**1 FIND – S**

**Implement and demonstrate the FIND-S algorithm for finding the most specific hypothesis based on a given set of training data samples. Read the training data from a .CSV file.**

import csv

with open('find-s-training-examples.csv') as csvfile:

data = [line[:-1] for line in csv.reader(csvfile) if line[-1] == 'Y']

print('Positive training examples are: {}'.format(data))

S = ['$'] \* len(data[0])

print('\nOutput at each step is \n{}'.format(S))

for example in data:

i = 0

for feature in example:

**S[i] = feature if S[i] == '$' or S[i] == feature else '?'**

i += 1

print(S)

**2 CANDIDATE ELIMINATION**

**For a given set of training data examples stored in a .CSV file, implement and demonstrate the Candidate-Elimination algorithm to output a description of the set of all hypotheses consistent with the training examples.**

**def consistent(h1, h2)**

**def candidateElimination()**

import csv

with open('candidate-elimination-training-examples.csv') as file:

data = [**tuple(line)** for line in csv.reader(file)]

D = []

for i in range(len(data[0])):

**D.append(list(set([ele[i] for ele in data])))**

def consistent(h1, h2):

for x, y in zip(h1, h2):

**if not (x == '?' or (x != '$' and (x == y or y == '$'))):**

return False

return True

def candidateElimination():

G = {('?',) \* (len(data[0]) - 1),}

S = ['$'] \* (len(data[0]) - 1)

num = 0

print('G[{0}]:'.format(num), G)

print('S[{0}]:'.format(num), S)

for item in data:

num += 1

inp, res = item[:-1], item[-1]

if res in 'Yy':

G = {g for g in G if consistent(g, inp)}

**i = 0**

for s, x in zip(S, inp):

if s != x:

S[i] = '?' if s != '$' else x

**i += 1**

else:

S = S

Gprev = G.copy()

for g in Gprev:

#if g not in G:

#continue

**for i in range(len(g)):**

**if g[i] == '?':**

**for val in D[i]:**

**if val != inp[i] and S[i] == val:**

g\_new = g[:i] + (val, ) + g[i + 1:]

G.add(g\_new)

else:

G.add(g)

**G.difference\_update([h for h in G if any([consistent(h, g1) for g1 in G if g1 != h])])**

print('G[{0}]:'.format(num), G)

print('S[{0}]:'.format(num), S)

candidateElimination()

**3 ID3**

**Write a program to demonstrate the working of the decision tree based ID3 algorithm. Use an appropriate data set for building the decision tree and apply this knowledge to classify a new sample.**

**def \_\_init\_\_(self, data, attribute = None)**

**def calculateEntropy(probs)**

**def splitData(aList, attribute, Class)**

**def entropy(aList, attribute = ‘PlayTennis’, Gain = False)**

**def informationGain(data)**

**def id3(root)**

import math

import pandas as pd

from collections import Counter

tennis = pd.DataFrame.from\_csv('id3-training-examples.csv')

tennis = tennis.sample(frac = 1).reset\_index(drop = True) #OPTIONAL - FOR SHUFFLE

**class Node:**

**def \_\_init\_\_(self, data, attribute = None):**

self.decision\_attribute = attribute

self.child = {}

self.data = data

self.decision = None

def calculateEntropy(probs):

return sum([-prob \* math.log(prob, 2) for prob in probs])

def splitData(aList, attribute, Class):

return **aList[aList[attribute] == Class]**

def entropy(aList, attribute = 'PlayTennis', Gain = False):

cnt = Counter(aList[attribute])

n = len(aList[attribute])

probs = [x / n for x in cnt.values()]

**if not Gain:**

**return calculateEntropy(probs)**

print(cnt.items())

gain = 0

**for Class, prob in zip(cnt.keys(), probs):**

**gain += -prob \* entropy(splitData(aList, attribute, Class))**

return gain

def informationGain(data):

maxGain = -1

maxGainAttribute = None

**for attribute in data.keys():**

**if attribute == 'PlayTennis':**

**continue**

**gain = entropy(data) + entropy(data, attribute, Gain = True)**

print("\n", attribute, ": %.4f" % gain)

if gain > maxGain:

maxGain = gain

maxGainAttribute = attribute

return maxGainAttribute

def id3(root):

**global nodes**

if len(root.data.keys()) == 1 or len(root.data) == 1:

**cnt = Counter(root.data['PlayTennis'])**

**root.decision = cnt.most\_common(1)[0][0]**

print("Decision =", root.decision)

return

maxGainAttribute = informationGain(root.data)

root.decision\_attribute = maxGainAttribute

print("------------------------------")

**for attribute in set(root.data[maxGainAttribute]):**

**childData = splitData(root.data, maxGainAttribute, attribute)**

**root.child[attribute] = Node(childData.drop([maxGainAttribute], axis = 1))**

**id3(root.child[attribute])**

**def predict(example, root):**

**if root.decision != None:**

**return root.decision**

**try:**

**prediction = predict(example, root.child[example[root.decision\_attribute]])**

**return prediction**

**except:**

**return "No"**

trainingData = tennis.iloc[: -4]

testData = tennis.iloc[-4: ]

root = Node(data = trainingData)

id3(root)

print('\nTest Data:\n', testData)

**predictions = [predict(test ,root) for \_,test in testData.iterrows()]**

correct = testData['PlayTennis']

print("Predictions:\n", predictions)

print("Actual:\n", testData['PlayTennis'])

**print("Accuracy: %.4f" % (sum([1 for x, y in zip(predictions, correct) if x == y]) / len(predictions)))**

**4 BACK PROPAGATION**

**Build an Artificial Neural Network by implementing the Backpropagation algorithm and test the same using appropriate data sets.**

**def initialize\_network(n\_inputs, n\_hidden, n\_outputs)**

**def activate(weights, inputs):**

**def forward\_propagate(network, row):**

**def backward\_propagate\_error(network, expected):**

**def update\_weights(network, row, l\_rate):**

**def train\_network(network, dataset, l\_rate, n\_epoch, n\_outputs):**

import random

import math

def initialize\_network(n\_inputs, n\_hidden, n\_outputs):

network = list()

hidden\_layer = [{'weights': [random.uniform(-0.5, 0.5) for i in range(n\_inputs + 1)]} for i in range(n\_hidden)]

network.append(hidden\_layer)

output\_layer = [{'weights': [random.uniform(-0.5, 0.5) for i in range(n\_inputs + 1)]} for i in range(n\_outputs)]

network.append(output\_layer)

i = 1

print("\nThe initialized neural network is:")

for layer in network:

j = 1

for neuron in layer:

print("\nLayer[%d] Node[%d]: \n" % (i, j), neuron)

j += 1

i += 1

return network

def activate(weights, inputs):

activation = weights[-1]

for i in range(len(weights) - 1):

activation += weights[i] \* inputs[i]

return activation

def forward\_propagate(network, row):

inputs = row

for layer in network:

new\_inputs = []

for neuron in layer:

activation = activate(neuron['weights'], inputs)

neuron['output'] = 1 / (1.0 + math.exp(-activation))

new\_inputs.append(neuron['output'])

inputs = new\_inputs

return inputs

def backward\_propagate\_error(network, expected):

for i in range(len(network) - 1, -1, -1):

layer = network[i]

errors = list()

if i != len(network) - 1:

for j in range(len(layer)):

error = 0.0

for neuron in network[i + 1]:

error += (neuron['weights'][j] \* neuron['delta'])

errors.append(error)

else:

for j in range(len(layer)):

errors.append(expected[j] - layer[j]['output'])

for j in range(len(layer)):

neuron = layer[j]

neuron['delta'] = errors[j] \* neuron['output'] \* (1 - neuron['output'])

def update\_weights(network, row, l\_rate):

for i in range(len(network)):

inputs = row[: -1]

if i != 0:

inputs = [neuron['output'] for neuron in network[i - 1]]

for neuron in network[i]:

for j in range(len(inputs)):

neuron['weights'][j] += l\_rate \* neuron['delta'] \* inputs[j]

neuron['weights'][-1] += l\_rate \* neuron['delta']

def train\_network(network, dataset, l\_rate, n\_epoch, n\_outputs):

print("\nNetwork Training Begins:")

for epoch in range(n\_epoch):

sum\_error = 0

for row in dataset:

outputs = forward\_propagate(network, row)

expected = [0 for i in range(n\_outputs)]

expected[row[-1]] = 1

sum\_error += sum([(expected[i] - outputs[i]) \*\* 2 for i in range(len(expected))])

backward\_propagate\_error(network, expected)

update\_weights(network, row, l\_rate)

print('>epoch = %d, l\_rate = %.3f, error = %.3f' % (epoch, l\_rate, sum\_error))

print('Network training ends.')

random.seed()

dataset = [ [2.7810836, 2.550537003, 0], [1.465489372, 2.362125076, 0], [3.396561688, 4.400293529, 0], [1.38807019, 1.850220317, 0], [3.06407232, 3.005305973, 0], [7.627531214, 2.759262235, 1], [5.332441248, 2.088626775, 1], [6.922596716, 1.77106367, 1], [8.675418651, -0.242068655, 1], [7.673756466, 3.508563011, 1] ]

print("The input dataset is:")

print(dataset)

n\_inputs = len(dataset[0]) - 1

print("\nNumber of inputs: ", n\_inputs)

n\_outputs = len(set([row[-1] for row in dataset]))

print("Number of outputs: ", n\_outputs)

network = initialize\_network(n\_inputs, 2, n\_outputs)

train\_network(network, dataset, 0.5, 20, n\_outputs)

print("\nFinal neural network:")

i = 1

for layer in network:

j = 1

for sub in layer:

print("\nLayer[%d] Node[%d]: " % (i, j), sub)

j += 1

i += 1

**5 NAÏVE BAYES**

**Write a program to implement the naïve Bayesian classifier for a sample training data set stored as a .CSV file. Compute the accuracy of the classifier, considering few test data sets.**

**def mean(numbers)**

**def stdev(numbers):**

**variance = sum([pow(x - avg, 2) for x in numbers]) / float(len(numbers) - 1)**

**def summarize(dataset)**

**def calcProb(summary, item)**

import csv

import math

def mean(numbers):

return sum(numbers) / len(numbers)

def stdev(numbers):

avg = mean(numbers)

**variance = sum([pow(x - avg, 2) for x in numbers]) / (len(numbers) - 1)**

return **math.sqrt(variance)**

def summarize(dataset):

summaries = [(mean(attribute), stdev(attribute)) for attribute in **zip(\*dataset)**]

del summaries[-1]

return summaries

def calcProb(summary, item):

prob = 1

for i in range(len(summary)):

x = item[i]

mean, stdev = summary[i]

**exponent = math.exp(-pow(x - mean, 2) / (2 \* stdev \*\* 2))**

**final = exponent / (math.sqrt(2 \* math.pi) \* stdev)**

prob \*= final

return prob

with open('naive-bayes-training-examples.csv') as file:

data = [line for line in csv.reader(file)]

**for i in range(len(data)):**

**data[i] = [float(x) for x in data[i]]**

split = int(0.90 \* len(data))

train = data[:split]

test = data[split:]

print('\nTotal number of hypotheses:', len(data))

print('Number of hypotheses in training data:', len(train))

print('Number of hypotheses in test data:', len(test))

print("\nThe values assumed for the concept learning attributes are:")

print("OUTLOOK: Sunny = 1, Overcast = 2 and Rain = 3\nTEMPERATURE: Hot = 1, Mild = 2 and Cool = 3\nHUMIDITY: High = 1 and Normal = 2\nWIND: Weak = 1 and Strong = 2")

print("TARGET CONCEPT: PlayTennis where Yes = 10 and No = 5")

print("\nTraining dataset:")

for x in train:

print(x)

print("\nTest dataset:")

for x in test:

print(x)

yes = []

no = []

for i in range(len(train)):

if data[i][-1] == 5.0:

no.append(data[i])

else:

yes.append(data[i])

yes = summarize(yes)

no = summarize(no)

predictions = []

for item in test:

yesProb = calcProb(yes, item)

noProb = calcProb(no, item)

predictions.append(10.0 if(yesProb > noProb) else 5.0)

correct = 0

for i in range(len(test)):

if(test[i][-1] == predictions[i]):

correct += 1

print("\nActual values are:")

for i in range(len(test)):

print(test[i][-1], end=" ")

print("\nPredicted values are:")

for i in range(len(predictions)):

print(predictions[i], end=" ")

print("\nAccuracy is %.1f%%" % ((correct / len(test)) \* 100))

**6 NAÏVE BAYES TEXT CLASSIFIER**

**Assuming a set of documents that need to be classified, use the naïve Bayesian Classifier model to perform this task. Built-in Java classes/API can be used to write the program. Calculate the accuracy, precision, and recall for your data set.**

from sklearn.datasets import fetch\_20newsgroups

from sklearn.metrics import **confusion\_matrix, accuracy\_score, classification\_report**

from sklearn.feature\_extraction.text import CountVectorizer, TfidfTransformer

from sklearn.naive\_bayes import MultinomialNB

train = fetch\_20newsgroups(subset = 'train', shuffle = True)

**print('The categories of 20NewsGroups are:')**

for cat in train.target\_names:

print(cat)

**categories = ['alt.atheism', 'soc.religion.christian', 'comp.graphics', 'sci.med']**

train = fetch\_20newsgroups(subset = 'train', categories = categories, shuffle = True)

test = fetch\_20newsgroups(subset = 'test', categories = categories, shuffle = True)

**countVectorizer = CountVectorizer()**

traintf = countVectorizer.**fit\_transform(train.data)**

print('\ntf train count:', traintf.shape)

testtf = countVectorizer.**transform(test.data)**

print('tf test count:', testtf.shape)

**tfidftransformer = TfidfTransformer()**

traintfidf = tfidftransformer.**fit\_transform(traintf)**

print('\ntf train count:', traintf.shape)

testtfidf = tfidftransformer.**transform(testtf)**

print('tf test count:', testtf.shape)

**model = MultinomialNB()**

**model.fit(traintfidf, train.target)**

**predicted = model.predict(testtfidf)**

**print('Accuracy score:', accuracy\_score(test.target, predicted))**

**print(classification\_report(test.target, predicted, target\_names = test.target\_names))**

**print('Confusion Matrix: ', confusion\_matrix(test.target, predicted))**

**7 BAYESIAN BELIEF NETWORK**

**Write a program to construct a Bayesian network considering medical data. Use this model to demonstrate the diagnosis of heart patients using standard Heart Disease Data Set. You can use Java/Python ML library classes/API.**

import numpy as np

import pandas as pd

import urllib

from urllib.request import urlopen

import pgmpy

**from pgmpy.models import BayesianModel**

**from pgmpy.estimators import MaximumLikelihoodEstimator**

**from pgmpy.inference import VariableElimination**

**url = 'http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/processed.hungarian.data'**

**np.set\_printoptions(threshold = np.nan)**

**names = ['age', 'sex', 'cp', 'trestbps', 'chol', 'fbs', 'restecg', 'thalach', 'exang', 'oldpeak', 'slope', 'ca', 'thal', 'heartdisease']**

heartDisease = **pd.read\_csv**(urlopen(url), names = names)

print(heartDisease.head())

del heartDisease['oldpeak']

del heartDisease['slope']

del heartDisease['ca']

del heartDisease['thal']

heartDisease = heartDisease.replace('?', np.nan)

print(heartDisease.dtypes)

**model = BayesianModel([('age', 'trestbps'), ('age', 'fbs'), ('sex', 'trestbps'), ('sex', 'trestbps'), ('exang', 'trestbps'), ('trestbps', 'heartdisease'), ('fbs', 'heartdisease'), ('heartdisease', 'restecg'), ('heartdisease', 'thalach'), ('heartdisease', 'chol')])**

model.fit(heartDisease, estimator = MaximumLikelihoodEstimator)

print(model.get\_cpds('age'))

print(model.get\_cpds('chol'))

print(model.get\_cpds('sex'))

**model.get\_independencies()**

**inference = VariableElimination(model)**

q = inference.query(variables = ['heartdisease'], **evidence = {'age': 28}**)

print(q['heartdisease'])

q = inference.query(variables = ['heartdisease'], **evidence = {'chol': 100}**)

print(q['heartdisease'])

**8 KMEANS AND EM**

**Apply EM algorithm to cluster a set of data stored in a .CSV file. Use the same data set for clustering using *k*-Means algorithm. Compare the results of these two algorithms and comment on the quality of clustering. You can add Java/Python ML library classes/API in the program.**

import numpy as np

import pandas as pd

**from sklearn.cluster import KMeans**

**from sklearn.mixture import GaussianMixture**

from sklearn.datasets import load\_iris

from sklearn.metrics import accuracy\_score, confusion\_matrix

**from sklearn import preprocessing**

**import matplotlib.pyplot as plt**

l1 = [0, 1, 2]

**def rename(S):**

**l2 = []**

**for i in S:**

**if i not in l2:**

**l2.append(i)**

**for i in S:**

**pos = l2.index(i)**

**i = l1[pos]**

**return S**

iris = load\_iris()

print('Data', iris.data)

print('Target names:', iris.target\_names)

print('Target:', iris.target)

**x = pd.DataFrame(iris.data)**

x.columns = ['SepalLength', 'SepalWidth', 'PetalLength', 'PetalWidth']

**y = pd.DataFrame(iris.target)**

y.columns = ['Targets']

**model = KMeans(n\_clusters = 3)**

**model.fit(x)**

plt.figure(figsize = (14, 7))

colormap = np.array(['red', 'lime', 'black'])

plt.subplot(1, 2, 1)

plt.scatter(x.PetalLength, x.PetalWidth, c = colormap[**y.Targets**], s = 40)

plt.title(**'Real Classification'**)

plt.subplot(1, 2, 2)

plt.scatter(x.PetalLength, x.PetalWidth, c = colormap[**model.labels\_**], s = 40)

plt.title(**'KMeans Classification'**)

plt.show()

**km = rename(model.labels\_)**

print('What KMeans thought:', km)

print('Accuracy score of KMeans:', accuracy\_score(y, km))

print('Confusion matris of KMeans:', confusion\_matrix(y, km))

**scaler = preprocessing.StandardScaler()**

**scaler.fit(x)**

**xsa = scaler.transform(x)**

**xs = pd.DataFrame(xsa, columns = x.columns)**

**print('\n', xs.sample(5))**

**gmm = GaussianMixture(n\_components = 3)**

**gmm.fit(xs)**

**y\_cluster\_gmm = gmm.predict(xs)**

plt.subplot(1, 2, 1)

plt.scatter(x.PetalLength, x.PetalWidth, c = colormap[**y\_cluster\_gmm**], s = 40)

plt.title(**'GMM Classification'**)

plt.show()

**em = rename(y\_cluster\_gmm)**

print('What EM thought:', km)

print('Accuracy score of EM:', accuracy\_score(y, em))

print('Confusion matris of EM:', confusion\_matrix(y, em))

**9 K NEAREST NEIGHBOUR**

**Write a program to implement *k*-Nearest Neighbour algorithm to classify the iris data set. Print both correct and wrong predictions. Java/Python ML library classes can be used for this problem.**

import numpy as np

from sklearn.datasets import load\_iris

**from sklearn.model\_selection import train\_test\_split**

**from sklearn.neighbors import KNeighborsClassifier**

dataset = load\_iris()

print('IRIS FEATURES | TARGET NAMES:', dataset.target\_names)

print('\nData:', dataset["data"])

print('\nTarget', dataset["target"])

**xtrain, xtest, ytrain, ytest = train\_test\_split(dataset["data"], dataset["target"], random\_state = 0)**

print("\nX TRAIN \n", xtrain)

print("\nX TEST \n", xtest)

print("\nY TRAIN \n", ytrain)

print("\nY TEST \n", ytest)

kn = KNeighborsClassifier(n\_neighbors = 1)

kn.fit(xtrain, ytrain)

predictions = kn.predict(xtest)

for i in range(len(xtest)):

print("\nActual: {0} {1} \nPredicted: {2} {3}".format(ytest[i], dataset["target\_names"][ytest[i]], predictions, dataset["target\_names"][predictions]))

print("\nTEST SCORE[ACCURACY]: {:.2f}\n".format(kn.score(xtest, ytest)))

**10 REGRESSION – BOKEH IS LIFE!!!**

**Implement the non-parametric Locally Weighted Regression algorithm in order to fit data points. Select appropriate data set for your experiment and draw graphs.**

import numpy as np

from bokeh.layouts import gridplot

from bokeh.plotting import figure, show

def local\_regression(x0, X, Y, tau):

x0 = np.r\_[1, x0]

X = np.c\_[np.ones(len(X)), X]

**xw = X.T \* np.exp(np.sum((X - x0) \*\* 2, axis = 1) / (-2 \* tau \*\* 2))**

**beta = np.linalg.pinv(xw @ X) @ xw @ Y**

return x0 @ beta

n = 1000

**X = np.linspace(-3, 3, num = n)**

print('The **dataset**(10 samples) X is:', X[1: 10])

**Y = np.log(np.abs(X \*\* 2 - 1) + 0.5)**

print('\nThe **fitted curve dataset**(10 samples) Y is:', Y[1: 10])

**X += np.random.normal(scale = 0.1, size = n)**

print('\nThe **normalized dataset**(10 samples) X is:', X[1: 10])

**domain = np.linspace(-3, 3, num = 300)**

print('\nThe domain (10 samples) is:', domain[1: 10])

def plot\_lwr(tau):

prediction = [local\_regression(x0, X, Y, tau) for x0 in domain]

plot = figure(plot\_width = 400, plot\_height = 400)

plot.title.text = 'tau: %g' % tau

plot.scatter(X, Y, alpha = 0.3)

**plot.line(domain, prediction, line\_width = 2, color = 'red')**

return plot

**show(gridplot([[plot\_lwr(10), plot\_lwr(1)], [plot\_lwr(.1), plot\_lwr(.01)]]))**