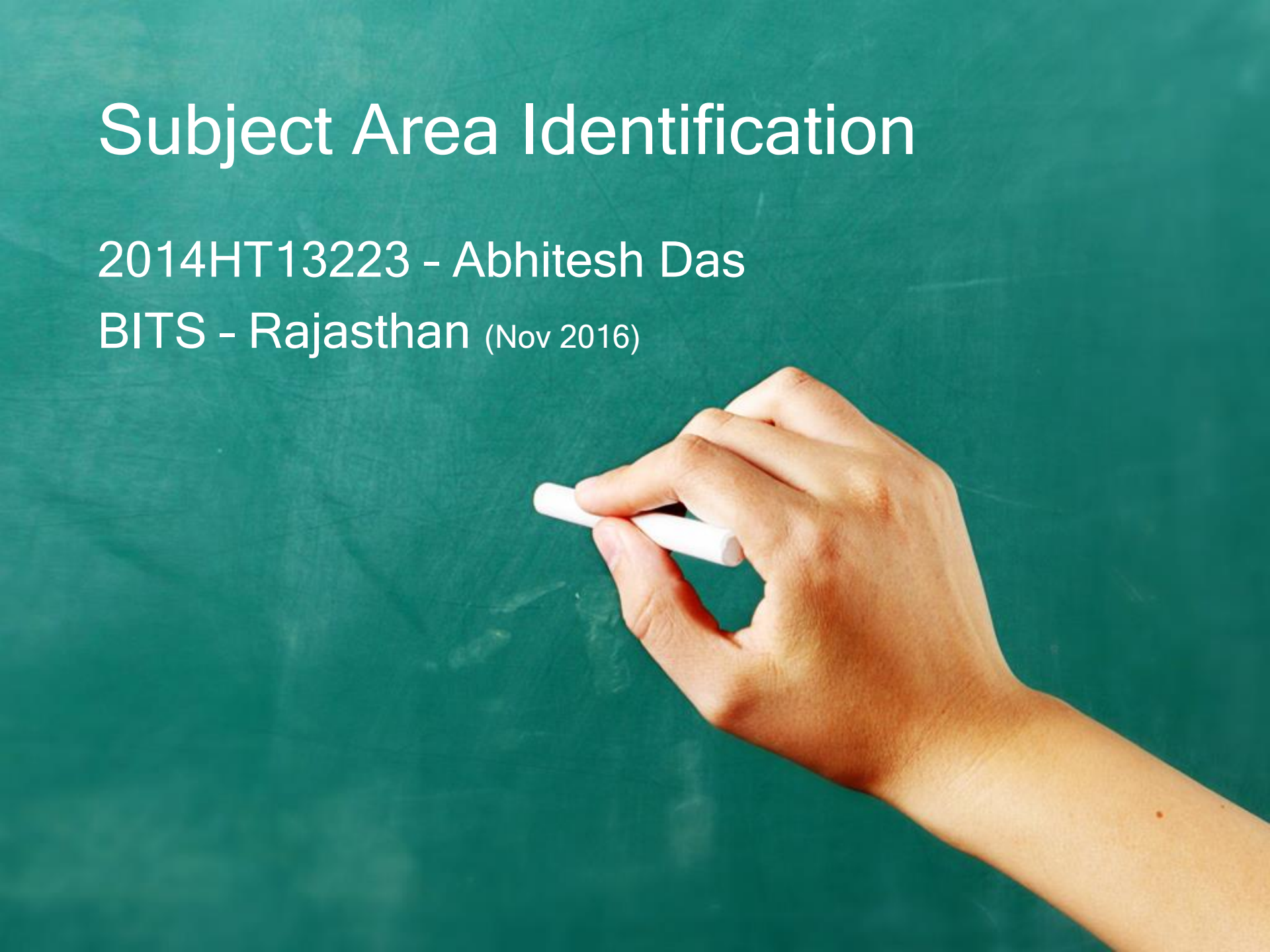
Subject Area Identification

2014HT13223 - Abhitesh Das
BITS - Rajasthan (Nov 2016)



Acknowledgements

I express my sincere thanks to my Supervisor Mr. Aditya Vadrevu and my Additional Examiner Mr. Praveen Sharma for their constant guidance, and encouragement.

I would also like to thank my mentor Mr Gokul Kanan Sadasivam, for valuable feedback during the course.

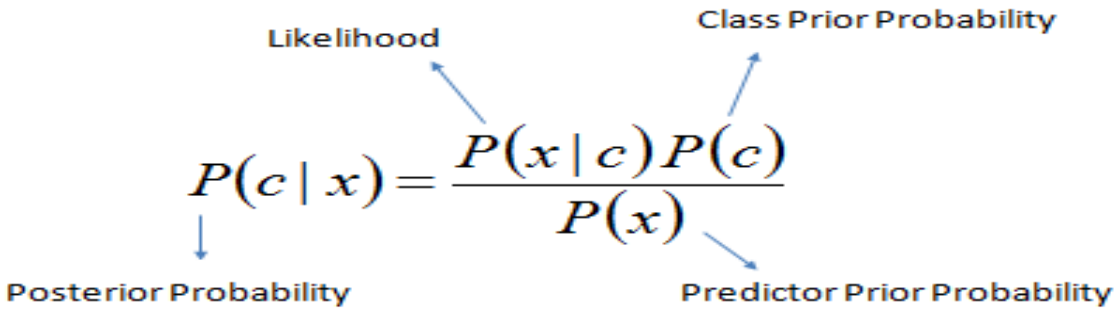
Agenda

- Introduction
- List of algorithms
- Conclusion
- Future work

Introduction

- We receive hundreds of paper daily for editing
- These paper needs to be edited by a specific editor who is well-versed with the subject area of the paper, so that we can deliver the quality output to the client.
- For this to happen, we need to determine the subject area of the document.
- As of now, this is done by humans, which takes a lot of time and effort.
- Automation of this subject area identification will save us a lot of time and effort and will also get us rid of human error

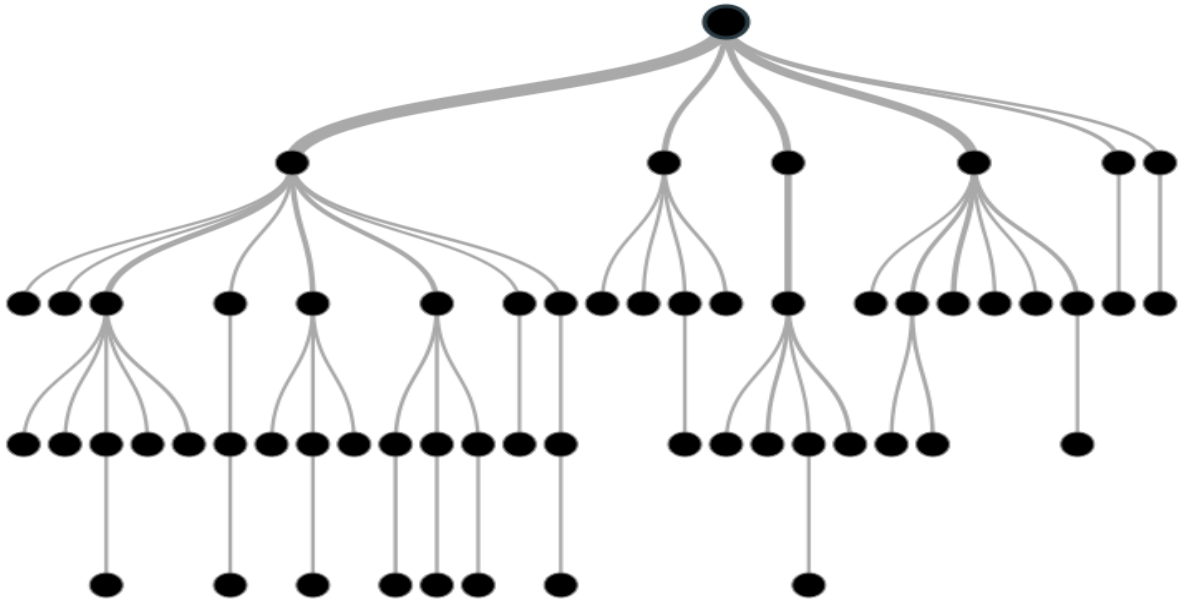
Algorithm – Naïve Bayes



$$P(c \mid \mathbf{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

- Its one of the simplest algorithm to start with
- It simply uses the bag of words to predict the class of the new data
- Its called naïve, because of the assumption that all the features are independent of each other i.e they can exists irrespective of the occurrence of the one another.
- Its one of the most efficient text classification algorithm.
- Simple to implement

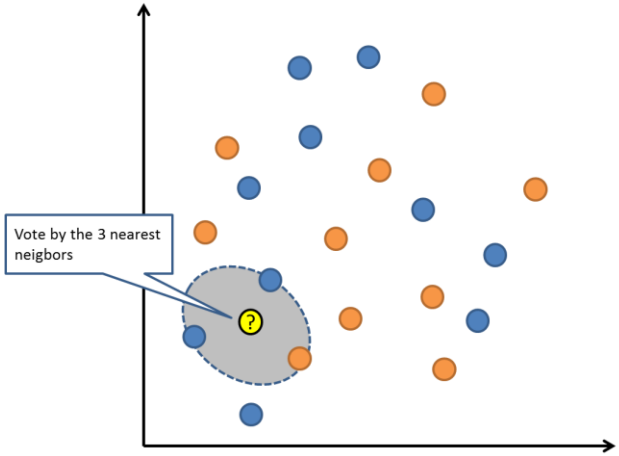
Algorithm – Decision Tree



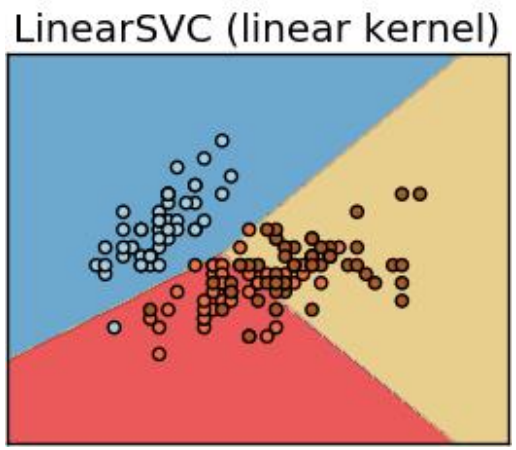
- It gives a very good visual representation of data set
- Tree is created by splitting data up by variables and then counting the no of items that fall in each bucket
- It is prone to over fitting
- Building a model takes more time as compared to Naïve Bayes

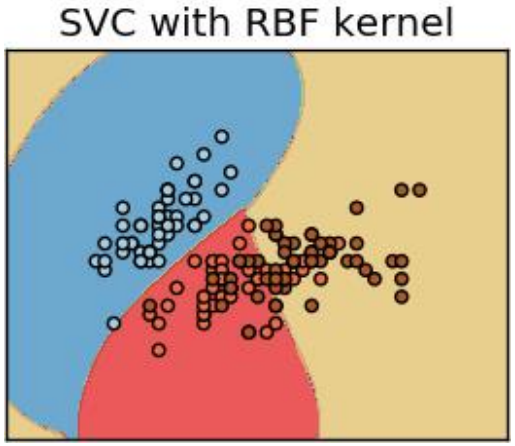
k - Nearest neighbor

- It doesn't builds a prediction model while training, rather simply records the class labels for each input data item
- Hence, takes less time to train
- However, while determining the class of new data set, it tries to find the k-nearest neighbors and then classify the new data set as the class label with max votes.
  (k determines the no. of neighbors to find)
- While voting, the class labels nearer to the new data point are weighted higher than that of those which are far away from the new data point
- Hence, takes more time to predict the class of new data point
- It can be computation expensive



Support Vector Machines (aka SVM)





- Its simply a line separating two different classes
- However, at time a simple *(linear)* line wont be enough to separate the classes.
- In such cases, we need to fit the curve i.e achieved by changing the kernel to rbf or poly, etc
- In a nutshell, SVM helps find the optimal decision boundary
- It maximizes the margin of decision boundary
- SVM is also very well know for its ability to deal with outliers.
- It takes a little more time to train, but its worth it.

Conclusion

	Documents
Training	10774
Test	4618
Total	**15392**

Algorithm	Accuracy
Naïve Bayes	51 %
Decision Tree	33 %
k-NN	10%
Support Vector Machines	**60%**

For our data set we achieve maximum accuracy with SVM

Future work

- Process all the papers received in last one year ~1 lakh
- Detect the classes for which the algo is not classifying accurately with the help of confusion matrix and fix them by adding more training materials or by tuning the algo parameters
- Fine tune the algorithm with help of grid search
- Implement the self learning for the selected algorithm – so it rebuilds itself every week with the new knowledge
- Determine the hardware requirements for the final algo implementation
- Design a interface to easily change the parameters of the algorithm

Thank you