

Text Categorization of South Asian Last Names

Sabyasachi Guharay
Advisor: Prof. D. Madigan

Statistics Dept.
Rutgers University

July 24, 2003

INTRODUCTION

I. Text Categorization

a) Deterministic algorithms to classify text

Ex.: “To be or not to be”?

Y

Shakespearean or NOT?

1 = TRUE; 0 = FALSE

II) Easiest way to solve cases such as above is to implement binary conditions

III) Further sophisticated Machine learning Algorithms for non-Binary classification

IV) This project deals with Binary classification of South Asian Last Names.

DATA SET

I) 5000 South Asian surnames

II) 5000 non-South A surnames

III) Training Data

4000 SA surnames; 4000 non-SA surnames

IV) Testing Data: Remaining cases

ALGORITHM POSSIBILITIES

I) Support Vector Machines

II) Lightweight Induction rules

III) Naïve-Bayes Classifier

COMMON IN ALL METHODS

Error Analysis

| | Rule True | Rule False |
|------------------------|----------------------|-----------------------|
| Class True | True + | True - |
| Class False | False + | False - |

NAÏVE BAYES

**I) Generate triplets or Quadruplets
(keys)**

Ex.: “ngh” → Singh

Ex.: “chak” → Chakraborty

“chak” → Chaki

II) Develop exhaustive list of keys

NAÏVE BAYES CONTD.

III) Use Training Data

| Name | “ingh” | “chak” | “guh” |
|--------------|--------|--------|-------|
| Singh | 1 | 0 | 0 |
| Chakra-Borty | 0 | 1 | 0 |
| Guharay | 0 | 0 | 1 |
| Guha | 0 | 0 | 1 |

1 = TRUE; 0 = FALSE

NAÏVE BAYES CONTD.

**IV) Simple Model of using
Training Data → Predict**

Let Y = Binary Predictor variable SA or non-SA

| | X₁ | X₂ | X₃ | Y |
|----------|----------------------|----------------------|----------------------|----------|
| 1 | 1 | 0 | 1 | 1 |
| 2 | 0 | 1 | 1 | 1 |
| 3 | 1 | 0 | 0 | 0 |
| 4 | 0 | 0 | 1 | 0 |

$$\text{Prob}(Y=1 | X_1 = x_1, X_2 = x_2, \dots, X_n = x_n)$$



Estimate Y

NAÏVE BAYES CONTD.

V) Idea from Bayes Theorem

$$\text{Prob}(Y=1 | X_1 = x_1) = \frac{[\text{Prob}(X_1 = x_1 | Y=1) * \text{Prob}(Y=1)]}{[\text{Prob}(X_1 = x_1)]}$$



Bayes Theorem

→ $\text{Prob}(Y=1 | X_1 = x_1) \text{ prop. } \text{Prob}(X_1 = x_1 | Y=1) * \text{Prob}(Y=1)$

$\text{Prob}(Y=1) \rightarrow$ frequency count

Extend above to n variables

NAÏVE BAYES CONTD.

VI) Main Advantage:

$$\text{Prob}(Y=1 | X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) \sim \text{Prob}(X_1 = x_1 | Y=1) * \text{Prob}(X_2 = x_2 | Y=1) * \dots * \text{Prob}(X_n = x_n | Y=1) * \text{Prob}(Y=1)$$

i.e. INDEPENDENT!!!!

$\text{Prob}(A, B) = \text{Prob}(A) * \text{Prob}(B) \rightarrow A \text{ \& B are statistically independent}$

NAÏVE BAYES CONTD.

Small Example:

| | X_1 | X_2 | X_3 | Y |
|----------|----------|----------|----------|----------|
| 1 | 1 | 0 | 1 | 1 |
| 2 | 0 | 1 | 1 | 1 |
| 3 | 1 | 0 | 0 | 0 |
| 4 | 0 | 0 | 1 | 0 |

$\text{Prob}(Y=1) = 1/2$; $\text{Prob}(X_1 = 1 | Y= 1) = 1/2$; $\text{Prob}(X_2 = 1 | Y= 0) = 0/2 \leftarrow \text{????}$

NAÏVE BAYES CONTD.

VII) Fix Previous issue regarding probability of ZERO

$$\text{Prob}(X_1 = 1 | Y = 1) = [1 + d] / [2 + d]$$

$$\text{Prob}(X_2 = 1 | Y = 0) = [0 + d] / [2 + d]$$

It has been determined that $d \sim 1/2$.

NAÏVE BAYES CONTD.

VIII) Convert to Log Scale:

$$\mathbf{X}^* = (\mathbf{X}_1 = \mathbf{x}_1, \mathbf{X}_2 = \mathbf{x}_2, \dots, \mathbf{X}_n = \mathbf{x}_n)$$

$$\log \frac{\text{Prob}(Y = 1 | \mathbf{X}^*)}{\text{Prob}(Y = 0 | \mathbf{X}^*)} = \log \frac{\text{Prob}(Y = 1)}{\text{Prob}(Y = 0)} + \sum_{i=1}^d \log \frac{\text{Prob}(X_x = x_i | Y = 1)}{\text{Prob}(X_i = x_i | Y = 0)}$$

↑↑

| | \mathbf{X}_1 | \mathbf{X}_2 | \mathbf{X}_3 | \mathbf{Y} |
|-------|----------------|----------------|----------------|--------------|
| y_1 | 1 | 1 | 0 | ? |
| y_2 | . | . | . | . |
| ... | . | . | . | . |
| y_n | . | . | . | ? |

With these results for the actual data sets, the next question involves determining the accuracy of y_i .

THRESHOLDING

Q How accurate do we want?

$p = 0.8?$

$p = 0.7?$

$p = 1? \rightarrow$ ALL names are SA

$p = 0? \rightarrow$ NO names are SA

How to accurately choose p ?

Further Work