K2 Analytics
Building Skills, Building Individuals

Naïve Bayes

- Rajesh Jakhotia

Earning is in Learning
- Rajesh Jakhotia

About K2 Analytics

At K2 Analytics, we believe that skill development is very important for the growth of an individual, which in turn leads to the growth of Society & Industry and ultimately the Nation as a whole. For this it is important that access to knowledge and skill development trainings should be made available easily and economically to every individual.

Our Vision: *"To be the preferred partner for training and skill development"*

Our Mission: *"To provide training and skill development training to individuals, make them skilled & industry ready and create a pool of skilled resources readily available for the industry"*

*We have chosen Business Intelligence and Analytics as our focus area. With this endeavour we make this presentation on "**Naïve Bayes**" accessible to all those who wish to learn this technique using R. We hope it is of help to you. For any feedback / suggestion feel free to write back to us at ar.jakhotia@k2analytics.co.in*

You can also write to us for job opportunities on analytics on our email ar.jakhotia@k2analytics.co.in

Welcome to Logistic Regression using R!!!

Agenda

- Naïve Bayes
- Navie Bayes Algorithms
- Advantages and Disadvantages

Naïve Bayes

- In machine learning, **naive Bayes classifiers** are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naive) independence assumptions between the features.

- Naïve – Because it makes a strong assumption that all the Independent Variables i.e. attributes /features are independent and do not have any relationship with each other

- Bayes – Because it is based on the Bayes Theorem

https://en.wikipedia.org/wiki/Naive_Bayes_classifier

Bayes’ Theorem

- Bayes’ Theorem describes the probability of an event, based on prior knowledge of conditions that might be related to the event

- Bayes’ Theorem is an extension of Conditional Probability

P (A | B) = P (A ∩ B) / P (B) …………………… Eq 1

P (B | A) = P (B ∩ A) / P (A) …………………… Eq 2

However, P (B ∩ A) = P (A ∩ A)
As such from Eq 1 and Eq 2, we can write either of the below equation and this is Bayes’ Theorem

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

Where P (B) ≠ 0

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

Where P (A) ≠ 0

Naïve Bayes derivation…

- Abstractly, naive Bayes is a conditional probability model: given a problem instance to be classified, represented by a vector $\mathbf{x} = (x_1, \ldots, x_n)$ representing some n features (independent variables), it assigns to this instance probabilities

  $$p(C_k \mid x_1, \ldots, x_n)$$

  for each of K possible outcomes or *classes* C_k.[7]

- The problem with the above formulation is that if the number of features n is large or if a feature can take on a large number of values, then basing such a model on probability tables is infeasible. We therefore reformulate the model to make it more tractable. Using Bayes' theorem, the conditional probability can be decomposed as

  $$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})}$$

- In plain English, using Bayesian probability terminology, the above equation can be written as

  $$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

- In practice, there is interest only in the numerator of that fraction, because the denominator does not depend on C and the values of the features x_i are given, so that the denominator is effectively constant. The numerator is equivalent to the joint probability model

  $$p(C_k, x_1, \ldots, x_n)$$

Naïve Bayes derivation…

- $$p(C_k, x_1, \dots, x_n)$$

  which can be rewritten as follows, using the chain rule for repeated applications of the definition of conditional probability:

  $$\begin{aligned} p(C_k, x_1, \dots, x_n) &= p(x_1, \dots, x_n, C_k) \\ &= p(x_1 \mid x_2, \dots, x_n, C_k) p(x_2, \dots, x_n, C_k) \\ &= p(x_1 \mid x_2, \dots, x_n, C_k) p(x_2 \mid x_3, \dots, x_n, C_k) p(x_3, \dots, x_n, C_k) \\ &= \dots \\ &= p(x_1 \mid x_2, \dots, x_n, C_k) p(x_2 \mid x_3, \dots, x_n, C_k) \dots p(x_{n-1} \mid x_n, C_k) p(x_n \mid C_k) p(C_k) \end{aligned}$$

- Now the "naive" conditional independence assumptions come into play: assume that each feature x_i is conditionally independent of every other feature x_j for $j \neq i$, given the category C_k. This means that

  $$p(x_i \mid x_{i+1}, \dots, x_n, C_k) = p(x_i \mid C_k) \, .$$

- Thus, the joint model can be expressed as

  $$\begin{aligned} p(C_k \mid x_1, \dots, x_n) &\propto p(C_k, x_1, \dots, x_n) \\ &= p(C_k) \, p(x_1 \mid C_k) \, p(x_2 \mid C_k) \, p(x_3 \mid C_k) \, \cdots \\ &= p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k) \, , \end{aligned}$$

  where $\propto$ denotes proportionality.

Naïve Bayes Algorithms

- Bernoulli Naïve Bayes – Used when feature (Independent) variables are all binary
- Multinomial Naïve Bayes – Useful when features describe discrete frequency counts (i.e. they are not simply binomial – True / False)
- Gaussian Naïve Bayes – Good for features which are normally distributed (i.e. continuous variables can be considered)

Bernoulli Naïve Bayes calculations

Is_Male	Is_Self_Emp	count_T0	count_T1	obs	Proportions
0 - No	0	3,126	220	3,346	0.0658
	1	555	82	637	0.1287
1 - Yes	0	7,632	642	8,274	0.0776
	1	1,452	291	1,743	0.1670
Column Total		12,765	1,235	14,000	0.0882

	Target		Row Total
Male	0	1	
0 - No	3,682	303	3,985
1 - Yes	9,085	934	10,019
Col. Total	12,767	1,237	14,004

Note : All the cell in the intersection of Is_Male and Target crosstab has been incremented by 1 Unit to avoid divide by 0 error

	Target		Row Total
Self-Employed	0	1	
0 - No	10,759	863	11,622
1 - Yes	2,008	374	2,382
Col. Total	12,767	1,237	14,004

Note : All the cell in the intersection of Is_Male and Target crosstab has been incremented by 1 Unit to avoid divide by 0 error

Simple Probabilities		
P(M = 1)	= # Male / Total Obs	0.72
P(M = 0)	= 1 - P(M = 1)	0.28
P(Is_Self_Emp = 1)	= # Self-Emp Count / Total Obs	0.17
P(Is_Self_Emp = 0)	= 1 - P(Self_Emp = 1)	0.83
P(Target = 1)	= # Target 1 / Total Obs	0.09
P(Target = 0)	= 1 - P(Target = 1)	0.91

Conditional Probabilities		
P(M = 1 | T = 1)	= 934 / 1237	0.76
P(M = 0 | T = 1)	= 1 - P(M = 1 | T = 1)	0.24
P(Self_Emp = 1 | T = 1)	= (374) / 1237	0.30
P(Self_Emp = 0 | T = 1)	= 1 - P(Self_Emp = 1 | T = 1)	0.70

Naïve Bayes Calculation			Corresponding Values based on BernoulliNB Python Package
P(T = 1 | M=1 . Self_Emp=1)	P(M=1 | T=1) * P(Self_Emp=1 | T=1) * P(T=1) / P(M=1) * P(Self_Emp=1)	0.165562	0.164815
P(T = 1 | M=0 . Self_Emp=1)	P(M=0 | T=1) * P(Self_Emp=1 | T=1) * P(T=1) / P(M=0) * P(Self_Emp=1)	0.135077	0.136413
P(T = 1 | M=1 . Self_Emp=0)	P(M=1 | T=1) * P(Self_Emp=0 | T=1) * P(T=1) / P(M=1) * P(Self_Emp=0)	0.078247	0.078329
P(T = 1 | M=0 . Self_Emp=0)	P(M=0 | T=1) * P(Self_Emp=0 | T=1) * P(T=1) / P(M=0) * P(Self_Emp=0)	0.063840	0.063694

Gaussian Naïve Bayes

Gaussian naive Bayes [edit]

When dealing with continuous data, a typical assumption is that the continuous values associated with each class are distributed according to a Gaussian distribution. For example, suppose the training data contains a continuous attribute, x. We first segment the data by the class, and then compute the mean and variance of x in each class. Let μ_k be the mean of the values in x associated with class C_k, and let σ_k^2 be the variance of the values in x associated with class C_k. Suppose we have collected some observation value v. Then, the probability *distribution* of v given a class C_k, $p(x = v \mid C_k)$, can be computed by plugging v into the equation for a Normal distribution parameterized by μ_k and σ_k^2. That is,

$$p(x = v \mid C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}}\, e^{-\frac{(v-\mu_k)^2}{2\sigma_k^2}}$$

See Naïve_Bayes.ipynb file for Gaussian Naïve Bayes Model Development in Python

Another common technique for handling continuous values is to use binning to discretize the feature values, to obtain a new set of Bernoulli-distributed features; some literature in fact suggests that this is necessary to apply naive Bayes, but it is not, and the discretization may throw away discriminative information.[4]

https://en.wikipedia.org/wiki/Naive_Bayes_classifier

Advantages and Disadvantages of Naïve Bayes

- Advantages
  - Too fast in processing at time of prediction
  - Very Simple to Understand
  - Can be trained with small data

- Dis-advantages
  - Makes strong assumption that any two features are independent
  - Continuous variable has to be discretized. Alternatively one may use Gaussian Distribution for the likelihoods

Thank you

Contact us:

ar.jakhotia@k2analytics.co.in