Naive Bayes

SGA07_DATASCI

5th March 2020

Module Overview

- Review of Probability
- Baye's Theorem
- Naive Bayesian Classification

Book Keeping

- Group task submission: Submission by to 6:00pm today
- Next week starts the last quarter of the course modules
- Do not forget to turn-in your daily challenge from yesterday

Outcome

After this Module, you will;

- Review the topics of probability - random variable and probability distribution
- Understand how to extend probability rules into Bayes' theorem
- Learn how to apply and evaluate Naive Bayesian Classification model

Probability (Def.)

- Study of uncertainty
- Usually a number between 0 and 1
- Toss of a fair coin or roll of a 6-sided dice

> Probability refers to an assessment of the likelihood of the various possible outcomes in an experiment or some other situation with a "random" outcome.

Probability (Concepts)

- A random variable is the outcome of a natural process that can not be predicted with certainty.
- A sample space is the set of all possible outcomes for a random variable
- An event space is a set whose elements are subset of the sample space
- A probability distribution is the sample space together with all probabilities

Frequency Tables

- The frequency of a particular data value is the number of times that data value occurs
- A frequency table is constructed by arranging collected data values in ascending order of magnitude with their corresponding frequencies
- For example, consider the marks awarded for an assignment set for a Year 8 class of 20 students were as follows:

6	7	5	7	7	8	7	6	9	7
4	10	6	8	8	9	5	6	4	8

Frequency Tables

Mark	Frequency	Probability
4	2	0.1
5	2	0.1
6	4	0.2
7	5	0.25
8	4	0.2
9	2	0.1
10	1	0.05
Sum	20	1

Probability (Maths)

- Given a random variable A, for an event F which is a subset of sample space Ω
  - $P(A) \geq 0$, for all $A \in F$
  - If $A_1, A_2, \ldots$ are disjoint events (i.e $A_i \cap A_j = \emptyset$ whenever $i \neq j$), then $P(\cup_i A_i) = \sum_i P(A_i)$
  - $P(\Omega) = 1$

Probability (Maths)

- $A \subseteq B \implies P(A) \leq P(B)$
- $P(A \cap B) = min(P(A), P(B))$
- Union Bound: $P(A \cup B) \leq P(A) + P(B)$
- $P(\Omega - A) = 1 - P(A)$
- Law of Total Probability: If $A_1, \ldots, A_k$ are a set of disjoint events such that $\cup_{i=1}^{k} A_i = \Omega$, then $\sum_{i=1}^{k} P(A_i) = 1$

Conditional Probability

- Let B be an event with non-zero probability. The conditional probability of any event A given B is defined as

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

- In other words, $P(A \mid B)$ is the probability measure of the event A after observing the occurrence of event B.

Chain Rule

- Let $S_1, \ldots, S_k$ be events, $P(S_i) > 0$. Then the chain rule states that:

$$P(S_1 \cap S_2 \cap \ldots \cap S_k) = P(S_1)P(S_2 | S_1)P(S_3 | S_2 \cap S_1) \ldots P(S_k | S_1 \cap S_2 \cap \ldots \cap S_{k-1})$$

- Note that for $k = 2$ events, this is just the definition of conditional probability

$$P(S_1 \cap S_2) = P(S_1)P(S_2 | S_1)$$

- In general, the chain rule is derived by applying the definition of conditional probability multiple times

Independence

- Two events are called independent if $P(A \cap B) = P(A)P(B)$, or equivalently, $P(A \mid B) = P(A)$. Intuitively, A and B are independent means observing B does not have any effect on the probability of A

Probability Functions

- Cumulative distribution function
- Probability mass function
- Probability density function
- Expectation
- Variance

Common Random Variables

- Discrete random variables
  - Bernoulli
  - Binomial
  - Geometric
  - Poisson
- Continuous random variables
  - Uniform
  - Exponential
  - Normal

Bayes' Theorem

- Posterior probability of A given B = probability of B given A $\cdot$ prior probability of A / probability of B
- I.e Prior probability $P(A)$ updated in light of new evidence B

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

Bayes' Theorem Example 1

- Given:
  - Probability of seeing a black sheep $P(B) = 0.1$
  - Probability of seeing a white sheep $P(W) = 0.9$
  - Probability of long hair when sheep is black $P(L|B) = 0.3$
  - Probability of long hair when sheep is white $P(L|W) = 0.2$
- What is probability of a long haired sheep being black?

Bayes' Theorem Example I

- $P(B \mid L) = P(L \mid B) \cdot \dfrac{P(B)}{P(L)}$
  - $= P(L \mid B) \cdot \dfrac{P(B)}{(P(L \mid W)P(W) + P(L \mid B)P(B))}$
  - $= 0.3 \times \dfrac{0.1}{((0.2 \times 0.9) + (0.3 \times 0.1))}$
  - $= \dfrac{0.03}{0.18 + 0.03}$
  - $= 0.143$

Bayes' Theorem Example II

- Given:
  - Meningitis causes stiff neck in 50% of cases
  - Prior probability of any patient having meningitis = 1/50,000
  - Prior probability of any patient having stiff neck = 1/20
- If patient has stiff neck, what is probability he/ she has meningitis?

$$P(M|S) = \frac{P(S|M)P(M)}{P(S)} = \frac{0.5 \times 1/50000}{1/20} = 0.0002$$

Bayes' Theorem Example III

- Given:
  - Probability of lab test giving +ve result P(+ve|Dis) if disease present = 98%
  - Probability of giving –ve result P(-ve|~Dis) if disease absent = 97%
  - Probability of disease P(Dis) = 0.8%
- If result is +ve, what is probability P(Dis|+ve) that patient has disease?

Bayes' Theorem Example III

- $P(Dis \mid +ve) = P(+ve \mid Dis) \cdot \dfrac{P(Dis)}{P(+ve)}$
  - $= P(+ve \mid Dis) \cdot \dfrac{P(Dis)}{(P(+ve \mid Dis)P(Dis) + P(+ve \mid \sim Dis)P(\sim Dis))}$
  - $= 0.98 \times \dfrac{0.008}{((0.98 \times 0.008) + (0.03 \times 0.992))}$
  - $= \dfrac{0.00784}{0.00784 + 0.02976}$
  - $= 0.21$

Naive Bayes Classifier

- Probabilistic classifier
- Based on Bayes' theorem
- Comparable in performance to decision tree and neural networks
- High accuracy and speed
- Assumes class-conditional independence

> "Naive Bayes Classifiers predict class membership probabilities such as the probability that a given tuple belongs to a particular class."

Naive Bayes Classifier

- $P(class \mid x) = P(x \mid class) \cdot \dfrac{P(class)}{P(x)}$
- Classification of new instance x, where x represented by conjunction of attribute values.
  - If $x = < a_1 = v_1, a_2 = v_2, \ldots, a_j = v_j >$
  - $P(x \mid class) = P(a_1 = v_1 \mid class) \cdot \ldots . \cdot P(a_j = v_j \mid class)$

STUTERN

Naive Bayes Classifier

- Set C $< c_1, c_2, \ldots, c_i >$ of mutually exclusive classes with prior probabilities $P(C_1) \ldots P(C_i)$ dependent on attributes $a_1, \ldots, a_n$ with values $v_1, \ldots, v_n$ for given instance x
- Conditional or posterior probability of c_i

$$P(c_i | x) = P(c_i) \cdot \frac{P(a_1 = v_1 \wedge a_2 = v_2 \ldots \wedge a_n = v_n | c_i)}{P(x)}$$

- Assuming conditional independence of attributes:

$$P(c_i | x) = P(c_i) \cdot \frac{P(a_1 = v_1 | c_i) \cdot P(a_2 = v_2 | c_i) \cdot \ldots \cdot P(a_n = v_n | c_i)}{P(x)}$$

Problems with Naive Bayes

- Estimating probabilities by relative frequencies can give poor estimate if number of instances with given attribute value combination is small.
- In extreme case, posterior probability of some attribute values may be zero.

Practice Lab

Implement a Naive Bayes classification model in R

Use the following Instructions:

- Use the Iris Dataset in R
- Explore the dataset to get some intuition
- Partition your data into train and test sets
- Build your naive bayes model using 'e1071' package
- Evaluate your model on the test set

Recap/Summary

At the end of this Module, you should understand;

- Review the topics of probability - random variable and probability distribution
- Understand how to extend probability rules into Bayes' theorem
- Learn how to apply and evaluate Naive Bayesian Classification model

Suggested Material

- https://ermongroup.github.io/cs228-notes/preliminaries/probabilityreview/
- Machine Learning by Tom Mitchell Chapter 6
- Data Mining Concepts and Techniques (3rd Edition) by Jiawei Han, Micheline Kamper and Jian Pei: Chapter 8 (Section 3)
- https://www.kaggle.com/chinki/naive-bayes-classification-for-iris-dataset
- https://www.youtube.com/watch?v=RmajweUFKvM