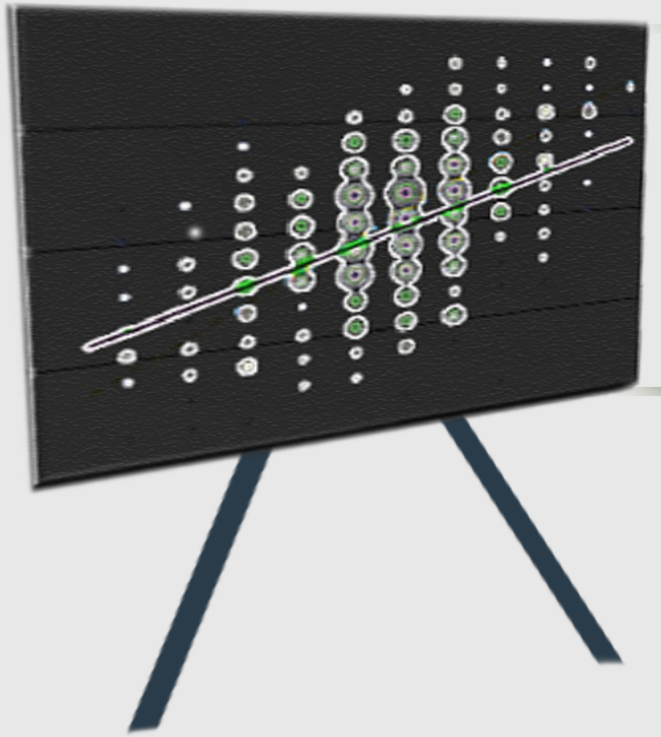


Classification Problems

U. DINESH KUMAR

IIMBX

IIMB
तेजस्वि नावधीतमस्तु

Decision Trees

- Introduction to Decision Trees.
- Decision Tree approach for classification.
- Chi-Square Automatic Interaction Detection (CHAID)
- Classification and Regression Tree (CART)

CHAID (Chi-square Automatic Interaction Detection)

Introduction to CHAID

- CHAID is a decision tree algorithm used in classification problems.
- Initial models of CHAID used chi-square test of independence for splitting.
- CHAID was first presented in an article titled, "An exploratory technique for Investigating large quantities of categorical variables", by G V Kass in Applied Statistics (1980)

CHAID

- CHAID partitions the data into mutually exclusive, exhaustive, subsets that best describe the dependent categorical variable.

- CHAID is an iterative procedure that examines the predictors (or classification variables) and uses them in the order of their statistical significance.

CHAID Splitting Rule

- The splitting in CHAID is based on the type of the dependent variable (or target variable).
  - For continuous dependent variable F test is used.
  - For categorical dependent variable chi-square test of independence is used.

CHAID Example

- German Credit Rating Data Set.
- Predictor Variable – Checking Account Balance

Chi-Square test of independence

- Chi-square test of independence starts with an assumption that there is no relationship between two variables.

- For example, we assume that there is no relationship between checking account balance and default.

Chi-Square test of Independence
German Credit Case

- H_0: There is no relationship between checking account balance and default.
- H_A: There is a relationship between checking account balance and default.

Contingency Table

Checking account balance	Default		
	1	0	Total
0 DM	135	139	274
Other than 0 DM	165	561	726
Total	300	700	1000

Chi-Square test of Independence in German Credit Data

- H_0: Checking account balance and default are independent.
- H_A: Checking account balance and default are dependent.

Chi-Square test of Independence – Test Statistic

$$\chi^2 - \text{statistic} = \sum\sum\left(\frac{(O-E)^2}{E}\right)$$

$$\text{Expected frequency } E = \frac{\text{row sum} \times \text{column sum}}{\text{total sum}}$$

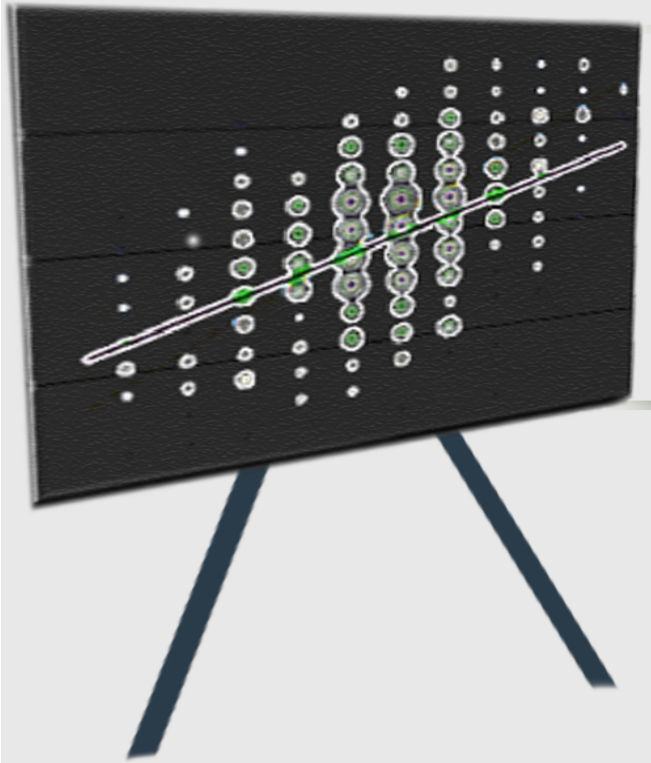
Chi-Square test of Independence

	Observed frequency	Expected Frequency		(O-E)^2/E
0DM-1	135	82.2		33.91
0DM-0	139	191.8		14.53
N0DM-1	165	217.8		12.8
N0DM-0	561	508.2		5.48
	1000	1000		66.73
			P-value	3.10E-16

p-value is less than 0.05, we reject the null hypothesis.

IIMBX

IIMB
तेजस्वि नावधीतमस्तु



CHAID Model

U. DINESH KUMAR

IIMBX

IIMB
तेजस्वि नावधीतमस्तु

CHAID Procedure

- Step 1: Examine each predictor variable for its statistical significance with the dependent variable using F test (for continuous dependent) or chi-square test for categorical dependent).
- Step 2: Determine the most significant among the predictors (predictor with smallest p value after Bonferonni correction).
- Step 3: Divide the data by levels of the most significant predictor Each of these groups will be examined individually further.
- Step 4: For each sub-group, determine the most significant variable from the remaining predictor and divide the data again.
- Step 5: Repeat step 4 till stopping criteria is reached.

Merging - CHAID

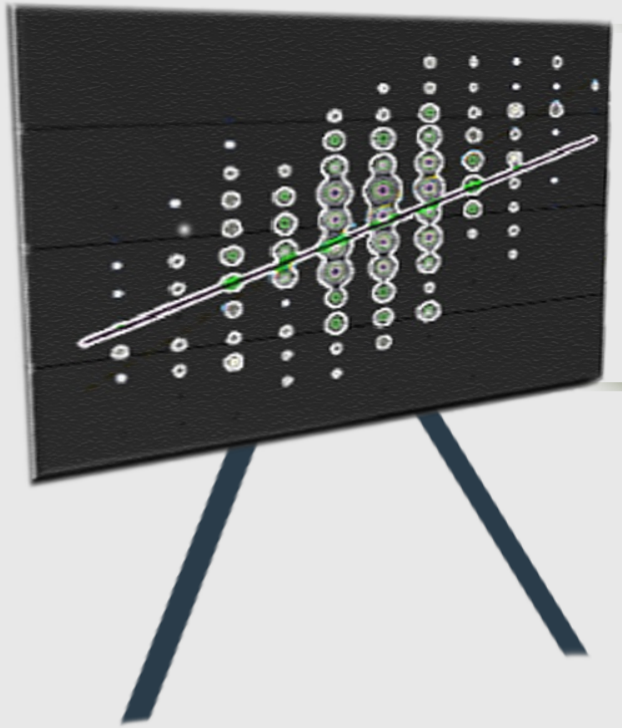
- CHAID uses both splitting and merging steps.
- In merging, least significantly different groups are merged to form one class.

CHAID Stopping Criteria

- Maximum tree depth is reached (which is pre-defined).
- Minimum number of a cases to be a parent node is reached (again pre-defined)
- Minimum number of to be a child node is reached.

CHAID Input

- Significance level for partitioning a variable.
- Significance level for merging.
- Minimum number of records for the cells.



CHAID Working: German Credit Rating

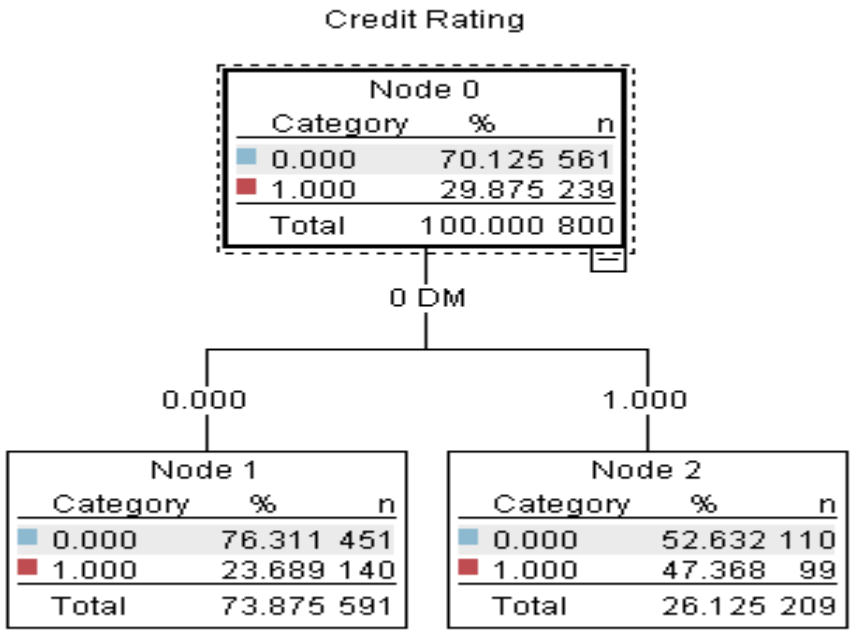
U. DINESH KUMAR

IIMBX

IIMB
तेजस्वि नावधीतमस्तु

CHAID Example: German Credit Data

First Split – ODM



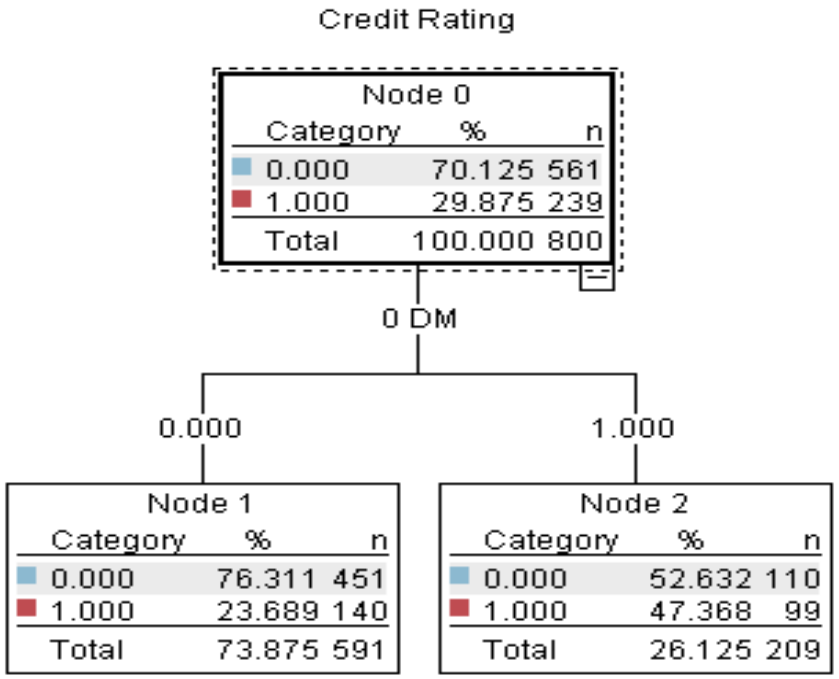
CHAID Tree

	OBSERVED		
Predictor	Response variable		Total
	Y = 1	Y = 0	
0DM = 1	99	110	209
0DM = 0	140	451	591
Total	239	561	800

	Expected		
Predictor	Response Variable		Total
	Y = 1	Y = 0	
0DM = 1	62.43875	146.56	209
0DM = 0	176.5613	414.44	591
Total	239	561	800

P-value = 1.28×10^{-10}

First Split – ODM



CHAID Tree

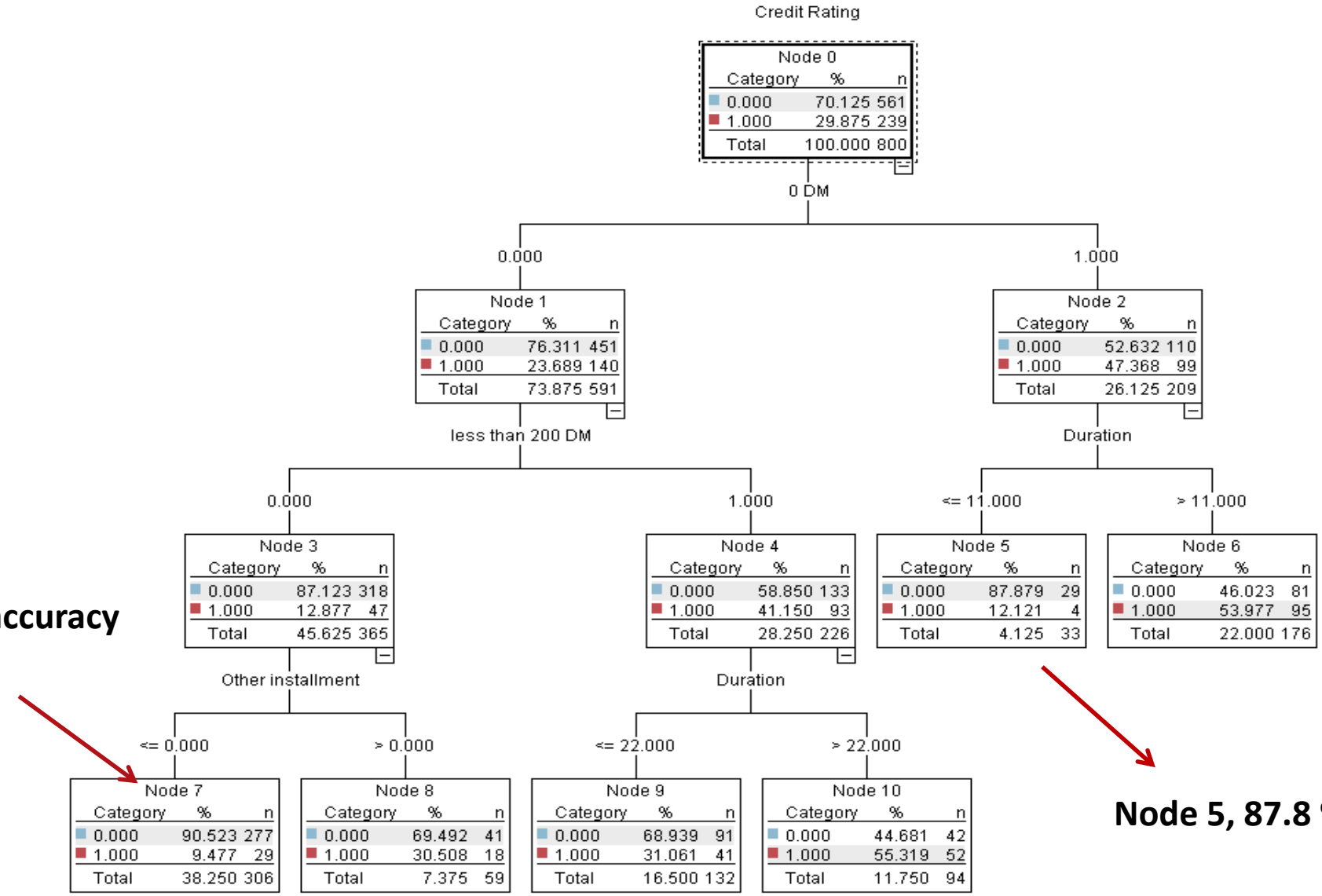
Business Rules

Business rules are always derived for leaf nodes (node without any branch)

Node 1: If 0 DM = 0 then classify the observation as 0 (good credit) which will be accurate 76% of the times

IIMBX

IIMB
तेजस्वि नावधीतमस्तु



Node 7, 90% accuracy

Node 5, 87.8 % accuracy

IIMBX

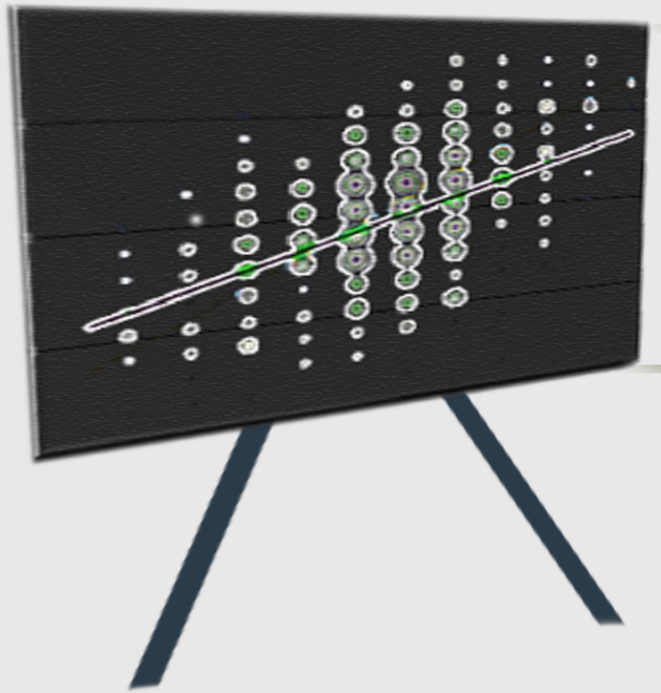
IIMB
तेजस्वि नावधीतमस्तु

Business Rule

- **Node 7**: If the Checking account balance is not less than 200 DM and there are no other installments, then classify the observation as Y = 0.

Tree Pruning

- Tree pruning is a process of reducing the size of the tree using different strategies.
- Pruning is important since large tree can result in over fitting.
- Pruning can be either based on the size of the or purity.



CART Model

U. DINESH KUMAR

IIMBX

IIMB
तेजस्वि नावधीतमस्तु

Classification and Regression Tree (CART)

Classification and Regression Trees (CART)

- Splits are chosen based either on Gini Index or twoing criteria.
- CART is a binary tree, whereas CHAID can split the initial node into more than 2 branches.

Gini Index (Classification Impurity)

- Gini Index is used to measure the impurity at a node (in classification problem) and is given by:

$$\text{Gini(k)} = \sum_{j=1}^{J} P(j|k)(1-P(j|k))$$

where P(j|k) is the proportion of category j in node k

"**Smaller Gini Index implies less impurity**"

Entropy (Impurity Measure)

- Entropy is another impurity measure that is frequently used. Entropy at node "k" is given by:

$$\text{Entropy}(k) = -\sum_{j=1}^{c} P(j \mid k)\ \mathbf{log_2}(P(j \mid k))$$

Gini Index Calculation

Number of classes 2 (say 0 and 1)

Consider node label k with 10 1s and 90 0s.

$$\text{Gini(k)} = \sum_{j=1}^{2} P(j|k)(1-P(j|k)) = 2P(j|k)(1-P(j|k))$$

$$\text{Gini(k)} = 2 \times 0.1 \times 0.9 = 0.18$$

Smaller number implies less impurity!

IIMBX

Entropy Calculation

Number of classes 2 (say 0 and 1)

Consider node label k with 10 1s and 90 0s.

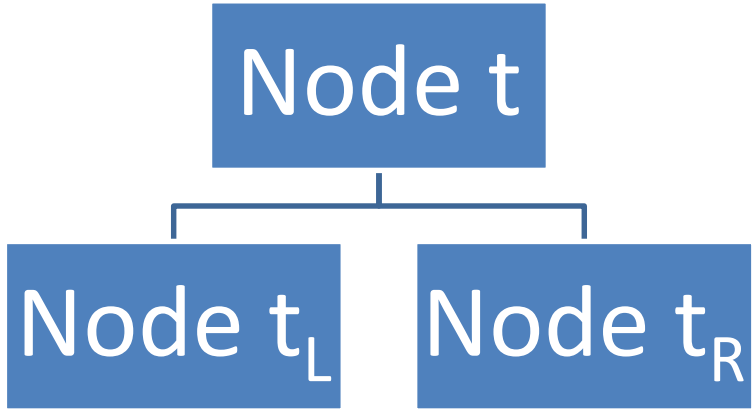
$$\text{Entropy(k)} = -\sum_{j=1}^{2} P(j|k)\log(P(j|k))$$

$$\text{Entropy(k)} = -0.1\times\log(0.1) - 0.9\times\log(0.9) = 0.4689$$

Higher than Gini coefficient

IIMBX

Classification Tree Logic



Reduction in impurity

$$\text{Max}\left[i(t) - N_L \times i(t_L) - N_R \times i(t_L)\right]$$

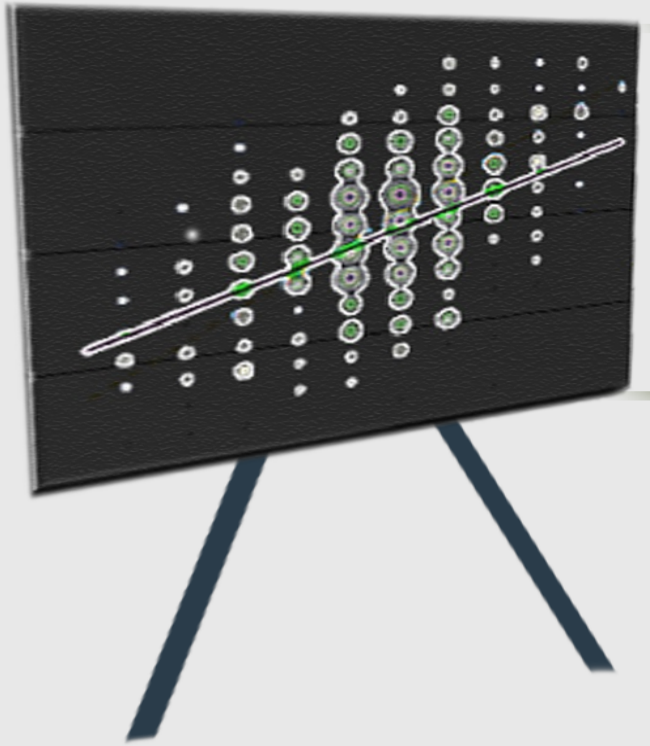
$i(.) = \text{Impurity at node } (.)$

$P_L = \text{Proportion of observations in the left node}$

$P_R = \text{Proportion of observations in the right node}$

IIMBX

IIMB
तेजस्वि नावधीतमस्तु



CART Working: German Credit Rating

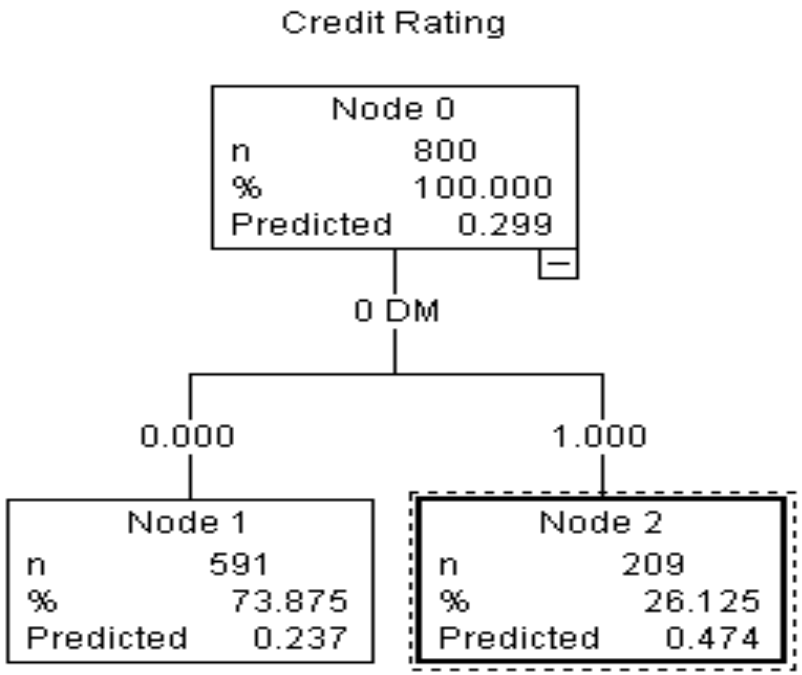
U. DINESH KUMAR

IIMBX

IIMB
तेजस्वि नावधीतमस्तु

CART Example: German Credit Data

First Split – ODM



CART Tree

Impurity at Node 0 = 2 x 0.299 x 0.701 = 0.4191

Impurity at Node 1 = 2 x 0.237 x 0.763 x 0.73875 = 0.2671

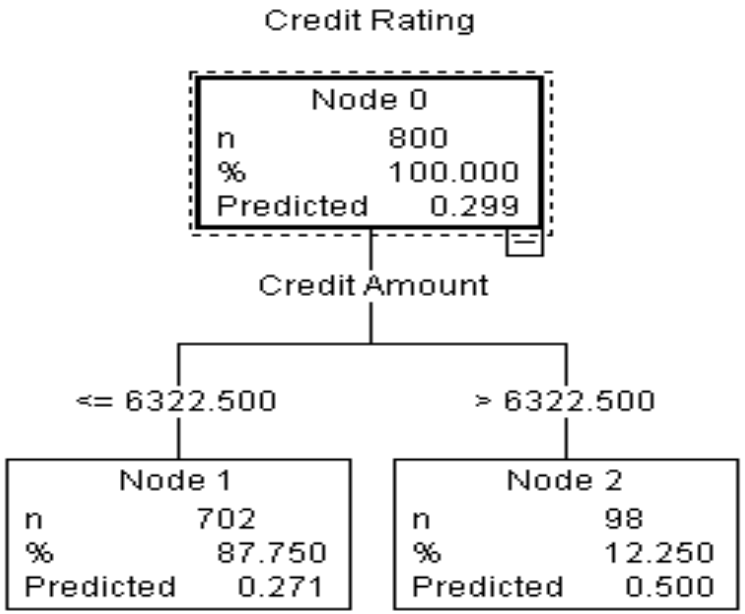
Impurity at Node 2 = 2 x 0.474 x 0.526 x 0.26125 = 0.130272

Change in impurity = 0.4191 – 0.2671 – 0.130271 = 0.0217

IIMBX

IIMB
तेजस्वि नावधीतमस्तु

First Split – Credit Amount



CART Tree

Impurity at Node 0 = 2 x 0.299 x 0.701 = 0.4191

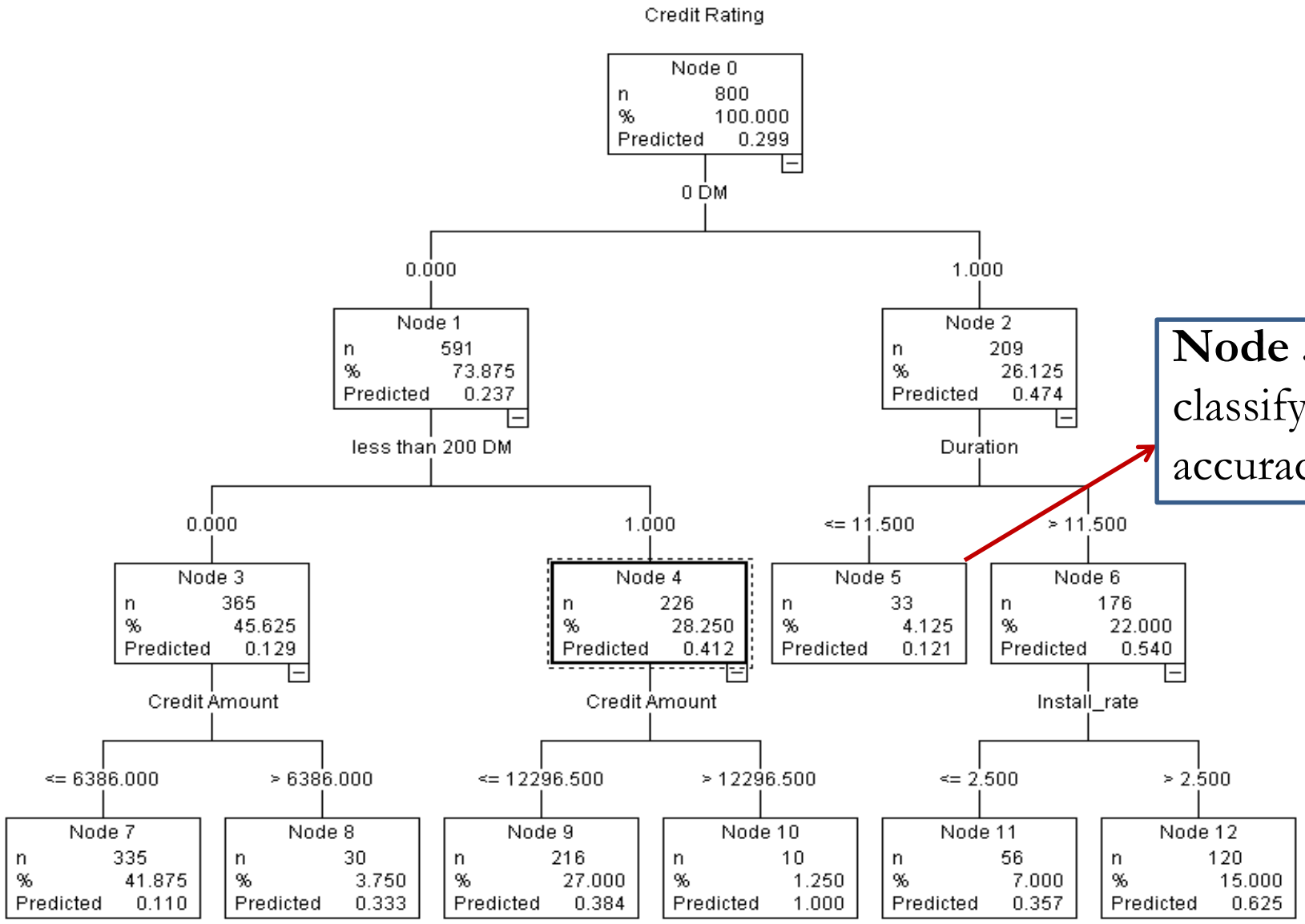
Impurity at Node 1 = 2 x 0.271 x 0.729 x 0.8775 = 0.3467

Impurity at Node 2 = 2 x 0.5 x 0.5 x 0.1225 = 0.06125

Change in impurity = 0.4191 – 0.3467 – 0.06125 = 0.0112

IIMBX

IIMB
तेजस्वि नावधीतमस्तु



Node 5: 0DM = 1 and duration is <= 11.5, classify the observation Y as 0 (88% accuracy)

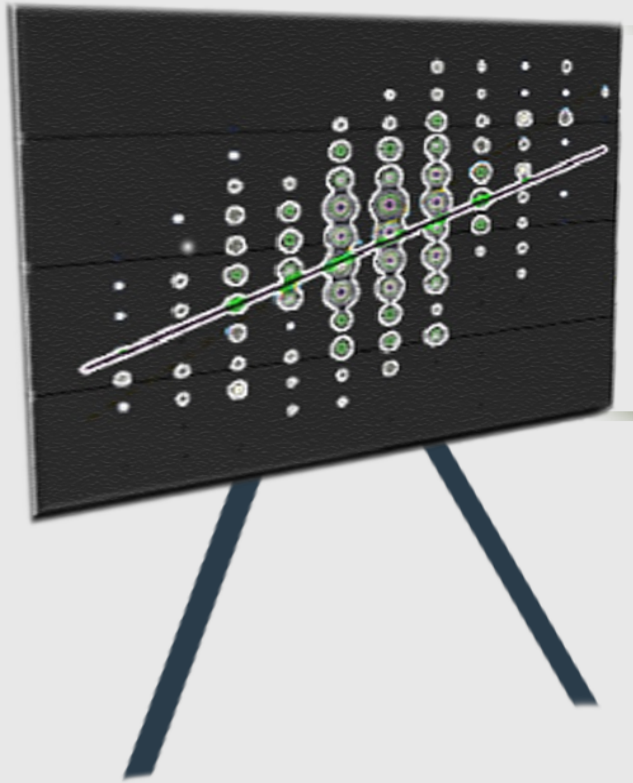
Ensemble Methods

- Ensemble methods is a majority voting method in which models are developed using various techniques such as logistic regression, discriminant analysis, CHAID, CART etc.
- For each observation, the classification is done based on majority voting.
- Ensemble methods in general improve the accuracy of prediction.

Random Forest

- Random forest is an ensemble method in which several trees are generated (thus the name forest) by sampling the predictor variable.

- The sampling can also be on the training data set.



Analysis of Unstructured data

U. DINESH KUMAR

IIMBX

IIMB
तेजस्वि नावधीतमस्तु

Analysis of unstructured data: Text Mining/Sentiment Analysis/Social Media Analytics

IIMBX

IIMB
तेजस्वि नावधीतमस्तु

Why Text Mining and Sentimental Analysis?

- **Science of Social Influence:**
  - 70% of the sales decisions were made before engaging a sales representative (HP).
  - Sales force is becoming irrelevant since customers are engaged through social media.

Importance of Social Listening

- A study by McKinsey revealed that electronic word-of-mouth leads to twice the sales as compared to paid advertising.

Data Sources for unstructured data and Social Media Analytics

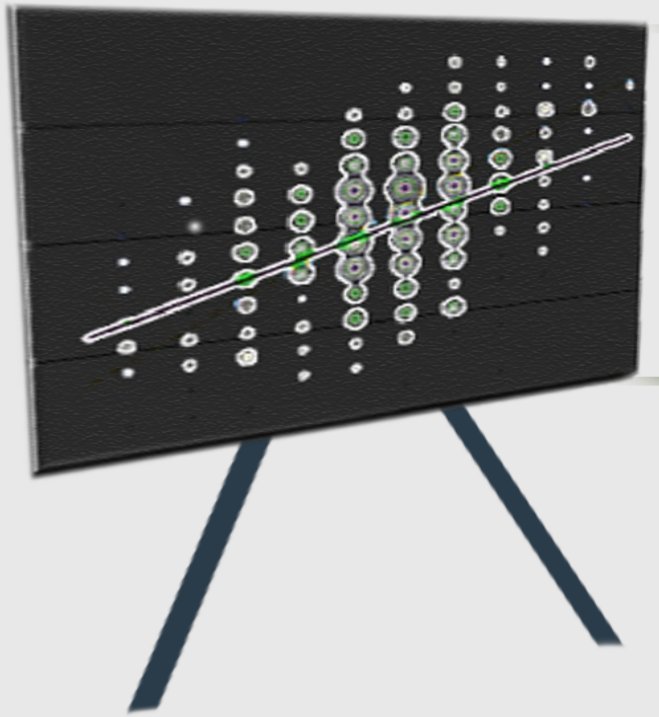
- Facebook
- Blogs and Microsites
- YouTube, Twitter, Instagram

Text Classification and Sentiment Analysis

- Text Classification Approaches
  - Rule based classification
  - Naïve Bayes Algorithm

Classification Methods: Rules

- Rules based on combinations of words present in the documents.
  - Spam e-mail classification
  - ("Hello Dear") OR ("You Won a Lottery")
  - Accuracy can be low



Introduction to Naïve Bayes Algorithm

U. DINESH KUMAR

IIMBX

IIMB
तेजस्वि नावधीतमस्तु

Naïve Bayes Algorithm

Naïve Bayes Algorithm

- Naïve Bayes algorithm (or naïve Bayes classifier) is a classification techniques that uses Bayes theorem.

- One of the popular tools for sentiment analysis.

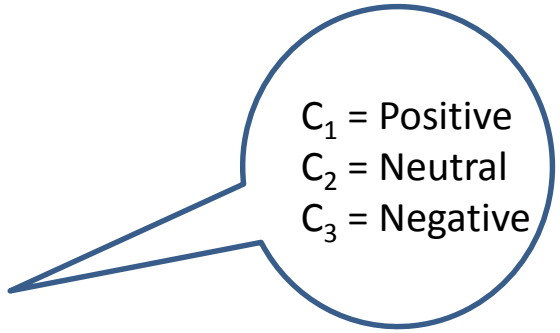
Comments on the Bollywood Movie "Chennai Express"

Some films are hard to make sense. Others are just nonsense. Chennai Express, directed by Rohit Shetty, ticks both boxes.

The film is filled with gigantic men whose size functions as a punch line. Yes, some of it is funny. The locations are beautiful.

Classification Methods: Naïve Bayes Algorithm

- ***Input:***
  - a document set D = {*d1, d2, ….*}
  - a fixed set of classes $C = \{c_1, c_2, \ldots, c_n\}$
  - A training set of k documents $(d_1,c_1),\ldots,(d_k, c_k)$



- ***Output:***
  - a learned classifier Y: d_{new} → c

Naïve Bayes Model

- Documents are treated as bag of words.
- Consider a **document D**, whose **class is given by C**. We classify D as the class which has the highest posterior probability P(C|D).
- P(C|D) can be expressed using **Bayes' theorem**:

$$P(C \mid D) = \frac{P(D \mid C)P(C)}{P(D)} \propto P(D \mid C)P(C)$$

Naïve Bayes Classification for Text Mining

- Document is represented as bag of words.
- We create a vocabulary V containing **N** word types (features).
- Documents are represented using feature vectors.

Vocabulary

V = {Awesome, humor, action, nonsense, romance, stereotype, Mindless , Boring, Overacting, trash }

N = 10 (Number of words in the Vocabulary)

Document Model

- **Bernoulli document model:** A document is represented by a feature vector with binary elements taking value 1 if the corresponding word is present and 0 otherwise.

- **Multinomial document model**: A document is represented by a feature vector with integer whose value is the frequency of that word in the document.

Bernoulli Document Model

- V = {awesome, humor, action, nonsense, romance, stereotype, Mindless, Boring, Overacting, trash}

- Some films are hard to make sense of. Others are just nonsense. Chennai Express, directed by Rohit Shetty, ticks both boxes. More than a quarter of the film is in Tamil, and hence incomprehensible if you're unfamiliar with the language. The rest is a stew of puerile humor, lazy stereotypes, and overacting from a star who appears to be trying too hard.

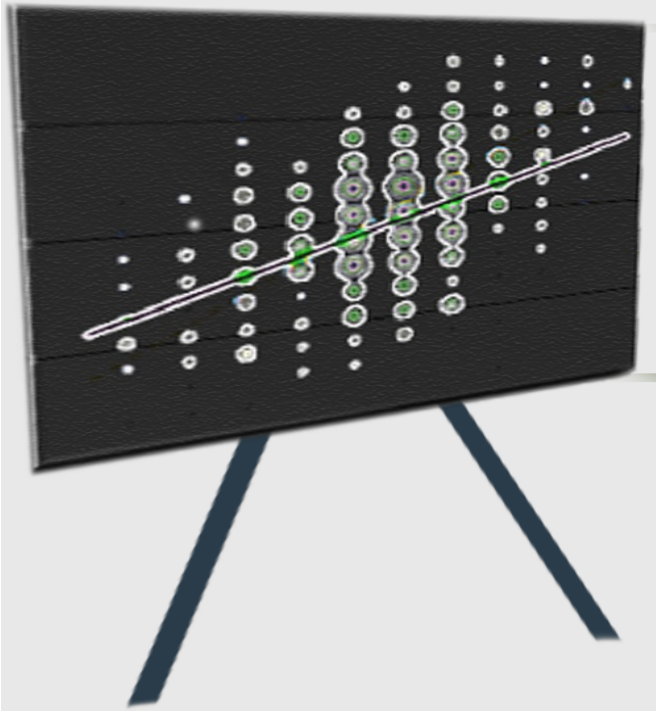
$$D = \{0, 1, 0, 1, 0, 1, 0, 0, 1, 0\}$$

Bernoulli Document Model

- V = {awesome, humor, action, nonsense, romance, stereotype, Mindless, Boring, Overacting, trash}

- Chennai Express plays neither to Rohit's strengths nor to Shah Rukh's. It's a strangely sloppy mishmash of cheesy humour, half-hearted romance, half-baked emotion and head-banging action. The film is filled with gigantic men whose size functions as a punch line. Yes, some of it is funny. The locations are beautiful.

$$D = \{0, 1, 0, 0, 1, 0, 0, 0, 0, 0\}$$



Text Analytics

U. DINESH KUMAR

IIMBX

IIMB
तेजस्वि नावधीतमस्तु

Naïve Bayes Algorithm

Training Data Set (Chennai Express from Bollywood Hungama)

what a awseome movie....start fr the DDLJ train scene..till the climax...too funny...conversation with singing a song..haha mind blowing... Mina washing powder mina...mina..ting tong...halirous man...SRK you proof once again that you are a Baap of acting and bollwood...

i like the movie v.much. its train secquence srk acting in train, converstion betwn SRK and depica by singing hindi songs it is amaizing.its drama,its music and its climax is v.good , proud of such a lover rahol..Ce has good entertainment story love to watch it again & again.

wow mind blowing...superb.. truly enjoy to watch this. CE rocksssssssssssssssss

Training Data Set (Chennai Express from Bollywood Hungama)

trash in one word.

Shame on the film industry. An ugly blot on the blockbuster list. SRK how do u manage to face yourself with films such as this. U used to be an actor of quality with movies such as Daar, swadesh, dil se, He Ram, to back you. The makers of the film have made a mockery out of Indians. And what is worse is we further demean ourselves by going to watch such movies.

Quiet shocking to see actor of Shahrukh s caliber and seniority doing such an ordinary role. Perfomance wise much more expected from Shahrukh .In fact his sarcasm was felt very repetitive and far from funny.Only good thing about movie is Deepika s Perfomance .She has done extraordinary job and should be given the full credit for the success of movie.

Ref: Hiroshi Shimodaira, "Text Classification using Naïve Bayes", Learning and Data Note, University of Edinburgh

Naïve Bayes Model

- P(C|D) can be expressed using **Bayes' theorem**:

$$P(C \mid D) = \frac{P(D \mid C)P(C)}{P(D)} \propto P(D \mid C)P(C)$$

Bernoulli Document Model – Probability Estimation

- If there are N documents in total in the training set, then the prior probability of class C = k may be estimated as the relative frequency of documents of class C=k:

$$\hat{P}(C = k) = \frac{N_k}{N}$$

$$P(C \mid D) = \frac{P(D \mid C)P(C)}{P(D)} \propto P(D \mid C)P(C)$$

Bernoulli Document Model – Probability Estimation

Number of documents = 500

Number of positive documents = 200

$$\hat{P}(C = k) = \frac{N_k}{N} = \frac{200}{500} = 0.4$$

Naïve Bayes Model

- P(C|D) can be expressed using **Bayes' theorem**:

$$P(C \mid D) = \frac{P(D \mid C)P(C)}{P(D)} \propto P(D \mid C)P(C)$$

Documents and converted into feature vectors

b_i= {0, 1, 0, 0, 1, 0, 0, 0, 0, 0}

Bernoulli Document Model

- Let b_i be the feature vector for the i^{th} document D_i

$$b_{it} = \begin{cases} 1 \text{ if word } w_t \text{ is present in the document} \\ 0 \text{ otherwise} \end{cases}$$

Bernoulli Document Model

- $P(D_i | C)$ in terms of the individual word likelihoods $P(w_t | C)$ is given by:

$$P(D_i | C) \approx P(b_i | C) = \prod_{i=1}^{N} [b_{it} P(w_t | C) + (1 - b_{it})(1 - P(w_t | C))]$$

Probability of finding w_t given class C

Probability of not finding w_t given class C

Bernoulli Document Model – Probability Estimation

- Let $n_k(w_t)$ be the number of documents of class $C = k$ in which w_t is observed; and let N_k be the total number of documents of that class.

- Then estimate $P(W_t | C = k)$ is given by:

$$\hat{P}(w_t | C = k) = \frac{n_k(w_t)}{N_k}$$

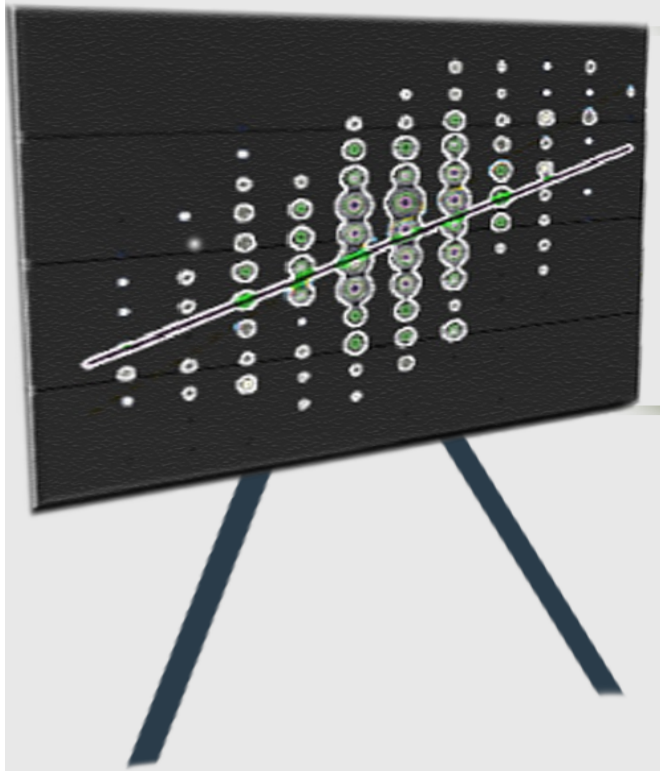
Bernoulli Document Model – Probability Estimation

Class k = Positive

N_K = 200 (total number of positive documents in the training data)

$N_k(w_t)$ = 50 (that is word w_t appears in 50 training documents)

$$\hat{P}(w_t \mid C = k) = \frac{n_k(w_t)}{N_k} = \frac{50}{200} = 0.25$$



Naïve Bayes Algorithm

U. DINESH KUMAR

IIMBX

IIMB
तेजस्वि नावधीतमस्तु

Naïve Bayes Algorithm

V = {awesome, humor, action, nonsense, romance, stereotype, Mindless, Boring, Overacting, trash}

$$T^P = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 1 & 11 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 01 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 & 00 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 11 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 11 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 & 11 & 1 \end{pmatrix}$$

T^P = Documents that are classified as positive sentiment in the training dataset.

$$T^N = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

T^N = Documents that are classified as negative sentiment in the training dataset.

IIMBX

IIMB
तेजस्वि नावधीतमस्तु

Naïve Bayes Model

- P(C|D) can be expressed using **Bayes' theorem**:

$$P(C \mid D) = \frac{P(D \mid C)P(C)}{P(D)} \propto P(D \mid C)P(C)$$

Naïve Bayes Algorithm

$$P(C \mid D) = \frac{P(D \mid C)P(C)}{P(D)} \propto P(D \mid C)P(C)$$

P(P) = 6/12 = 0.5 P(N) = 6/12 = 0.5

$$P(D_i \mid C) \approx P(b_i \mid C) = \prod_{i=1}^{N} \left[b_{it} P(w_t \mid C) + (1 - b_{it})(1 - P(w_t \mid C)) \right]$$

$N_p(w_1) = 4$ $N_p(w_2) = 1$ $N_p(w_3) = 2$ $N_p(w_4) = 3$ $N_p(w_5) = 3$

$N_p(w_6) = 4$ $N_p(w_7) = 4$ $N_p(w_8) = 4$ $N_p(w_9) = 5$ $N_p(w_{10}) = 4$

$N_p(w_t)$ = Number of times word w_t has occurred in positive documents in the training data.

$P(w_1 \mid P) = 4/6$ $P(w_2 \mid P) = 1/6$ $P(w_3 \mid P) = 2/6$ $P(w_4 \mid P) = 3/6$

$P(w_5 \mid P) = 3/6$ $P(w_6 \mid P) = 4/6$ $P(w_7 \mid P) = 4/6$ $P(w_8 \mid P) = 4/6$

$P(w_9 \mid P) = 5/6$ $P(w_{10} \mid P) = 4/6$

$P(w_t \mid P)$ = Conditional probability of observing word w_t given it is a positive sentiment.

IIMBX

IIMB
तेजस्वि नावधीतमस्तु

Naïve Bayes Algorithm

$N_N(w_1) = 2$ $N_N(w_2) = 3$ $N_N(w_3) = 3$ $N_N(w_4) = 1$ $N_N(w_5) = 1$

$N_N(w_6) = 1$ $N_N(w_7) = 3$ $N_N(w_8) = 1$ $N_N(w_9) = 2$ $N_N(w_{10}) = 1$

$N_N(w_t)$ = Number of times word w_t has occurred in negative documents in the training data.

$P(w_1|N) = 2/6$ $P(w_2|N) = 3/6$ $P(w_3|N) = 3/6$ $P(w_4|N) = 1/6$

$P(w_5|N) = 1/6$ $P(w_6|N) = 1/6$ $P(w_7|N) = 3/6$ $P(w_8|N) = 1/6$

$P(w_7|N) = 2/6$ $P(w_8|N) = 1/6$

$P(w_t|N)$ = Conditional probability of observing word w_t given it is a negative sentiment.

IIMBX

Naïve Bayes Algorithm

- Classify the following documents into positive and negative sentiments:

$$D1 = (1, 1, 1, 1, 1, 1, 1, 1, 1, 1)$$

D1 = (1, 1, 1, 1, 1, 1, 1, 1)

$$P(P \mid D_1) \propto P(P) \prod_{i=1}^{|V|} \left[b_{it} P(w_t \mid C) + (1 - b_{it})(1 - P(w_t \mid C)) \right]$$

$$= (6/12) * \left[\frac{4}{6} \times \frac{1}{6} \times \frac{2}{6} \times \frac{3}{6} \times \frac{3}{2} \times \frac{4}{6} \times \frac{4}{6} \times \frac{4}{6} \times \frac{5}{6} \times \frac{4}{6} \right] = 0.046$$

$$P(N \mid D_1) \propto P(N) \prod_{i=1}^{|V|} \left[b_{it} P(w_t \mid C) + (1 - b_{it})(1 - P(w_t \mid C)) \right]$$

$$= (6/12) * \left[\frac{2}{6} \times \frac{3}{6} \times \frac{3}{6} \times \frac{1}{6} \times \frac{1}{2} \times \frac{1}{6} \times \frac{3}{6} \times \frac{1}{6} \times \frac{2}{6} \times \frac{1}{6} \right] = 8.9 \times 10^{-7}$$

The document will be classified as P since it has the higher probability

Challenges

- Difficult to classify sarcastic comments
- All my life I thought Air was free… until I bought a bag of chips.
- You have been so incredibly helpful and thanks (for nothing)

Text Pre-Processing

- Feature Extraction
  - Feature extractor is used to convert each comment to a feature set.
  - Can be used for creating vocabulary

- Stemming
  - Different forms of the same word. Stemming is a process of transforming a word into its stem.

- N Gram Analysis
  - N-Gram is a sequence of n consecutive words(e.g. “machine learning” is 2-gram)