Lab # 12

National University of Computer & Emerging Sciences

Lab Manual

CS461: Artificial Intelligence Lab

Course Instructor	Dr. Hafeez-Ur-Rehman
Lab Instructor	Muhammad Hamza
Semester	Spring 2022

Lab # 12

Machine Learning

Machine Learning

Machine learning is subtype of Artificial Intelligence.
"[Machine Learning is the] field of study that gives computers the ability to learn without being explicitly programmed."

"A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E." -- Tom Mitchell, Carnegie Mellon University

ML solves problems that cannot be solved by numerical means alone.

Supervised machine learning

The program is "trained" on a pre-defined set of "training examples" with given class labels, which then facilitate its ability to reach an accurate conclusion when given new data.

- Supervised machine learning is reliable.

Examples:

1. Logistic Regression (Output is discrete e-g 0 or 1)
2. Linear Regression (Output is continuous e-g 2.34, 122)
3. Decision Trees
4. K – Nearest Neighbors (KNN)
5. Support Vector Machines (SVMs)

Unsupervised machine learning

The program is given a bunch of data and must find patterns and relationships therein.

- Unsupervised machine learning is quick.

Examples:

1. Clustering
2. Autoencoders
3. GANs
4. Dimensionality reduction

Supervised Learning (Classification)

K-Nearest Neighbor Algorithm: Numerical Example of K Nearest Neighbor Algorithm

Here is step by step on how to compute K-nearest neighbors KNN algorithm:

1. Determine parameter K = number of nearest neighbors. "K" should be an Odd, it helps in picking majority votes. If K=4 => 2 rows have label '0" and 2 rows have label "1", so it is very difficult to pick majority label.
2. Calculate the distance between the query-instance and all the training samples
3. Sort the distance and determine nearest neighbors based on the K-th minimum distance
4. Gather the Y (labels) of only nearest neighbors. Use simple majority of the Y (labels) of nearest neighbors as the prediction value of the query instance

Lab # 12

We have data from the questionnaires survey (to ask people opinion) and objective testing with two attributes (acid durability and strength) to classify whether a special paper tissue is good or not. Here is four training samples.

Rows = Instances = Records = Training Examples
Columns = Features = Attributes
Column Y = Output Variable = Classification Label (could be binary/multi-class)
Euclidian distance measure is used in this example

Lab # 12

X1 = Acid Durability	X2 = Strength	Y = Classification
7	7	Bad
7	4	Bad
3	4	Good
1	4	Good

Now the factory produces a new paper tissue that pass laboratory test with X1 = 3 and X2 = 7. Without another expensive survey, can we guess what the classification of this new tissue is?

1. Determine parameter K = number of nearest neighbors; Suppose use K = 3
2. Calculate the distance between the query-instance and all the training samples
   a. Coordinate of query instance is (3, 7), instead of calculating the distance we compute square distance which is faster to calculate (without square root)

X1 = Acid Durability	X2 = Strength	Euclidian Distance with query (3,7)
7	7	$(7-3)^2+(7-7)^2=16$
7	4	$(7-3)^2+(4-7)^2=25$
3	4	$(3-3)^2+(4-7)^2=9$
1	4	$(1-3)^2+(4-7)^2=13$

3. Sort the distance and determine nearest neighbors based on the K-th minimum distance

X1 = Acid Durability	X2 = Strength	Euclidian Distance with query (3,7)	Rank Min. Distance	Included
7	7	$(7-3)^2+(7-7)^2=16$	3	Yes
7	4	$(7-3)^2+(4-7)^2=25$	**4**	**No**
3	4	$(3-3)^2+(4-7)^2=9$	1	Yes
1	4	$(1-3)^2+(4-7)^2=13$	2	Yes

4. Gather the category of the nearest neighbors. Notice in the second-row last column that the category of nearest neighbor (Y) is not included because the rank of this data is more than 3 (=K).

X1 = Acid Durability	X2 = Strength	Euclidian Distance with query (3,7)	Rank Min. Distance	Included	Y = Label
7	7	$(7-3)^2+(7-7)^2=16$	3	Yes	Bad
7	4	$(7-3)^2+(4-7)^2=25$	**4**	**No**	**-**
3	4	$(3-3)^2+(4-7)^2=9$	1	Yes	Good
1	4	$(1-3)^2+(4-7)^2=13$	2	Yes	Good

5. Use simple majority of the category of nearest neighbors as the prediction value of the query instance. We have 2 good and 1 bad, since 2 > 1, then we conclude that a new paper tissue that pass laboratory test with X1 = 3 and X2 = 7 is included in Good category.

Advantages

- K-NN is simple:
- K-NN has no assumptions:
- K-NN is a non-parametric algorithm which means there are assumptions to be met to implement K-NN. Parametric models like linear regression have lots of assumptions to be met by data before it can be implemented which is not the case with K-NN.
- No Training Step
- Very easy to implement for multi-class problem
- Variety of distance criteria to be choose from: K-NN algorithm gives user the flexibility to choose distance while building K-NN model.
  - ✓ Euclidean Distance
  - ✓ Hamming Distance
  - ✓ Manhattan Distance
  - ✓ Minkowski Distance

Disadvantages

- Only works for numerical data
- K-NN is a slow algorithm
- Curse of Dimensionality: For high dimensional data, it's a bad choice.
- K-NN needs homogeneous features
- Optimal number of neighbors
- Imbalanced data causes problems
- Cannot handle outlier
- Missing Value treatment

Task

a. Download the dataset from **Google classroom** folder i.e. fruit_data_with_colors.

b. Remove the features having text/categorical values.

c. Fill the missing values by mean value of each column separately, if any.

d. Select the value of "K" as any even number and observe the difference i.e., 4,6,8 etc.

e. Implement KNN algorithm using python from scratch, you can only use Numpy or Pandas.

f. Use first 50 rows as training samples and remaining 10 rows for Testing to predict their labels?