Naive Bayes-01

Sat, 12 Nov 2022 11:00

Q- Guess the word?

Ans: "Classification word"

Agenda

→ Problem Statement : Spam Classification

→ Naive Bayes.

→ Mathematical Intuition
- Prior
- Conditional Probabilities

→ Naive Assumption

→ Time & Space Complexity

Problem Statement

[Spam & Ham] / Text Classification

"Email" "Text" "Review"

"Sentiment Analysis"

random

Amazon Review

Movie Review

Ex-1.) "I, Nguyen know web help. free money"

Ex-2.) "money Goldman U.S. Mortgage"

Report: "Nguyen know", "free money", "Money", "web-dev" free free free

Data.

"Twitter data"

Spy of emails → s/o

$Data^{train} \rightarrow w_1^{s/o}, w_2^{s/o}, w_3^{s/o}, w_4^{s/o} \ldots w_n^{s/o}$

Email are of different length.

Data - Preprocessing

① Lower → "Free", "Free", "Hello", "Hello"

Index & unique word

② Tokenization → "Hello"

["H", "e", "l", "l", "o"]

③ Punctuation → @ ! , _ =

④ Stopwords Removal → → ns (ns) or, 1 ns

Index & unique words.

"free" → "Free"

Vector Intuition

[Bag of Words]

"Bag of words" "bowwords"

Vocabulary → "Unique words"

	a	an	apple	bag	big	for	gift	money	month	pen	...	zoo
1st mail	0	0	0	0	0	1	1	0	0	0		0
2nd mail	1	0	0	1	0	0	0	1	0	1		0
3rd mail												
4th mail												

Naive Bayes

Bayes Rule

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

$P(Y=1|\text{Mail}) \quad \text{and} \quad P(Y=0|\text{Mail})$

$P(Y=1|\text{money, ..., ns}) \quad \text{and} \quad P(Y=0|\text{money, ..., ns})$

$P(Y=1|X)$ $\qquad$ $1 - P(Y=1|X)$

Posterior

$P(Y=1|X) = \frac{P(X|Y=1) \cdot P(Y=1)}{P(X)}$

Posterior Likelihood Prior Evidence

Posterior Probability (Ranking)

$P(Y=1|X) = \frac{P(Y=1|X) \cdot P(Y=1)}{P(X)}$

$P(Y=0|X) = \frac{P(Y=0|X) \cdot P(Y=0)}{P(X)}$

$P(Y=1|X) = 0.92$

$P(Y=0|X) = 0.12$ → ? Assign to class with Max Prob

Multinomial

"Spam" → "Social", "Promotion", "Primary"

$P(Y=\text{Spam}|X)$ $\qquad$ $P(Y=\text{Social}|X)$

$P(Y=\text{Promo}|X)$ $\qquad$ $P(Y=\text{Primary}|X)$

↑ Posterior Probability

$P(Y=1|X) \propto P(\text{money, ..., ns}|Y=1) \cdot P(Y=1)$

$P(Y=0|X) \propto P(\text{money, ..., ns}|Y=0) \cdot P(Y=0)$

$0.92 < 0.98$

Prior Probability.

$P(Y=1) = \frac{\#\text{ham msg}}{\#\text{ham msg} + \#\text{spam msg}}$

$P(Y=0) = \frac{\#\text{ham msg}}{\#\text{ham msg} + \#\text{spam msg}}$

Likelihood / Conditional Probability.

$P(X|Y=1)$ & $P(X|Y=0)$

$P(\text{money, ..., ns}|Y=1)$ & $P(\text{money, ..., ns}|Y=0)$

Naive Assumption

Naive → "very very very naive"

$P(X|Y=1) \approx$

$P(Y=1|X)$

$P(\text{money, ..., ns}|Y=1)$

Big data, very very very

Naive assumption

money
free
free
free