Module 5 - Contents

1) Text Mining
2) Naive - Bayes Analysis
3) Support Vector Machines
4) Web Mining
5) Social Network Analysis
NAIVE-Bayes Analysis

➢ NB tech is a **supervised learning** tech that uses probability theory based analysis.
➢ It computes the **probability of an instance** belonging to each one of many target classes.
➢ Example: classifying text documents.
➢ Probability: from **past records**, the probability of something **happening in the future** can be reliably assessed.
➢ Example: the probability of **dying from an airline accident**, by the **total no:of airline accident related deaths in a time period** by the **total no:of flying during that time**.
Advantages and Disadvantages

- The NB Logic is simple for classification of instances.
- Conditional Probability can be computed for discrete data.
- It computes the probability of a new occurrence not only on the recent record, but also on the basis of prior experience.
- Posterior probability computation required time.
- When there are no variables in the vector X then the problem can be modeled using probability function. Such as: normal, lognormal, gamma and poisson.
NAIVE-BAYES MODEL

\[ P(c | x) = \frac{P(x | c)P(c)}{P(x)} \]

Posterior Probability

\[ P(c | X) = P(x_1 | c) \times P(x_2 | c) \times \cdots \times P(x_n | c) \times P(c) \]
Suppose a salon needs to predict the service required by the incoming customer.

If there are only two services offered: Hair Cut (R) and Manicure-pedicure (M)

Predict: whether the next customer will be R or M

Let the number of classes $K = 2$

Data of one year: 2500 customers for R and 1500 customers for M

Ans: the default probability for the next customer to be for R is $2500/4000$ or $5/8$ and for M is $1500/4000$ or $3/8$; next customer would likely be for R.

Example Sequence: R, M, R, M, M

NB Posterior Probability: $P(R) = 5/8 \times 2/5 = 10/40 = 0.25$

NB Posterior Probability: $P(M) = 3/8 \times 3/5 = 9/40 = 0.225$
## TEXT CLASSIFICATION EXAMPLE

<table>
<thead>
<tr>
<th>TRAINNING SET</th>
<th>DOCUMENT ID</th>
<th>KEYWORDS IN THE DOCUMENT</th>
<th>CLASS = H (HEALTY)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>Love Happy Joy Joy Love</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Happy Love Kick Joy Happy</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Love Move Joy Good</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Love Happy Joy Pain Love</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>Joy Love Pain Kick Pain</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>Pain Pain Love Kick</td>
<td>No</td>
</tr>
<tr>
<td>TEST DATA</td>
<td>7</td>
<td>Love Pain Joy Love Kick</td>
<td>?</td>
</tr>
</tbody>
</table>
SUPPORT VECTOR MACHINE (SVM)

SVM is a **supervised Machine Learning Algorithm** which can be used for both **classification** and **regression** challenges. However, it is most widely used in classification problem.

- **Application**: Face Detection, Text and Image Classification, Handwriting Recognition.

- In this algorithm, we plot each data item as point in a **n-dimensional space** with the value of each feature being the value of a particular coordinate.

  We perform the classification by finding the hyper-plan that differentiate the two classes very well called segments.
\[
\frac{w}{\|w\|} \cdot (x_2 - x_1) = \text{width} = \frac{2}{\|w\|}
\]

\[
w \cdot x_2 + b = 1
\]

\[
w \cdot x_1 + b = -1
\]

\[
w \cdot x_2 + b - w \cdot x_1 - b = 1 - (-1)
\]

\[
w \cdot x_2 - w \cdot x_1 = 2
\]

\[
\frac{w}{\|w\|} (x_2 - x_1) = \frac{2}{\|w\|}
\]
SVM MODEL (How it Works)

- An SVM is a classifier function in a high-dimensional space that defines the decision boundary between two classes.
- Classify the label set of points into two classes called segments. The goal is to find the best classifier between the two points of the two types.
- SVM takes the widest street approach to demarcate the two classes and thus finds the hyperplane that has the widest margin.
- In diagram, the dotted line is the optimal hyperplane. The hardlines are the gutters on the sides of the two classes.
  - The gap between the gutters is the maximum margin.
  - The classifier is defined by only those points that fall on the gutters on both sides. These points are called Support Vector.
  - Rest of the data points are irrelevant for defining the classifier.
SVM MODEL

- Notations
  - The training data of n-points : \((X_1, Y_1), \ldots, (X_n, Y_n)\)
  - \(X\) will represent two binary class value 1 or -1
  - The classifier hyperplan that satisfies the equation \(W \cdot X + b = 0\); where \(W\) = normal vector to the hyperplan
  - The Hard Margin can be defined ; \(W \cdot X + b = 1\) and \(W \cdot X + b = -1\)
  - The width of the hard margin can be calculated \(\frac{2}{|W|}\)
THE KERNEL METHOD

● The heart of an SVM algorithm is the kernel method. Most kernel algorithm is based on optimization in a convex space.

● Kernel methods operate using what is called the “Kernel trick”.

● Trick involves computing and working with the inner products of only the relevant pairs of data in the feature space.

● The kernel trick makes the algorithm much less demanding in computational and memory resources.
There are several types of support vector models including linear, polynomial, and sigmoid.

In all types, we assign high weights to the abnormal situation and very low weight to the normal situation.

SVM is more flexible and be able to tolerate some amount of misclassification. By categorising Soft Margin and Hard Margin.
ADVANTAGES AND DISADVANTAGES

- SVM work very well when no: of features are larger then the instances.
- It can work on data set with huge features space; example: spam filtering.
- SVM are easy to understand. They create an easy-to-understand linear classifier.
- They are computationally efficient.
- SVM are now available with almost all data analytics toolsets.
- It works well only with real numbers. All the data should be defined in numerical values.
- It works only with binary classification problem.
- Training the SVM is an inefficient and time consuming process. When the data is large.
- SVM does not work well when there is much noise in the data.
- SVM will also not provide a probability estimate of classification.
Text Mining is the art and science of discovering patterns from an organized collection of textual database.

Textual mining can help with frequency analysis of important terms and their semantic relationship.

Text mining can help be applied to large scale social media data for gathering preferences and measuring emotional sentiments.
5.1.1: TEXT MINING - EXAMPLE

- **Format**: Word Documents, PDF Files, XML Files, Text messages etc
- **In Legal Profession**: text sources would include law, court delibration, court orders etc
- **Academic Research**: published papers and articles
- **World of Finance**: statutory reports, internal histories, discharge summaries etc
- **Medicine**: medical journals, patient history, discharge summaries
- **Marketing**: advertisement, customers comments etc.
5.1.2: Text Mining Applications

- Marketing
  - The voice of the customer can be captured in its native and raw format and then analyzed for customer preference and complaints.
- Social Personas are a clustering technique to develop the customer segments of interest.
- Listening platform is a text mining application, that in real time gathers social media, blogs and other textual feedback and filters out the chatter to extract true consumer sentiments.
- The BPO conversation and records can be analysed for pattern of customers complaints.
Text Mining Applications

- **Business Operations**
  - Social Network Analysis and Text Mining can be applied to email, blogs and social media to measure the emotional status and the mood of employee populations.
  - Studying people as emotional investors and using text analysis of the social internet.

- **Legal**
  - Lawyers can more easily search case histories and law for relevant documents in a particular case
  - E-discovery platforms that help in minimizing risk in the process of sharing legal documents.
Text Mining Applications

- Governance and Politics
  - Social Network analysis and text mining of large scale social media data can be used for measuring emotional state and the mood of constituent populations.
  - In geo political security, internet chatter can be proceed for real time information and to connect the dots on any emerging threats.
There are three levels

- **Level 1 : Identifying frequent words**
  - This create a bag of important words. Text documents or smaller messages – can then be ranked on how they match to a particular bag-of-words
  - *Establishing the corpus of text and organized.*

- **Level 2 : Identifying the meaningful phrases from the words.**
  - Example : ice and cream will be two different key words that often come together. However there is a more meaningful phrases by combining the two words into “ice cream”. This is also called “Structure Using Term Document Matrix”

- **Level 3 : Multiple Phrases can be combined or Mine TDM for Patterns**
  - The two phrases can be put into a common bucket and this bucket is called “Desserts”

Refer diagram in page no : 138
This is the heart of structuring process. Text can be converted into numerical data, which can then be mined using regular data mining technique.

One could call key words, phrases or a topic as a term of interest. This approach measure the frequencies of selecting important term occurring in each document. This create a txd, where t=no:of terms and d= no:of documents.
The TDM can be mined to extract the patterns/knowledge. The variety of tech can be applied.

- Visualize the highest frequency term, this can be done using wordcloud tech.
- Predictors of desirable terms, example the word “profit” is a desirable word in a document. The no:of occurrence of the word profit in a document could be regressed against many other terms in the TDM.
Web Mining is the art and Science of discovering patterns from WWW so as to improve.

The WWW is at the heart of the digital revolution, billions of users are using it every day for a variety of purposes.

The web is used for electronic commerce, business communication, and many other applications.

Data for web mining is collected via web crawlers, web logs and other means.
Characteristics of web sites

- **Appearance**: Attractive Design, well-formatted content, easy to scan and navigate, good color contrasts.
- **Content**: well-planned information architecture with useful contents, Fresh content, SEO, link to other good sites.
- **Functionality**: Accessible to all the authorized users, fast loading time, usable form, mobile enabled.
- **FeedBack Analysis**: FeedBack data can be used for commercial advertisement and even for social engineering.
TYPES OF WEB MINING (3 TYPES)
A web site is designed in the form of pages with distinct URL. A large website may contain thousand of web pages.

These pages are managed by a specialized software called Content Management Systems.

Every page may have text, graphics, audio, video, forms, applications, and more kind of content.

The website keeps a record of all request received for its URL, including the requester information using cookies.

The log of these requests could be analyzed to gauge the popularity of those pages among different segment of the population (PageRank Algorithm).
Web Content Mining

- The text and application content on the pages could be analyzed for its usage by visit counts.
- Use quality contents to attract the users.
- Unwanted and unpopular pages could be weeded out or they can be transfered with different contents and style.
- Assign more resources to keep more popular pages fresh and inviting.
WEB STRUCTURE MINING

- The web works through a system of hyperlinks using the http.
- Any page can create a hyperlink to any other pages (Link to another pages), self referral nature of web lends itself to some unique network analytical algorithms.
- There are two basic strategic models for successful websites - Hubs and Authorities.
WEB STRUCTURE MINING

- **Hubs (Gathering Point):**
  - These are pages with large number of interesting links, media site like yahoo.com or government site could serve that purpose.
  - More focus site like [www.google.com](http://www.google.com) could aspire to become hub for new emerging areas.

- **Authorities:**
  - The page which provide the most complete and authoritative information on a particular subject.
  - Example: news, advice, user reviews etc
  - These web site have maximum number of inbound link from other web sites.
WEB USAGE MINING

- The goal of the web usage mining is to extract useful information and patterns from data generated through web page visits and transactions.
- The activity data comes from data stored in server access log, referrer logs, agent logs, and client-side cookies.
- The user characteristics and usage profile are also gathered directory or indirectly through syndicate data.
- Meta data such as page attributes, content attribute and usage data are also gathered.
The web content could be analyzed at multiple levels (Figure 14.2).

**Figure 14.2** Web Usage Mining Architecture

- **Website Visitors, Users, Customers**
- **Web Logs**
  - Web Click-streams
- **Gather and Prepare Data for Analysis**
  - Identify users
  - Identify visits
  - Identify page views
  - Summarize activity
- **Mine Data for Patterns**
  - Usage patterns
  - Usage profiles
  - Web page profiles
  - Web optimization

The server side analysis would show the relative popularity of the web pages. Visitors could be hubs and authorities.
There are two types of web usage mining

- **The Server Side Analysis**
  - Display the relative popularity of the web pages accessed. Those web sites could be hubs and authorities.

- **The client-side Analysis**
  - Focus on the usage pattern or the actual content consumed and created by users.
  - Clickstream analyse web activity for pattern of sequence of clicks and location and duration of the visit on web sites.
  - Clickstream analysis can be useful for web activity analysis, software testing, market research and analysing employee productivity.
There are two types of web usage mining

- **client-side Analysis**
  - Textual information can be analyzed using text mining technique called bag-of-words.
  - Bag-of-words matrix mine using cluster analysis and association rules for patterns such as popular topic and sentiment analysis.
  - Application: it can help to predict user behavior based on previously learned rules and users profile and can help to determine lifetime value of clients.
  - Can be used to designed cross marketing strategy by using association rule.