Pfa assignment of worksheet 7 solutions

Statistics

1)b

2)d

3)c

4)b

5)a

6)a

7)c

8)b

9)a

10)a

11)a

12)

13)

14)

15)b

Sql

1)b

2)b

3)c

4)c

5)c

6)c

7)c

8)b

9)b

10)a

11) A SQL Join statement is used to combine data or rows from two or more tables based on a common field between them.

12) **Different Types of SQL JOINs**

* (INNER) **JOIN** : Returns records that have matching values in both tables.
* LEFT (OUTER) **JOIN** : Returns all records from the left table, and the matched records from the right table.
* RIGHT (OUTER) **JOIN** : Returns all records from the right table, and the matched records from the left table.

13) SQL Server is a relational database management system developed by Microsoft. As a database server, it is a software product with the primary function of storing and retrieving data as requested by other software applications—which may run either on the same computer or on another computer across a network

14) The PRIMARY KEY constraint uniquely identifies each record in a table. Primary keys must contain UNIQUE values, and cannot contain NULL values. A table can have only ONE primary key; and in the table, this primary key can consist of single or multiple columns (fields).

15)ETL stands for Extract, Transform and Load, which is a process used to collect data from various sources, transform the data depending on business rules/needs and load the data into a destination database.

Machine learning

1)d

2)d

3)b

4)c

5)c

6)c

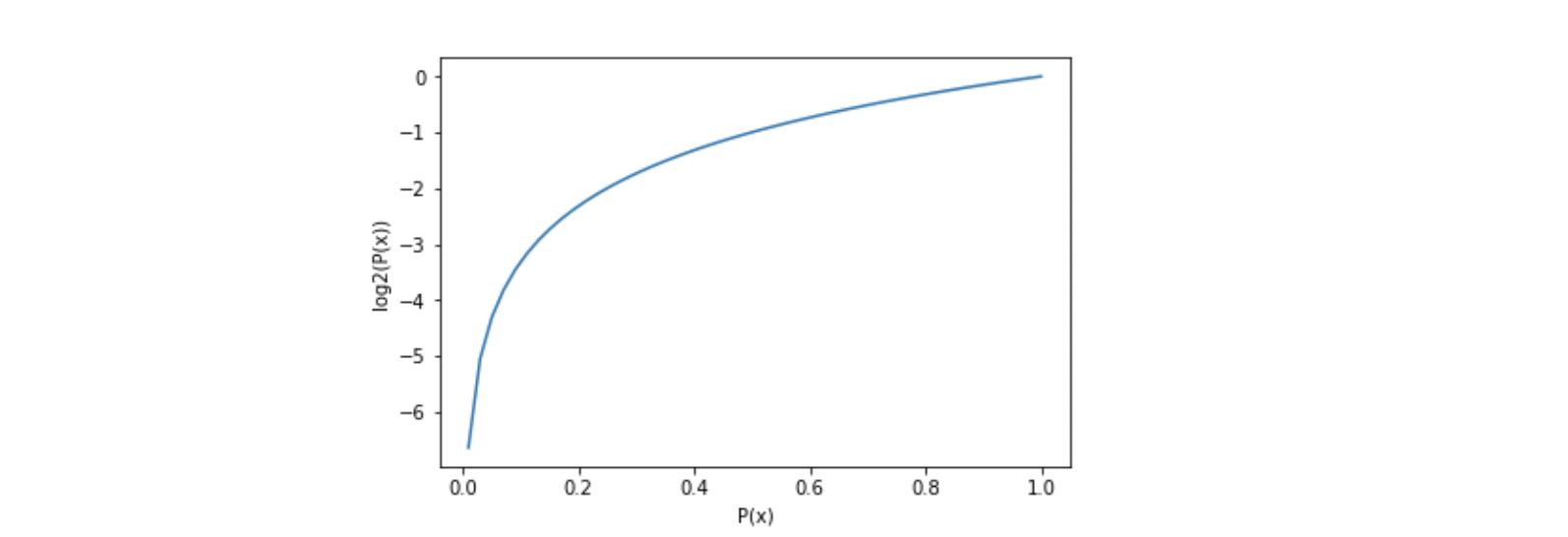
7)a

8)a

9) **Entropy**: It is used to measure the impurity or randomness of a dataset. Imagine choosing a yellow ball from a box of just yellow balls (say 100 yellow balls). Then this box is said to have 0 entropy which implies 0 impurity or total purity.

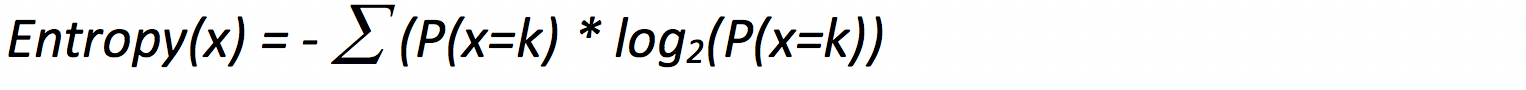
Now, let’s say 30 of these balls are replaced by red and 20 by blue. If we now draw another ball from the box, the probability of drawing a yellow ball will drop from 1.0 to 0.5. Since the impurity has increased, entropy has also increased while purity has decreased. Shannon’s entropy model uses the logarithm function with base 2 (log2(P(x)) to measure the entropy because as the probability P(x) of randomly drawing a yellow ball increases, the result approaches closer to binary logarithm 1 as shown in the graph below.





When a target feature contains more than one type of element (balls of different colors in a box), it is useful to sum up the entropies of each possible target value and weigh it by the probability of getting these values assuming a random draw. This finally leads us to the formal definition of Shannon’s entropy which serves as the baseline for the information gain calculation:





Where P(x=k) is the probability that a target feature takes a specific value, k.

Logarithm of fractions gives a negative value and hence a ‘-‘ sign is used in entropy formula to negate these negative values. The maximum value for entropy depends on the number of classes.

* 2 classes: Max entropy is 1
* 4 Classes: Max entropy is 2
* 8 Classes: Max entropy is 3
* 16 classes: Max entropy is 4

**Information Gain**: To find the best feature which serves as a root node in terms of information gain, we first use each descriptive feature and split the dataset along the values of these descriptive features and then calculate the entropy of the dataset. This gives us the remaining entropy once we have split the dataset along the feature values. Then, we subtract this value from the originally calculated entropy of the dataset to see how much this feature splitting reduces the original entropy which gives the information gain of a feature and is calculated as:



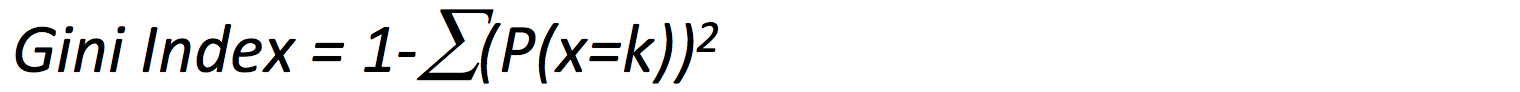


* The feature with the largest information gain should be used as the root node to start building the decision tree.

ID3 algorithm uses information gain for constructing the decision tree.

**Gini Index:** It is calculated by subtracting the sum of squared probabilities of each class from one. It favors larger partitions and easy to implement whereas information gain favors smaller partitions with distinct values.





A feature with a lower Gini index is chosen for a split.

10)Random forests consist of multiple single trees each based on a random sample of the training data. They are typically more accurate than single decision trees. The following figure shows the decision boundary becomes more accurate and stable as more trees are added

11) The most common techniques of feature scaling are Normalization and Standardization. Normalization is used when we want to bound our values between two numbers, typically, between [0,1] or [-1,1]. While Standardization transforms the data to have zero mean and a variance of 1, they make our data unitless.

12) Feature scaling is an idea that makes sure that the features involved in the gradient descent computation are on the similar scale. This implies that the features involved should take similar ranges of values. Feature scaling aims to speed up the process of convergence of gradient descent.

13)About the challenge of choosing metrics for classification, and how it is particularly difficult when there is a skewed class distribution.

How there are three main types of metrics for evaluating classifier models, referred to as rank, threshold, and probability.

How to choose a metric for imbalanced classification if you don’t know where to start.

14) the traditional **F** measure is calculated as follows: **F**-Measure = (2 \* Precision \* Recall) / (Precision + Recall)

15) The fit() function calculates the values of these parameters. The transform function applies the values of the parameters on the actual data and gives the normalized value. The fit\_transform() function performs both in the same step. Note that the same value is got whether we perform in 2 steps or in a single step.