BHARATHIDASAN UNIVERSITY

Tiruchirappalli- 620024

Tamil Nadu, India.

Programme: M.Sc. Statistics

Course Title: Introduction to Machine learning

Course Code: 20ST01VAC

Unit-IV

Chi-Square Automatic Interaction Detector

Dr. T. Jai Sankar
Associate Professor and Head
Department of Statistics

Ms. E. Devi
Guest Faculty
Department of Statistics

UNIT IV

Chi-square Automatic Interaction Detector:

CHAID (Ch**i-square** A**utomatic** I**nteraction** D**etector)** analysis is an algorithm used for discovering relationships between a categorical response variable and other categorical predictor variables. It is useful when looking for patterns in datasets with lots of categorical variables and is a convenient way of summarising the data as the relationships can be easily visualised.

In practice, CHAID is often used in direct marketing to understand how different groups of customers might respond to a campaign based on their characteristics.

Example:

Gender, socio-economic status, geographic location, etc. are associated with the response rate achieved. We build a CHAID "tree" showing the effects of different customer characteristics on the likelihood of response.

The process repeats to find the predictor variable on each leaf that is most significantly related to the response, branch by branch, until no further factors are found to have a statistically significant effect on the response (e.g., likelihood of responding to the marketing campaign). The results can be visualised with a so-called tree diagram.

What statistical techniques are used?

As indicated in the name, CHAID uses Person's Chi-square tests of independence, which test for an association between two categorical variables. A statistically significant result indicates that the two variables are not independent, i.e., there is a relationship between them.

Generally a large sample size is needed to perform a CHAID analysis. At each branch, as we split the total population, we reduce the number of observations available and with a small total sample size the individual groups can quickly become too small for reliable analysis.

How does the decision tree work?

There are different algorithm written to assemble a decision tree, which can be utilized by the problem.

A few of the commonly used algorithms are listed below:

- CART
- ID3
- C4.5

• CHAID Now we will explain about CHAID Algorithm step by step. Before that, we will discuss a little bit about chi square.

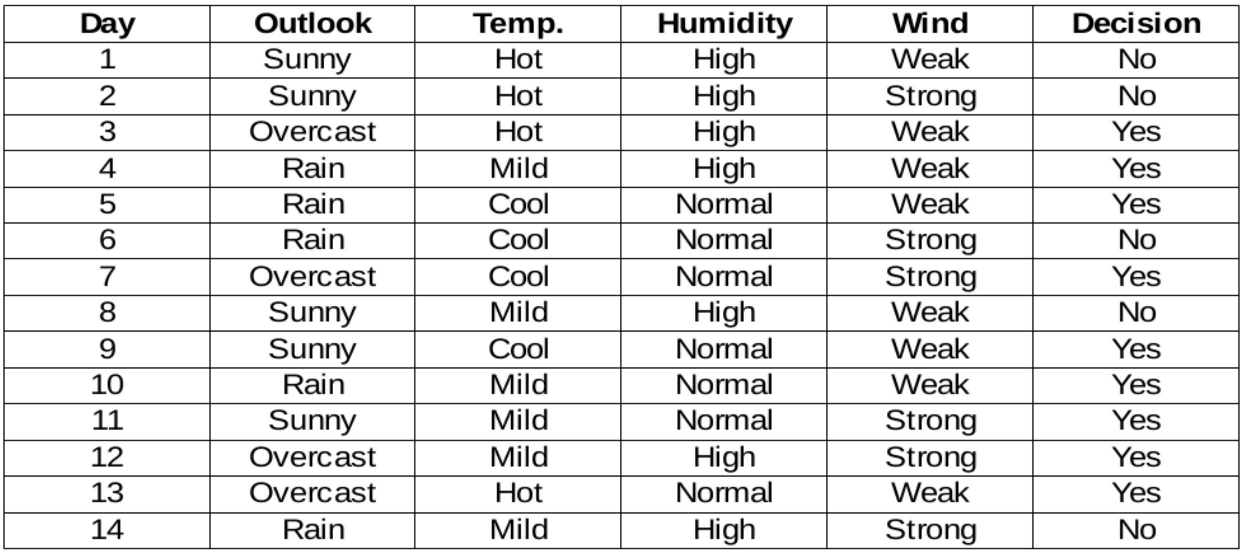
Chi – Square :

Chi-Square is a statistical measure to find the difference between child and parent nodes. To calculate this we find the difference between observed and expected counts of target variable for each node and the squared sum of these standardized differences will give us the Chi-square value.

Formula:

The formula of chi-square:- $\sqrt{((y - y')^2 / y')}$

Where y is actual and y' is expected.



Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

We need to find the most important feature w.r.t target columns to choose the node to split data in this data set.

Humidity feature:

There are two types of the class present in humidity columns such that high and normal. Now we will calculate the chi square values for them.

	Yes	**No**	**Total**	**Expected**	**Chi-square Yes**	**Chi-square No**
High	3	4	7	3.5	0.267	0.267
low	6	1	7	3.5	1.336	1.336

For each row, the total column is the sum of yes and no decisions. **Half of the total column is called Expected values** because there are 2 classes in the decision. It is easy to calculate the chi-squared values based on this table.

For example,

chi-square yes for high humidity is $\sqrt{((3 - 3.5)^2 / 3.5)} = 0.267$

where as actual is 3 and expected is 3.5.

So, the chi-square value of the humidity feature is = 0.267 + 0.267 + 1.336 + 1.336 = 3.207

The feature having the maximum chi-square value will be the decision point.

Wind feature:

There are two types of the class present in wind columns such that weak and strong. The following table is the below table.

	Yes	No	Total	Expected	Chi-square Yes	Chi-square No
Weak	5	2	7	3.5	0.802	0.802
Strong	3	3	6	3	0.000	0.000

The chi-square test value of the wind feature is= 0.802 + 0.802 + 0 + 0

= 1.604

This is less value than the chi-square value of humidity as well.

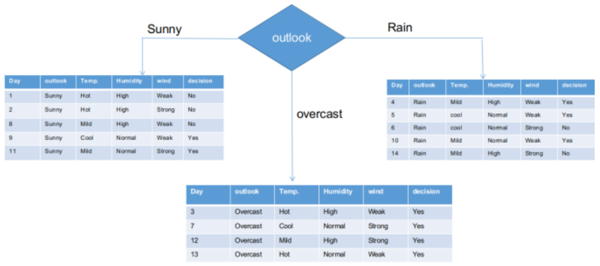
Temperature feature:

There are three types of the class present in temperature columns such that hot, cool and mild. The following table is the below table.

	Yes	No	Total	Expected	Chi-square Yes	Chi-square No
Hot	2	2	4	2	0	0
Mild	4	2	6	3	0.577	0.577
Cool	3	1	4	2	0.707	0.707

The chi-square test value of the temperature feature is = 0 + 0 + 0.577 + 0.577 + 0.707 + 0.707 = 2.569

This is less value than the chi-square value of humidity and greater than the chi square value of wind as well.



Day	outlook	Temp.	Humidity	wind	decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes

Day	outlook	Temp.	Humidity	wind	decision
4	Rain	Mild	High	Weak	Yes
5	Rain	cool	Normal	Weak	Yes
6	Rain	cool	Normal	Strong	No
10	Rain	Mild	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Day	outlook	Temp.	Humidity	wind	decision
3	Overcast	Hot	High	Weak	Yes
7	Overcast	Cool	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes

Outlook feature:

There are three types of a class present in temperature columns such that sunny, rain, and overcast. The following table is the below table.

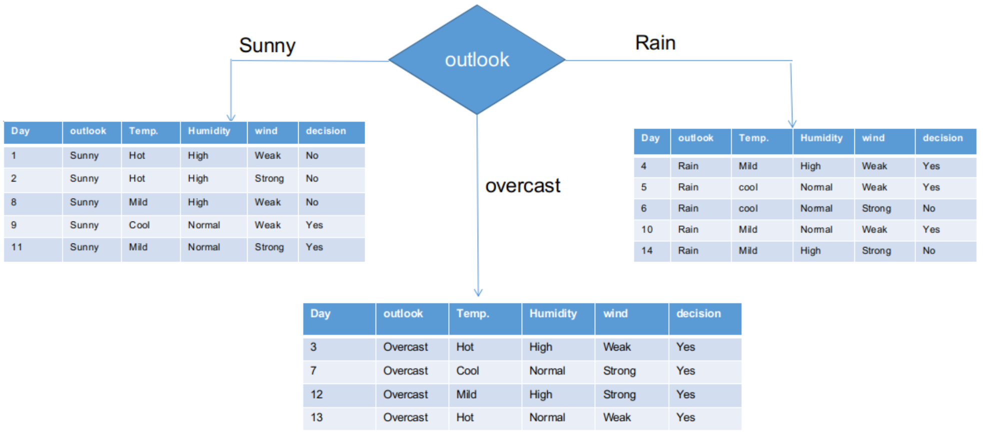
	Yes	No	Total	Expected	Chi-square Yes	Chi-square No
Sunny	2	3	5	2.5	0.316	0.316
Overcast	4	0	4	2	1.414	1.414
Rain	3	2	5	2.5	0.316	0.316

The chi-square test value of the outlook feature is = 0.316 + 0.316 + 1.414 + 1.414 + 0.316 + 0.316 = 4.092

We have calculated the chi-square values of all features.

Feature	Chi-square value
Outlook	4.092
Temperature	2.569
Humidity	3.207
Wind	1.604

The outlook column has the most elevated and highest chi-square value. This implies that it is the main component feature. Along with these values, we will put this feature to the root node.



Day	outlook	Temp.	Humidity	wind	decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes

Day	outlook	Temp.	Humidity	wind	decision
4	Rain	Mild	High	Weak	Yes
5	Rain	cool	Normal	Weak	Yes
6	Rain	cool	Normal	Strong	No
10	Rain	Mild	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Day	outlook	Temp.	Humidity	wind	decision
3	Overcast	Hot	High	Weak	Yes
7	Overcast	Cool	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes

Classification and Regression Trees:

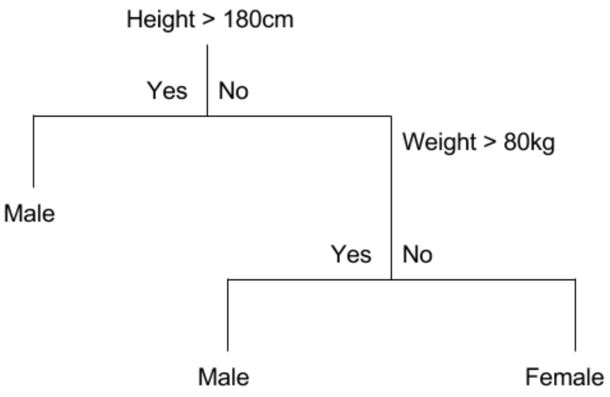
(i) Classification Trees

It is an algorithm where the target variable is always fixed or categorical. The algorithm is then used to identify the “class” within which a target variable would presumably fall into. An example of a classification-type problem would be determining who will or won’t subscribe to a digital platform; or who will or won’t graduate from high school.

These are samples of simple binary classifications where the specific variable can assume just one of two, mutually exclusive values.

In other cases, you would possibly need to predict among a variety of various variables. as an example, you’ll need to predict which sort of smartphone a consumer may plan to purchase. In such cases, there are multiple values for the specific variable. Here’s what a classic classification tree seems like.

(i)

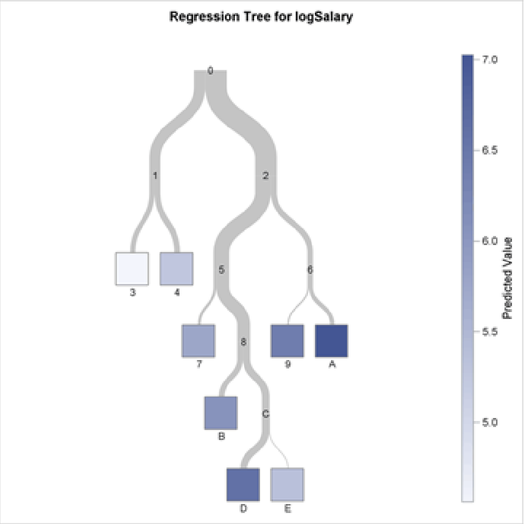


(ii)

(ii) Regression Trees:

A regression tree refers to an algorithm where the target variable is and therefore the algorithm is employed to predict it’s value. As an example of a regression type problem, you’ll want to predict the selling prices of a residential house, which may be a continuous variable.

This will depend upon both continuous factors like square footage also as categorical factors just like the sort of home, area during which the property is found then on.



Impurity measures:

  - The node impurity is a measure of the homogeneity of the labels at the node. The current implementation provides two impurity measures for classification.
  - (Gini impurity and entropy) and one impurity measure for regression (variance). fi is the frequency of label i at a node and C is the number of unique labels.
  - One way to measure impurity degree is **using entropy**.
  - The logarithm is base 2. Entropy of a pure table (consist of single class) is zero because the probability is 1 and log (1) = 0.
- Entropy reaches maximum value when all classes in the table have equal probability.

Gini Index:

The Gini Index is a summary measure of income inequality. The Gini coefficient incorporates the detailed shares data into a single statistic, which summarizes the dispersion of income across the entire income distribution. The Gini coefficient ranges from 0, indicating perfect equality (where everyone receives an equal share), to 1, perfect inequality (where only one recipient or group of recipients receives all the income).

Mathematically, The Gini Index is represented by

$$G = \sum_{i=1}^{C} p(i) * (1 - p(i))$$

The Gini Index works on categorical variables and gives the results in terms of "success" or "failure" and hence performs only binary split. It isn't computationally intensive as its counterpart – Information Gain.

Basic Mechanism:

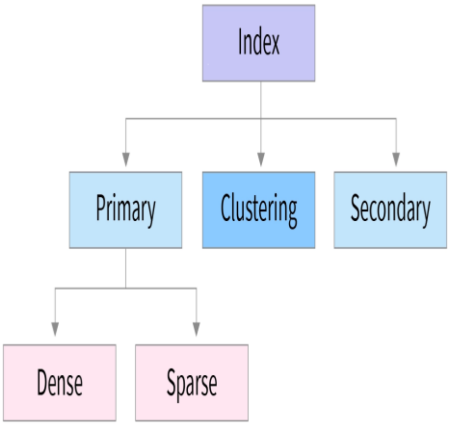
First, we shall randomly pick up any data point from the dataset

Then, we will classify it randomly according to the class distribution in the given dataset. In our dataset, we shall give a data point chosen with a probability of 5/10 for red and 5/10 for blue as there are five data points of each colour and hence the probability.

There are various algorithms designed for different purposes in the world of machine learning. The problem lies in identifying which algorithm to suit best on a given dataset. The decision tree algorithm seems to show convincing results too. To recognize it, one must think that decision trees somewhat mimic human subjective power.

Ordered Indexing:

The indices are stored in a sorted manner hence it is also known as ordered indices. Ordered Indexing is further divided into two categories: Dense Indexing: In dense indexing, the index table contains records for every search key value of the database



Types of indexes:

- Unique indexes enforce the constraint of uniqueness in your index keys.
- Bidirectional indexes allow for scans in both the forward and reverse directions.
- Clustered indexes can help improve the performance of queries that traverse the table in key order.
- Expression-based indexes efficiently evaluate queries with the indexed expression.

Unique and non-unique indexes

Unique indexes are indexes that help maintain data integrity by ensuring that no rows of data in a table have identical key values

Clustered and non-clustered indexes

Index architectures are classified as clustered or non-clustered. Clustered indexes are indexes whose order of the rows in the data pages corresponds to the order of the rows in the index. This order is why only one clustered index can exist in any table, whereas, many non-clustered indexes can exist in the table. In some database systems, the leaf node of the clustered index corresponds to the actual data, not a pointer to data that is found elsewhere.

Partitioned and non partitioned indexes

Partitioned data can have indexes that are partitioned and non partitioned.

Bidirectional indexes

Bidirectional indexes allow scans in both the forward and reverse directions.

Expression-based indexes

With expression-based indexes, you can create an index that includes expressions. The performance of queries that involves expressions is improved if the database manager chooses an index that is created on the same expressions.

*****