle between vector

- The dot product of two vectors is a real number, or scalar. This product is useful in finding the angle between two vectors and in determining whether two vectors are perpendicular.
- The dot product of two vectors $\vec{u} =< u_1, u_2 >$ and $\vec{v} =< v_1, v_2 >$ is $u.v = u_1.v_1 + u_2.v_2$
- If θ is the angle between two nonzero vectors u and v, then

$$cos\theta = \frac{u.v}{|u|\,|v|} \tag{1}$$

Examples

- Find a unit vector that has the same direction as the vector $w = \langle -3, 5 \rangle$.
- Find the angle between $u = \langle 3, 7 \rangle$ and $v = \langle -4, 2 \rangle$.

Application of vectors

- Cosine Similarity
  - Cosine Similarity is a metric that gives the cosine of the angle between vectors. It signfies the similarity and dissimilarity between two vectors.
- Text Vectorization
  - The process of converting or transforming a data set into a set of Vectors is called vectorization. It's easier to represent data set as vectors where attributes are already numeric

Matrix

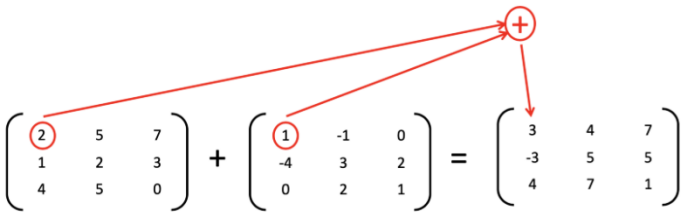
- Conventionally, the number of rows in a matrix is denoted by m and the number of columns by n. Since a rectangle's area is height x width, we denote a matrix's size by m x n. Thus is the matrix was to be called A, it would be written notationally as

$$A = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \end{bmatrix} \tag{2}$$

- A matrix with just one row is called a row matrix and a matrix with just one column is called a column matrix.

Operations on Matrix

- Matrix addition and Subtraction
  - Number of Rows of A = Number of Rows of B
  - Number of Columns of A = Number of Columns of B

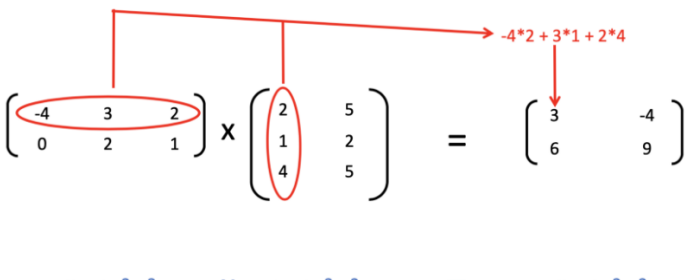


Operations on Matrix

- Matrix addition and Subtraction Properties
  - Addition of matrices is commutative which means A+B = B+A
  - Addition of matrices is associative which means A+(B+C) = (A+B)+C
  - Subtraction of matrices is non-commutative which means A-B $\neq$ B-A
  - Subtraction of matrices is non-associative which means A-(B-C) $\neq$ (A-B)-C
  - The order of matrices A, B, A-B and A+B is always the same
  - If the order of A and B is different, A+B, A-B can't be computed
  - The complexity of addition/subtraction operation is O(m*n) where m*n is order of matrices

Operations on Matrix

- The multiplication of two matrices A(m*n) and B(n*p) gives a matrix C(m*p). Notice that for multiplication you do not need the rows/columns of A and B to be the same. You only need
  - No. of Columns of A = No. of Rows of B
  - Or, No. of Columns of B = No. of Rows of A.



Georgian

Operations on Matrix

- Identity Matrix : It is the matrix equivalent of the number “1":
  - It is "square" (has same number of rows as columns),
  - It has 1s on the diagonal and 0s everywhere else.
  - Its symbol is the capital letter I.

$$I_{2X2} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \tag{3}$$

$$I_{3X3} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{4}$$

Determinant of Matrix

- The determinant is a special number that can be calculated from a matrix.

For a 2×2 Matrix

For a 2×2 matrix (2 rows and 2 columns):

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

The determinant is:

$$|A| = ad - bc$$

"The determinant of A equals a times d minus b times c"

For a 3×3 Matrix

For a 3×3 matrix (3 rows and 3 columns):

$$A = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}$$

The determinant is:

$$|A| = a(ei - fh) - b(di - fg) + c(dh - eg)$$

"The determinant of A equals ... etc"

Inverse of a matrix

- Why inverse ? Because with matrices we don't divide! Seriously, there is no concept of dividing by a matrix.

The inverse of A is A^{-1} only when:

$$AA^{-1} = A^{-1}A = \mathbf{I}$$

Sometimes there is no inverse at all.

Example

- Find the inverse of the given 3X3 matrix as follows :

$$M_{3X3} = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & 4 \\ 5 & 6 & 0 \end{bmatrix} \tag{5}$$

Eigen values and vectors

- Special properties in matrix

Let A be an $n \times n$ matrix and let $X \in \mathbb{C}^n$ be a **nonzero vector** for which

$$AX = \lambda X$$

for some scalar λ. Then λ is called an **eigenvalue** of the matrix A and X is called an **eigenvector** of A associated with λ, or a λ-eigenvector of A.

The set of all eigenvalues of an $n \times n$ matrix A is denoted by $\sigma(A)$ and is referred to as the **spectrum** of A.

Example

- Consider the matrix A and two column vector X_1 and X_2:

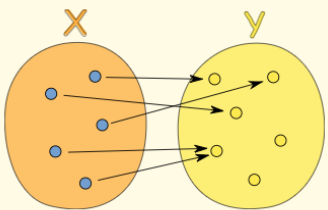
$$A_{3X3} = \begin{bmatrix} 0 & 5 & -10 \\ 0 & 22 & 16 \\ 0 & -9 & -2 \end{bmatrix} X_1 = \begin{bmatrix} 5 \\ -4 \\ 3 \end{bmatrix} X_2 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \quad (6)$$

Functions

- Function, in mathematics, an expression, rule, or law that defines a relationship between one variable (the independent variable) and another variable (the dependent variable). Functions are ubiquitous in mathematics and are essential for formulating physical relationships in the sciences.
- This relationship is commonly symbolized as $y = f(x)$ - which is said “f of x" - and y and x are related such that for every x, there is a unique value of y. That is, $f(x)$ can not have more than one value for the same x.

Functions

- Special rules of a function
  - It must work for every possible input value
  - And it has only one relationship for each input value



Formal Definition of a Function

A function relates **each element** of a set with **exactly one** element of another set (possibly the same set).

Functions

Example: $y = x^3$

- The input set "X" is all Real Numbers
- The output set "Y" is also all the Real Numbers

We can't show ALL the values, so here are just a few examples:

X: x	Y: x^3
-2	-8
-0.1	-0.001
0	0
1.1	1.331
3	27
and so on...	and so on...

Functions

Refer from the previous slide table

- the set "X" is called the Domain,
- the set "Y" is called the Codomain, and
- the set of elements that get pointed to in Y (the actual values produced by the function) is called the Range.

Thank You

Georgian

Optimization Techniques

Garima Malik

February 7, 2022

Outline

Functions

Function Optimization

Maxima and Minima

Unconstrained Optimization

Constrained Optimization

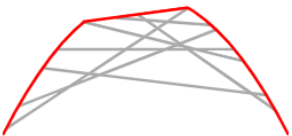
Convex and Concave Functions

- The second derivative of the function depicts how the function is curved, unlike the first derivative which tells us about the slope of the tangent function. A function that has an increasing first derivative bends upwards and is known as a convex function. On the other hand, a function, that has a decreasing first derivative is known as a concave function and bends downwards.

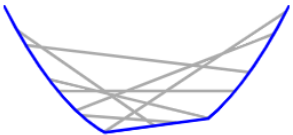
Theorem :

- If the function f and its derivative f' can be differentiated at a, then:
  - The f is convex at a if $f''(a) > 0$
  - The f is concave at a if $f''(a) < 0$

Convex and Concave Functions



A concave function:
no line segment joining two points on the graph lies above the graph at any point



A convex function:
no line segment joining two points on the graph lies below the graph at any point

A function that is neither concave nor convex:
the line segment shown lies above the graph at some points and below it at others

Example

- Identify the curve of the following function and determine whether it is a concave or a convex function:

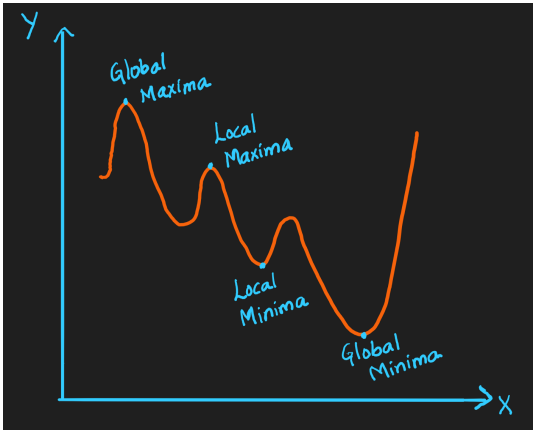
$$f(x) = 3x^2 + 7x - 9$$

Function Optimization

- Function optimization is a widely used tool bag of techniques employed in practically all scientific and engineering disciplines.
- Practically, function optimization describes a class of problems for finding the input to a given function that results in the minimum or maximum output from the function.
- Function Optimization involves three elements: the input to the function (e.g. x), the objective function itself (e.g. f()) and the output from the function (e.g. cost).
  - Input (x): The input to the function to be evaluated, e.g. a candidate solution.
  - Function (f()): The objective function or target function that evaluates inputs.
  - Cost: The result of evaluating a candidate solution with the objective function, minimized or maximized.

Maxima and Minima

- Maxima is the largest and Minima is the smallest value of a function within a given range. We represent them as below:



Maxima and Minima

- Global Maxima and Minima: It is the maximum value and minimum value respectively on the entire domain of the function
- Local Maxima and Minima: It is the maximum value and minimum value respectively of the function within a given range.
- There can be only one global minima and maxima but there can be more than one local minima and maxima.

Example

- Find the maximum of the function $f(x) = x^4 - 8x^2 + 3$ on the interval [-1, 3].

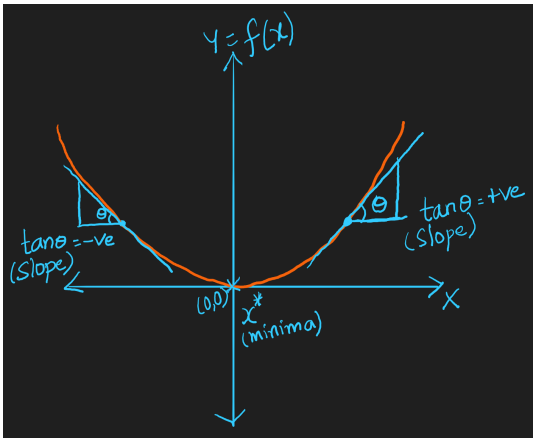
Unconstrained Optimization

- Unconstrained optimization problems consider the problem of minimizing an objective function that depends on real variables with no restrictions on their values.
- Mathematically, let $x \epsilon n$ be a real vector with $n \geq 1$ components and let $f : R^n \rightarrow R$ be a smooth function. Then, the unconstrained optimization problem is :

$$min_x f(x) \tag{1}$$

Gradient Descent

- Gradient Descent is an optimization algorithm and it finds out the local minima of a differentiable function. It is a minimization algorithm that minimizes a given function.



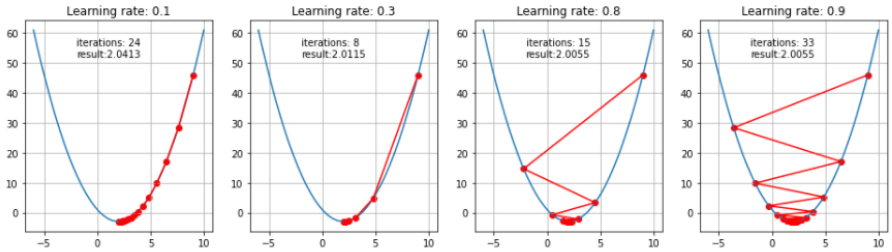
Georgian

Gradient Descent

- Gradient descent algorithm does not work for all functions. There are two specific requirements. A function has to be:
  - differentiable
  - convex
- Gradient Descent method's steps are:
  1. choose a starting point (initialisation)
  2. calculate gradient at this point
  3. make a scaled step in the opposite direction to the gradient (objective: minimise)
  4. repeat points 2 and 3 until one of the criteria is met:
     - maximum number of iterations reached
     - step size is smaller than the tolerance.

Learning Rate

- If the learning rate is too high, we might OVERSHOOT the minima and keep bouncing, without reaching the minima.
- If the learning rate is too small, the training might turn out to be too long.



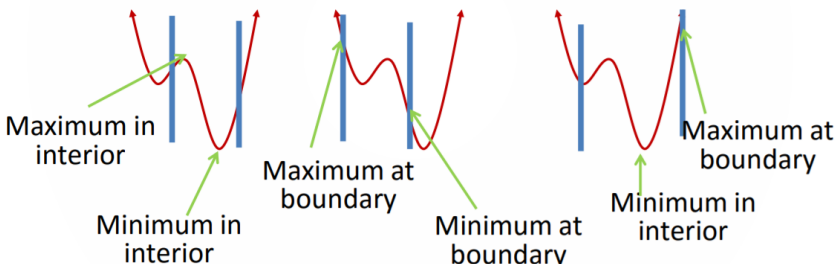
Georgian

Learning Rate

- Learning rate is optimal, model converges to the minimum
- Learning rate is too small, it takes more time but converges to the minimum
- Learning rate is higher than the optimal value, it overshoots but converges.
- Learning rate is very large, it overshoots and diverges, moves away from the minima, performance decreases on learning

Constrained Optimization

- In the previous slides, most of the functions we examined were unconstrained, meaning they either had no boundaries, or the boundaries were soft.
- A constraint is a hard limit placed on the value of a variable, which prevents us from going forever in certain directions.



Georgian

Examples

- If you are attempting to maximize the objective function, typical constraints might involve time, money, and resources. The amounts of these things are limited, and these limits also place limits on the best possible value of the objective function.
- if a single gizmo costs 4, then two gizmos cost 8, five gizmos cost 20, and g gizmos cost 4g. If you buy g gizmos at 4 and s sprockets at 2, then your total cost will be 4g + 2s. If you only have 70 dollars to spend at the gizmo-and sprocket store, then your total cost must be

$$4g + 2s \leq 70 \tag{2}$$

Graphical Method

- An airline offers coach and first-class tickets. For the airline to be profitable, it must sell a minimum of 25 first-class tickets and a minimum of 40 coach tickets. The company makes a profit of 225 dollar for each coach ticket and 200 dollar for each first-class ticket. At most, the plane has a capacity of 150 travelers. How many of each ticket should be sold in order to maximize profits?

Fundamental Theorem of Linear Programming

- If a solution exists to a bounded linear programming problem, then it occurs at one of the corner points.
- If a feasible region is unbounded, then a maximum value for the objective function does not exist.
- If a feasible region is unbounded, and the objective function has only positive coefficients, then a minimum value exist.

Thank You

Georgian

Bayes Theorem and its properties

Garima Malik

February 14, 2022

Outline

Conditional Probability

Bayes Theorem

Classification using Bayes Theorem

Naive Bayes Classifier

Laplace Correction

Conditional Probability

- Conditional probability is the probability of one event occurring with some relationship to one or more other events.
- $P(A|B) = P(AandB)/P(B)$ where P(B) is not equals to 0.
- $P(A|B) = P(A, B)/P(B)$
- $P(A|B) = P(A \bigcap B)/P(B)$

Example

- In a group of 100 sports car buyers, 40 bought alarm systems, 30 purchased bucket seats, and 20 purchased an alarm system and bucket seats. If a car buyer chosen at random bought an alarm system, what is the probability they also bought bucket seats?
- What is the probability a randomly selected person is male, given that they own a pet?

	Have pets	Do not have pets	Total
Male	0.41	0.08	0.49
Female	0.45	0.06	0.51
Total	0.86	0.14	1

Conditional Probability properties

- The joint probability can be calculated using the conditional probability; for example:
  - $P(A, B) = P(A|B) * P(B)$
- This is called the product rule. Importantly, the joint probability is symmetrical, meaning that:
  - $P(A, B) = P(B, A)$
- The conditional probability is not symmetrical; for example:
  - $P(A|B) \neq P(B|A)$

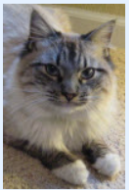
Bayes Theorem

- Bayes Theorem: Principled way of calculating a conditional probability without the joint probability.
  - $P(A|B) = P(B|A) * P(A)/P(B)$
  - $P(B) = P(B|A) * P(A) + P(B|notA) * P(notA)$

Examples

Example: Allergy or Not?

Hunter says she is itchy. There is a test for Allergy to Cats, but this test is not always right:

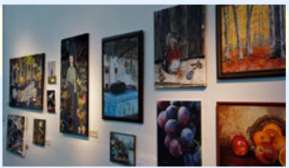


- For people that **really do** have the allergy, the test says "Yes" **80%** of the time
- For people that **do not** have the allergy, the test says "Yes" **10%** of the time ("false positive")

If 1% of the population have the allergy, and **Hunter's test says "Yes"**, what are the chances that Hunter really has the allergy?

Examples

Example: The Art Competition has entries from three painters: Pam, Pia and Pablo



- Pam put in 15 paintings, 4% of her works have won First Prize.
- Pia put in 5 paintings, 6% of her works have won First Prize.
- Pablo put in 10 paintings, 3% of his works have won First Prize.

What is the chance that Pam will win First Prize?

Naming the Terms in the Theorem

- Firstly, in general, the result P(A|B) is referred to as the posterior probability and P(A) is referred to as the prior probability.
- Sometimes P(B|A) is referred to as the likelihood and P(B) is referred to as the evidence.
- This allows Bayes Theorem to be restated as:
  - Posterior = Likelihood * Prior / Evidence

Binary Classification

$$P(C|x) = \frac{P(C)P(x|C)}{P(x)}$$

- $P(C)$: prior
- $P(x|C)$: likelihood
- $P(x)$: evidence
- $P(C|x)$: posterior

$$P(C = 0) + P(C = 1) = 1$$

$$P(x) = P(x|C = 1)P(C = 1) + P(x|C = 0)P(C = 0)$$

$$P(C = 0|x) + P(C = 1|x) = 1$$

Multi Label Classification

$$P(C_i|x) = \frac{P(C_i)P(x|C_i)}{P(x)}$$

$$= \frac{P(C_i)P(x|C_i)}{\sum_{k=1}^{K} P(C_k)P(x|C_k)}$$

$$P(C_i) \geq 0 \text{ and } \sum_{i=1}^{K} P(C_i) = 1$$

$$\text{choose } C_i \text{ if } P(C_i|x) = \max_k P(C_k|x)$$

Prior Probability

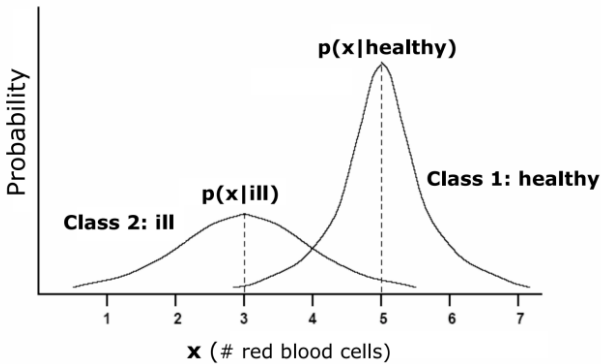
- A prior probability: Knowledge we have about one class before carrying out an experiment.
  - $P(c_i) = \pi_i$ where i = 1 ..... N
- Decision Rule : using only prior probabilities
  - Decide c_1 if $\pi_1 > \pi_2$
  - Decide c_2 if $\pi_2 > \pi_1$
- The probability of having an error in classification is the lower value of π_1 and π_2.

Classification with Prior Probability

- **Classification problem: Discriminate between healthy people and people with anemia.**
- Prior knowledge:
  - 90% of the people is healthy: $\pi_1 = 0.9$
  - 10% of the people is ill: $\pi_1 = 0.1$
- If we have to classify a new patient, which is his/her class?
  - Decide c_1 as $\pi_1 > \pi_2$
- If we have no other information, we have to take this decision. However, nobody will be satisfied if the doctor would decide the state of health without a checkup (or a blood test). It is necessary to use more information:
  - we need measurements relative to the patterns.

Classification with Class Conditional Probability

- Blood test reveals amount of red blood cells.
- The amount of red blood cells is the random variable (x) (We do not have the same number of red blood cells than other people).
- This variable has a Gaussian distribution.



Georgian

Classification with Class Conditional Probability

Classification problem: Discriminate between healthy people and people with anemia.

- Blood test: 4.5 million red blood cells.
- The patient is healthy.
  - $P(x = 4,500,000|c = \text{healthy}) > P(x = 4,500,000|c = \text{ill})$
- If we consider the patient is healthy, the probability he has 4.5 million red blood cells is higher than if we consider he is ill, with the given number of red blood cells

Classification with Class Conditional Probability

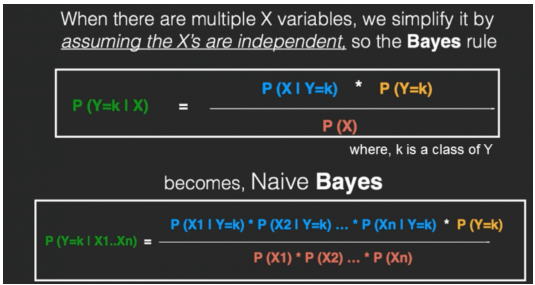
- Bayes Decision Rule :

$$\text{Decide } c_1 \text{ if } P(c_1|x) > P(c_2|x) \left(\text{ or } P(x|c_1)P(x_1) > P(x|c_2)P(x_2) \right)$$

$$\text{Decide } c_2 \text{ if } P(c_2|x) > P(c_1|x) \left(\text{ or } P(x|c_2)P(x_2) > P(x|c_1)P(x_1) \right)$$

Naive Bayes Classifier

- The Bayes Rule provides the formula for the probability of Y given X. But, in real-world problems, you typically have multiple X variables. When the features are independent, we can extend the Bayes Rule to what is called Naive Bayes. It is called 'Naive' because of the naive assumption that the X's are independent of each other. Regardless of its name, it's a powerful formula.



Georgian

Example

- Say you have 1000 fruits which could be either ‘banana’, ‘orange’ or ‘other’. These are the 3 possible classes of the Y variable. We have data for the following X variables, all of which are binary (1 or 0).
  - Long
  - Sweet
  - Yellow

Fruit	Long (x1)	Sweet (x2)	Yellow (x3)
Orange	0	1	0
Banana	1	0	1
Banana	1	1	1
Other	1	1	0
..	..	..	..

Example

- For the sake of computing the probabilities, let's aggregate the training data to form a counts table like this.

Type	Long	Not Long	Sweet	Not Sweet	Yellow	Not Yellow	**Total**
Banana	400	100	350	150	450	50	500
Orange	0	300	150	150	300	0	300
Other	100	100	150	50	50	150	200
Total	500	500	650	350	800	200	1000

Laplace Correction

- The value of P(Orange | Long, Sweet and Yellow) was zero in the above example, because, P(Long | Orange) was zero. That is, there were no ‘Long’ oranges in the training data.
- It makes sense, but when you have a model with many features, the entire probability will become zero because one of the feature’s value was zero. To avoid this, we increase the count of the variable with zero to a small value (usually 1) in the numerator, so that the overall probability doesn’t become zero.
- This correction is called ‘Laplace Correction’. Most Naive Bayes model implementations accept this or an equivalent form of correction as a parameter.

Laplace Correction

- It is also called zero frequency problem
- It is applied on categorical values.
- The generalised formula can be stated as :

$$P(X_i = v_j | C_k) = \frac{n_{ijk} + \lambda}{n_k + \lambda k} \tag{1}$$

- n_ijk is no of examples in C_k where $X_i = v_j$
- n_k is total number of examples in k class
- λ is usually 1
- k is no of classes

Gaussian Naive Bayes

So far we've seen the computations when the X's are categorical. But how to compute the probabilities when X is a continuous variable?

- If we assume that the X follows a particular distribution, then you can plug in the probability density function of that distribution to compute the probability of likelihoods.
- If you assume the X's follow a Normal or Gaussian Distribution, which is fairly common, we substitute the corresponding probability density of a Normal distribution and call it the Gaussian Naive Bayes. You need just the mean and variance of the X to compute this formula.

$$P(X|Y=c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} e^{\frac{-(x-\mu_c)^2}{2\sigma_c^2}}$$

Thank You