

Naive Bayes

Pritam Prakash Shete

Computer Division, BARC

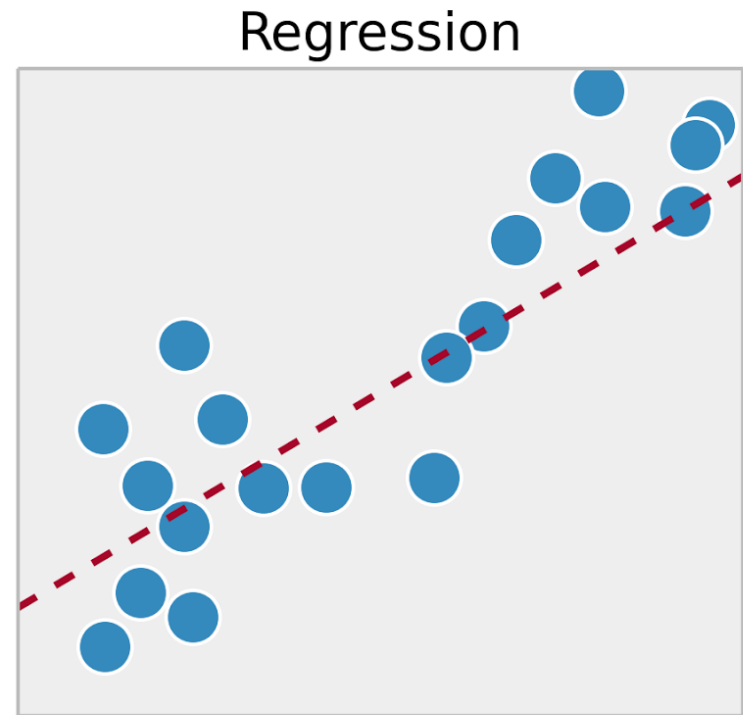
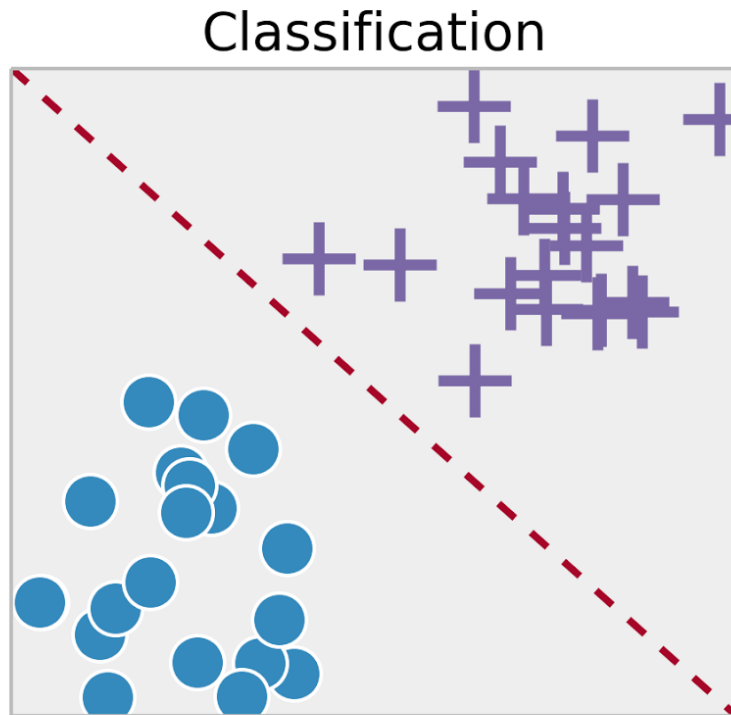
Centre for Excellence in Basic Sciences

Topics

- Supervised learning
- Conditional probability
- Bayes' rule
- Naive Bayes
- Laplacian smoothing
- Prior ratio
- Log likelihood
- Applications
- Advantages
- Disadvantages

Supervised Learning

- Training set – $\{ (x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(m)}, y^{(m)}) \}$
- Labeled dataset



Probability

- SMS spam detection
- Corpus of SMSs
- Events – Ham SMS or Spam SMS

$$P(\text{Spam}) = \frac{N_{\text{Spam}}}{N_{\text{Total}}} = 1 - P(\text{Ham})$$

$$P(\text{Ham}) = \frac{N_{\text{Ham}}}{N_{\text{Total}}} = 1 - P(\text{Spam})$$

Name	Value
N_{Total}	1000
N_{Ham}	800
N_{Spam}	200
$P(\text{Ham})$	0.8 (80%)
$P(\text{Spam})$	0.2 (20 %)

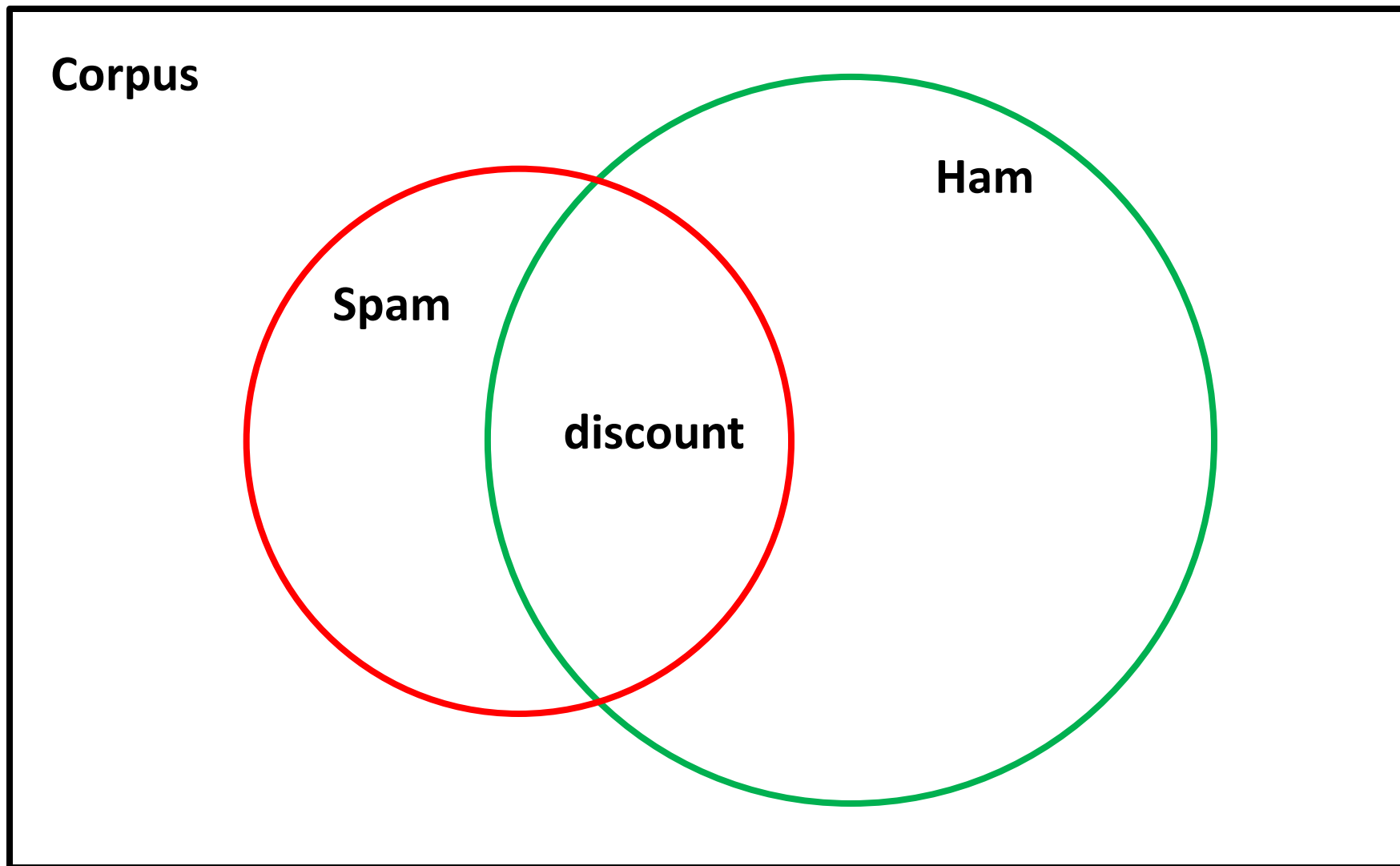
Probability

- SMS spam detection
- Corpus of SMSs
- Event – SMS contains word ‘discount’

$$P(\text{discount}) = \frac{N_{\text{discount}}}{N_{\text{Total}}}$$

Name	Value
N_{Total}	1000
N_{discount}	20
$P(\text{discount})$	0.02 (2 %)

Conditional Probability



Conditional Probability

- Event – SMS – Spam
- Event – SMS – ‘discount’
- Intersection – Spam and ‘discount’

$$P(\text{Spam} | \text{discount}) = \frac{P(\text{Spam} \cap \text{discount})}{P(\text{discount})}$$

- Probability that SMS is Spam, given SMS contains word ‘discount’.
- Probability that SMS is Spam and SMS contains word ‘discount’.

Bayes' Rule

- Event – SMS – Spam
- Event – SMS – 'discount'
- Intersection – Spam and 'discount'

$$P(\text{Spam} | \text{discount}) = \frac{P(\text{Spam} \cap \text{discount})}{P(\text{discount})}$$

$$P(\text{discount} | \text{Spam}) = \frac{P(\text{discount} \cap \text{Spam})}{P(\text{Spam})}$$

Bayes' Rule

- Event – SMS – Spam
- Event – SMS – 'discount'
- Intersection – Spam and 'discount'

$$P(\text{Spam} | \text{discount}) = P(\text{discount} | \text{Spam}) \times \frac{P(\text{Spam})}{P(\text{discount})}$$

Bayes' Rule

- Event – SMS – Spam
- Event – SMS – 'discount'
- Intersection – Spam and 'discount'

$$P(\text{Spam} | \text{discount}) = P(\text{discount} | \text{Spam}) \times \frac{P(\text{Spam})}{P(\text{discount})}$$

$$P(A | B) = P(B | A) \times \frac{P(A)}{P(B)}$$

Naive Bayes

- SMS spam detection
- Corpus
 - Spam corpus
 - Ham corpus
- Vocabulary
 - All words
 - Spam and Ham corpus

Word	Ham	Spam
I	400	400
am	400	400
because	50	0
meeting	200	100
company	350	100
discount	100	200
lottery	100	400
	1600	1600

Naive Bayes

- SMS spam detection
- Corpus
 - Spam corpus
 - Ham corpus
- Vocabulary
 - All words
 - Spam and Ham corpus

Word	Ham	Spam
I	0.25	0.25
am	0.25	0.25
because	0.03125	0
meeting	0.125	0.0625
company	0.21875	0.0625
discount	0.0625	0.125
lottery	0.0625	0.25
	1600	1600

$$P(\text{Word}_i \mid \text{Class})$$

Naive Bayes

- SMS spam detection
- Corpus
 - Spam corpus
 - Ham corpus
- Vocabulary
 - All words
 - Spam and Ham corpus

Word	Ham	Spam
I	400	400
am	400	400
because	50	0
meeting	200	100
company	350	100
discount	100	200
lottery	100	400
	1600	1600

$$P(I | \text{Ham}) = \frac{400}{1600} = 0.25$$

Naive Bayes

- SMS spam detection
- Corpus
 - Spam corpus
 - Ham corpus
- Vocabulary
 - All words
 - Spam and Ham corpus

Word	Ham	Spam
I	0.25	0.25
am	0.25	0.25
because	0.03125	0
meeting	0.125	0.0625
company	0.21875	0.0625
discount	0.0625	0.125
lottery	0.0625	0.25

Naive Bayes

- Identical probabilities
 - Neutral words
 - I, am
- Significant words
 - meeting, company
 - discount, lottery

Word	Ham	Spam
I	0.25	0.25
am	0.25	0.25
because	0.03125	0
meeting	0.125	0.0625
company	0.21875	0.0625
discount	0.0625	0.125
lottery	0.0625	0.25

Naive Bayes

$$\prod_{i=1}^m \frac{P(\text{word}_i | \text{Ham})}{P(\text{word}_i | \text{Spam})} > 1$$

I am in meeting.

$$\frac{0.25}{0.25} \times \frac{0.25}{0.25} \times \frac{0.125}{0.0625} = 2 > 1$$

Ham

Word	Ham	Spam
I	0.25	0.25
am	0.25	0.25
because	0.03125	0
meeting	0.125	0.0625
company	0.21875	0.0625
discount	0.0625	0.125
lottery	0.0625	0.25

Naive Bayes

$$\prod_{i=1}^m \frac{P(\text{word}_i | \text{Ham})}{P(\text{word}_i | \text{Spam})} > 1$$

You got discount.

$$\frac{0.0625}{0.125} = 0.5 < 1$$

Spam

Word	Ham	Spam
I	0.25	0.25
am	0.25	0.25
because	0.03125	0
meeting	0.125	0.0625
company	0.21875	0.0625
discount	0.0625	0.125
lottery	0.0625	0.25

Laplacian Smoothing

- Word count – 0
- Probability of word – 0
- Probability of sentence – 0
- e.g.
 - Word ‘because’ – Spam

Laplacian Smoothing

$$P(\text{Word}_i | \text{Class}) = \frac{C(\text{Word}_i, \text{Class})}{N_{\text{Class}}}$$

$$P(\text{Word}_i | \text{Class}) = \frac{C(\text{Word}_i, \text{Class}) + 1}{N_{\text{Class}} + U_{\text{Class}}}$$

$C(\text{Word}_i, \text{Class})$ = Number of Word_i words in Class

N_{Class} = Number of words in Class

U_{Class} = Number of unique words in Class

Word	Ham	Spam
I	0.2495	0.2495
am	0.2495	0.2495
because	0.0317	6.2e-4
meeting	0.125	0.0628
company	0.2184	0.0628
discount	0.0628	0.125
lottery	0.0628	0.2495

Prior Ratio

- Number of ham (positive) SMSs
- Number of spam (negative) SMSs
- Ratio of number of ham and spam SMSs

Log Likelihood

- Word classification
 - Neutral words – ~ 1
 - Positive words – > 1
 - Negative words – < 0
- Multiplication
 - Numerical underflow

Log Likelihood

- Logarithm of ration of probability
 - Log likelihood
- Addition of logarithms
 - No numerical underflow
- Word classification
 - Neutral words – < 1
 - Positive words – > 1
 - Negative words – ~ 0
- Logprior

Applications

- Spam detection
- Text classification
- Author identification
- Word disambiguation
- Sentiment analysis

Advantages

- Simple
- Easy to implement
- Real-time predictions
- Work with less training dataset
- Continuous and discrete data
- Not sensitive to irrelevant features

Disadvantages

- Assume independent features

Questions?

Thank you