Naïve Bayes and KNN

Learning Objectives

- Naïve Bayes theory
- Hands-on exercise Naïve Bayes
- KNN Theory
- Hands on exercise for KNN

Naïve Bayes theory

- Naïve Bayes classification is a form of classification that relies on the Bayes theorem.
- And Bayes theorem is a theorem in probability that tells us how we would re-visit the probability of an event given that we have more information.
- Bayes theorem takes information and constructs beliefs for the future.
- In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

Naïve Bayes Contd

A -> event

$P(A) = 0.5$ $P(\bar{A}) = 1 - P(A)$

For eg: Rolling a 6 face dice

A -> event that you roll a 1

B -> event that you roll an odd number

$P(A) = 1/6$ $P(B) = 3/6 = ½$

Conditional Probability : Probability of event A conditioned on the probability that event B happened

$P(A/B) = P(A \text{ and } B) / P(B) = (1/6) / (1/2) = 1/3$

Naïve Bayes contd.

A -> event that you roll a 1 on dice 1

B -> event that you roll a 1 on dice 2

P(A and B) = 1/36 = P(A) * P(B)

P(A/B) = 1/6 = P (A and B) / P(B) = (P(A)*P(B)) / P(B) = P(A)

P(B/A) = 1/6 = P(B)

Bayes Theorem

Bayes theorem

- P(A|B), reads “A given B,” represents the probability of A if B was known to have occurred.
- In many situations we would like to understand the relation between P(A|B) and P(B|A).
-

  You are planning an outdoor event tomorrow. When it actually rains, the weatherman correctly forecasts rain 90% of the time. When it doesn't rain, he incorrectly forecasts rain 10% of the time. Historically it has rained only 5 days each year. Unfortunately, the weatherman has predicted rain for tomorrow. What is the probability that it will rain tomorrow?

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

Naïve Bayes classification

Example

- Will my flight be on time? It is sunny, hot, Normal Humidity, and not windy!
- Data from the last several times we took this flight.

OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	Flight On Time
Rainy	Hot	High	0	No
Rainy	Hot	High	1	Yes
Overcast	Hot	High	0	Yes
Sunny	Mild	High	0	No
Sunny	Cool	Normal	0	Yes
Sunny	Cool	Normal	1	No
Overcast	Cool	Normal	1	Yes
Rainy	Mild	High	0	No
Rainy	Cool	Normal	0	Yes
Sunny	Mild	Normal	0	Yes
Rainy	Mild	Normal	1	Yes
Overcast	Mild	High	1	Yes
Overcast	Hot	Normal	0	Yes
Sunny	Mild	High	1	No

Sunny Hot Normal 0 ?

NB Classification contd.

P(Flight on time | Sunny, Hot, Normal humidity, 0) =

P(S, H, N, 0 | Time) P(Time) / P(S, H, N, 0)

P(Time) = 9/14

P(S, H, N, 0 | Time) = P(S|Time)* P(H|Time)* P(N|Time)* P(0|Time)

= 2/9 * 2/9 * 6/9 * 6/9

= 144/6561

P(S, H, N, 0 | Time) P(Time) = 0.0141

P(S, H, N, 0) = P(S, H, N, 0 | Time) P(Time) + P(S, H, N, 0 | $Time_{bar}$) P($Time_{bar}$)

= 0.0141 + (3/5 * 2/5 * 1/5 * 2/5 * 5/14) = 0.0068

P(Flight on time | Sunny, Hot, Normal humidity, 0) = 67%

Naïve Bayes Classifiers

- Probabilistic models based on Bayes' theorem.
- It is called "naive" due to the assumption that the features in the dataset are mutually independent.
- In real world, the independence assumption is often violated, but naïve Bayes classifiers still tend to perform very well.
- Idea is to factor all available evidence in form of predictors into the naïve Bayes rule to obtain more accurate probability for class prediction.
- It estimates conditional probability which is the probability that something will happen, given that something else has already occurred. For e.g. the given mail is likely a spam given appearance of words such as "prize"

Naïve Bayes Classifiers - Pros and Cons

- Advantages
  - Simple, Fast in processing and effective
  - Does well with noisy data and missing data
  - Requires few examples for training (assuming the data set is a true representative of the population)
  - Easy to obtain estimated probability for a prediction
- Dis-advantages
  - Relies on and often incorrect assumption of independent features
  - Not ideal for data sets with large number of numerical attributes
  - Estimated probabilities are less reliable in practice than predicted classes
  - If rare events are not captured in the training set but appears in the test set the probability calculation will be incorrect

Gaussian Naive Bayes classifier

- When some of our independent variables are continuous we cannot calculate conditional probabilities!

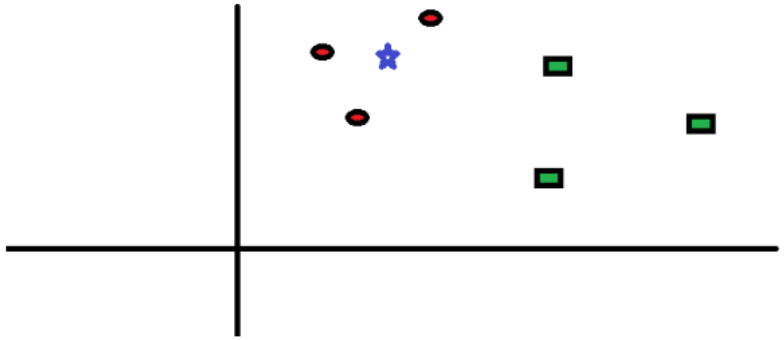
- In Gaussian Naive Bayes, continuous values associated with each feature (or independent variable) are assumed to be distributed according to a Gaussian distribution.

- All we would have to do is estimate the mean and standard deviation of the continuous variable.

KNN

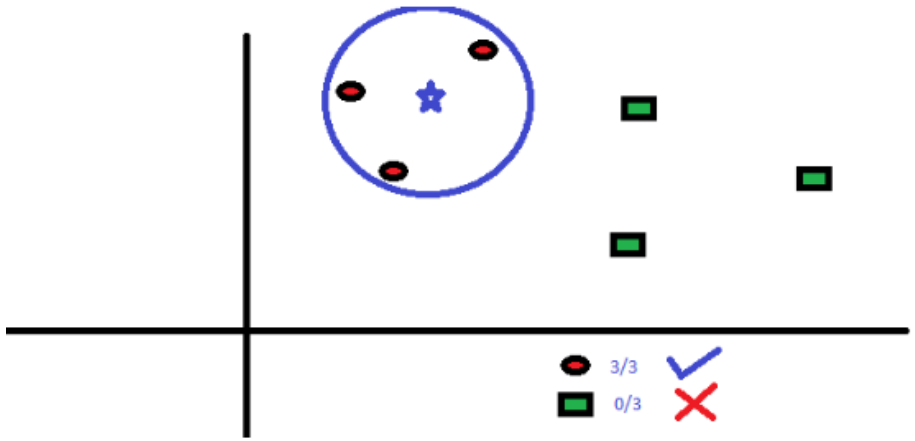
- KNN (K — Nearest Neighbors) is one of many (supervised learning) algorithms used in data mining and machine learning.
- It’s a classifier algorithm where the learning is based “how similar” is a data (a vector) from other.
- Let’s take a simple case to understand this algorithm. Following is a spread of red circles (RC) and green squares (GS) :

KNN Contd.



You intend to find out the class of the blue star (BS) . BS can either be RC or GS and nothing else. The “K” is KNN algorithm is the nearest neighbors we wish to take vote from. Let’s say K = 3. Hence, we will now make a circle with BS as center just as big as to enclose only three data points on the plane. Refer to following diagram for more details:

KNN Contd.



- The three closest points to BS is all RC. Hence, with good confidence level we can say that the BS should belong to the class RC. Here, the choice became very obvious as all three votes from the closest neighbor went to RC. The choice of the parameter K is very crucial in this algorithm.

Classification Steps

The KNN's steps are:

1. Receive an unclassified data;
2. Measure the distance (Euclidian, Manhattan, Minkowski or Weighted) from the new data to all others data that is already classified;
3. Gets the K(K is a parameter that you define) smaller distances;
4. Check the list of classes had the shortest distance and count the amount of each class that appears;
5. Takes as correct class the class that appeared the most times;
6. Classifies the new data with the class that you took in step 5;

Distance

Distance measure is important

- Most commonly distance is measured using Euclidean distance
- We should always Normalize data
- Other distance measurement methods include

  - Manhattan distance
  - Minkowski distance
  - Mahalanobis distance
  - Cosine similarity

Calculating distance

greatlearning

To calculate the distance between two points is very simple, there are several ways to get this value, here we will use the Euclidean distance.

The Euclidean distance's formula is:

$$d(\mathbf{p}, \mathbf{q}) = d(\mathbf{q}, \mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

Distances contd.

greatlearning

Manhattan distance

Definition: The **distance** between two points measured along axes at right angles. In a plane with $p_1$at (x_1, y_1) and p_2 at (x_2, y_2), it is $|x_1 - x_2| + |y_1 - y_2|$.

Minkowski distance

It is a metric in a normed vector space. Minkowski distance is used for distance similarity of vector. Given two or more vectors, find distance similarity of these vectors. Mainly Minkowski distance is applied in machine learning to find out distance similarity.

Cosine similarity

It is a measure of **similarity** between two non-zero vectors of an inner product space that measures the**cosine** of the angle between them. The **cosine** of 0° is 1, and it is less than 1 for any angle in the interval $(0,\pi]$ radians.

Other Variants

Radius Neighbor Classifier

- Implements learning based on number of neighbors within a fixed radius r of each training point, where r is a floating point value specified by the user
- May be a better choice when the sampling is not uniform. However, when there are many attributes and data is sparse, this method becomes ineffective due to curse of dimensionality.

KD Tree nearest neighbour

- Approach helps reduce the computation time.
- Very effective when we have large data points but still not too many dimensions

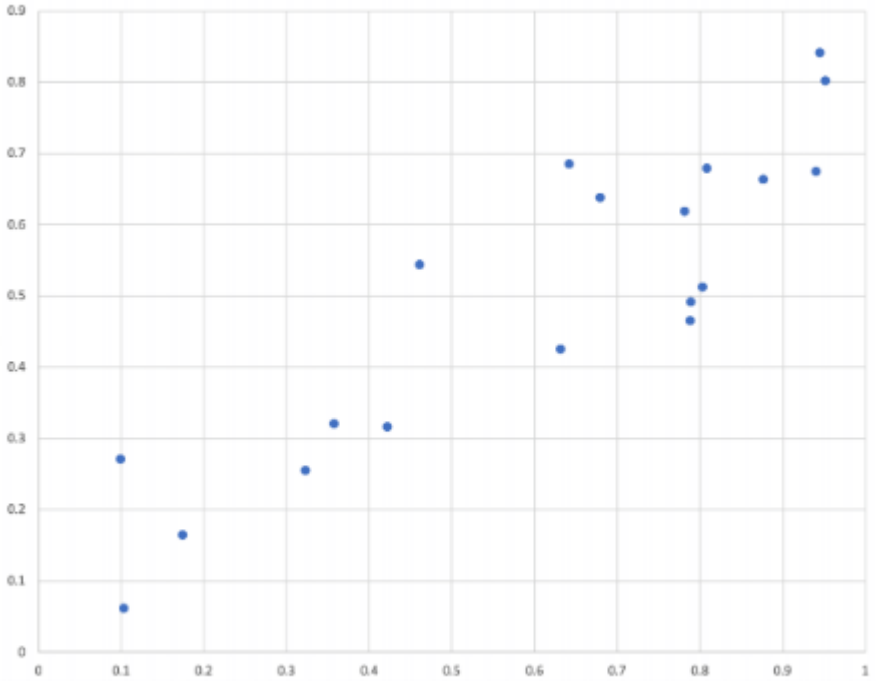
KNN features

K-NN

- It does not construct a “model”. Known as a non-parametric method.
- Classification is computed from a simple majority vote of the nearest neighbors of each point
- Suited for classification where relationship between features and target classes is numerous, complex and difficult to understand and yet items in a class tend to be fairly homogenous on the values of attributes
- Not suitable if the data is too noisy and the target classes do not have clear demarcation in terms of attribute values
- Can also be used for regression.

K-NN for regression

The Neighbors based algorithm can also be used for regression where the labels are continuous data and the label of query point can be average of the labels of the neighbors.



K Nearest Neighbours

Pros and **C**ons

Advantages

- Makes no assumptions about distributions of classes in feature space
- Can work for multi classes simultaneously
- Easy to implement and understand
- Not impacted by outliers.

Dis-advantages

- Fixing the optimal value of K is a challenge
- Will not be effective when the class distributions overlap
- Does not output any models. Calculates distances for every new point(lazy learner)•Computationally intensive

Case Studies

greatlearning

Let us now have case studies for the above two topics.

Case Study 1 - NB

greatlearning

Objective:

To predict whether income exceeds 50K/yr based on census data.

Feature description:

Age: continuous

Workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.

Fnlwgt: continuous.

Education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.

Education-num: continuous.

Marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.

greatlearning

Feature description contd.

Occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.

Relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.

Race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.

Sex: Female, Male.

Capital-gain: continuous.

Capital-loss: continuous.

Hours-per-week: continuous.

Native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, etc.

Steps to follow

greatlearning

1. Import the libraries.
2. Get the data.
3. Add headers to the dataframe
4. Handle missing data
5. Perform Data preprocessing by duplicating copy of the original dataframe.
6. Perform Hot encoding
7. Initialize the encoded categorical columns
8. Split the data into train and test
9. Implement Gaussian Naive Bayes
10. Calculate Accuracy

Case Study 2 - KNN

greatlearning

Context:

The dataset to be audited was provided which consists of a wide variety of intrusions simulated in a military network environment It created an environment to acquire raw TCP/IP dump data for a network by simulating a typical US Air Force LAN. The LAN was focused like a real environment and blasted with multiple attacks. For each TCP/IP connection, 41 quantitative and qualitative features are obtained from normal and attack data (3 qualitative and 38 quantitative features) .

The class variable has two categories:

- Normal
- Anomalous

Dataset Info

greatlearning

Dataset:

https://www.kaggle.com/what0919/intrusion-detection

Data basically represents the packet data for a time duration of 2 seconds.

1-9 Columns: basic features of packet (type 1)

10-22 columns: employ the content features (type 2)

23-31 columns: employ the traffic features with 2 seconds of time window (type 4)

32-41 columns: employ the host based features

Objective and Steps

greatlearning

To detect Network Intrusion using KNN

Steps:

1. Import Libraries and Data
2. Data Preparation and analysis(standardization)
3. Split the dataset into training and test datasets
4. Build the model and train and test on training and test sets respectively using scikit-learn. Print the Accuracy of the model with different values of k=3,5,9.
5. Cross Validation

greatlearning

HappyLearning

Questions?