UNIVERSITY OF NORTH CAROLINA AT CHARLOTTE

ITCS-6156-00

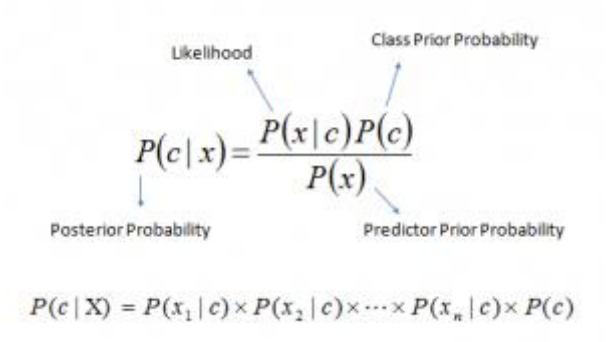
Naïve Bayes

Assignment 5 – Report

Archit Parnami

4/14/2017

Implementing a Naive Bayes classifier



The BayesClassifier class

- The input to the BayesClassifier are:
  - Data points labelled with a class (X, Y)
    - Where X are the input features
    - Y is the output class
- To Train call **Fit** method with X and Y.
- To predict call **predict** method with test input Xt and it should return the classified labels.

BayesClassifier
+number_of_examples : int +number_of_classes : int +number_of_attributes: int +occurances : {int : {int : {int:int}}}
-initialize([[int]],[int]): void - classify_data([[int]], [int]) : {int : [[int]]} - calculate_occurances({int:[[int]]}) : : {int : {int : {int:int}}} - calculate_probability([int]) : [(int, float)] - get_class_with_max_prob([(int, float)] : int - classify([int]) : int + **fit**([[int]], [int]) : void + **predict**([[int]]) : [int]

Dataset 1 - Optical Recognition of Handwritten Digits

1. Features
    - Each feature in the dataset represents an element of 8x8 matrix used to describe an Image.
    - Number of Features = 64
    - Range of values of each feature is 0 to 16
2. Output
    - Number ranging from 0 to 9

Distribution of Classes

Output Class	Frequency
0	376
1	389
2	380
3	389
4	387
5	376
6	377
7	387
8	380
9	382

Implementation Results

The model was trained using the custom implementation of Bayes Classifier and the accuracy of **89.7** was achieved.

The BayesGaussianClassifier class

- The input to the BayesGaussianClassifier are:
  - Data points labelled with a class (X, Y)
    - Where X are the input features
    - Y is the output class
- To Train call **Fit** method with X and Y.
- To predict call **predict** method with test input Xt and it should return the classified labels.
- It uses Gaussian probability density function to calculate the probability of the unseen samples.

BayesGaussianClassifier
+number_of_examples : int +number_of_classes : int +number_of_attributes: int +means : {int : {int : float}} +stddevs : {int : {int : float}} +output_classes : [int] +classified_data : {int : [[int]]}
-initialize([[int]],[int]): void - classify_data([[int]], [int]) : {int : [[int]]} - calculate_mean({int: [[int]]}) : {int : {int : float}} - calculate_stddev({int: [[int]]}, {int : {int : float}}) : {int : {int : float}} - calculate_gaussian_probability([int]) : [(int, float)] - get_class_with_max_prob([(int, float)] : int - classify([int]) : int + **fit**([[int]], [int]) : void + **predict**([[int]]) : [int]

The probability density of the normal distribution is:

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Where:

- μ is mean or expectation of the distribution (and also its median and mode).
- σ is standard deviation
- σ^2 is variance

Dataset 2 – Amazon reviews sentiment Analysis

1. Features
    - Product name and review
    - Number of features = 2
2. Output
    - Rating from 0 to 5
3. Number of Observations = **146824**

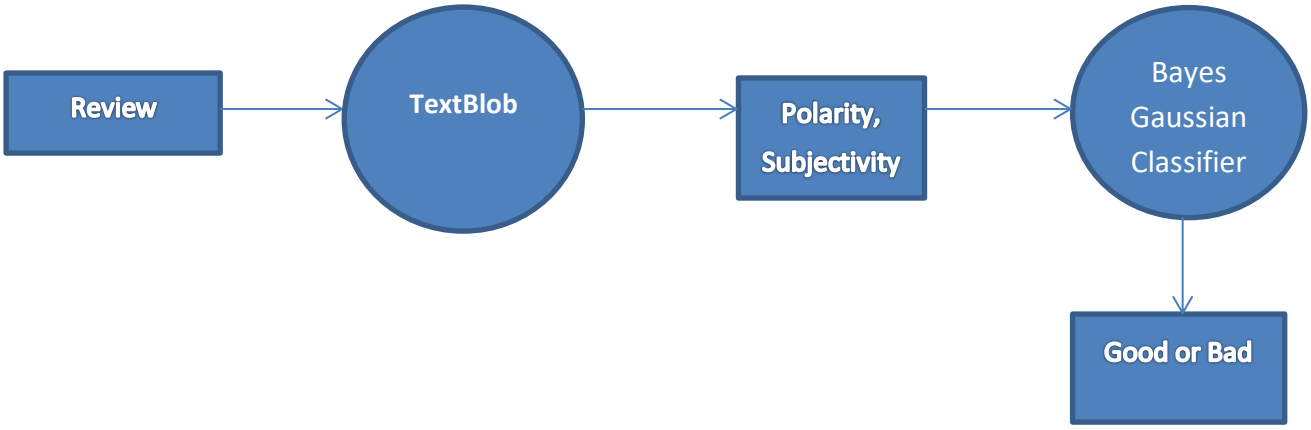
Original Problem: Given a review of a product predict the rating.
Modified Problem: Given a review of a product rate the product as **good** or **bad**.

Problem Solving Approach:

- The polarity and subjectivity of a review is obtained by performing sentiment analysis with a 3rd party library called **TextBlob**.
- If a review has a rating < 3 then it is considered bad(-1).
- If a review has a rating >=3 then it is considered good(1).
- Polarity & Subjectivity are then fed to a **BayesGaussianClassifier** as input features.
- While rating of -1 & 1 is used to represent negative & positive output respectively.

Implementation Steps

1. Sentiment Analysis & Input Generation
    - Input File: amazon_baby_train.csv, amazon_baby_test.csv
    - Output File: Train-SentimentAnalysis.csv, Test-SentimentAnalysis.csv
    - Library used for finding the Sentiment Analysis: TextBlob



Sample Input

Name	Review	Rating
Moby Wrap Original 100% Cotton Baby Carrier, Red	Bought this for my daughter….	5
Child to Cherish Handprints Tower Of Time Kit in Pink	It is very cute, and I got a lot of compliments….	4
JJ Cole Lite Embroidered Bundleme, Pink, Infant	This product is very pretty but does not fit the Graco Safe Seat	1

Sample Output

Polarity	Subjectivity	Rating
0.347	0.688	1
0.235	0.56	1
0.091	0.46	-1

2. Model Generation
    - Input Files: Train-SentimentAnalysis.csv, Test-SentimentAnalysis.csv
    - **Problem Statement**
        - Given the polarity, subjectivity and the rating of a review feed the data to BayesGaussianClassifier.
        - Use this BayesGaussianClassifier to predict the rating of new reviews

3. Implementation Results

The model was trained using the custom implementation of Bayes Gaussian Classifier and the accuracy of **85.64** was achieved.