Assignment Sheet 7

Assignment 7.1

Describe in a few words the goal of Semi-Supervised Learning. What problems are there?

What challenges?

Goal of the Semi-supervised learning is to-

Get better performance of a model than when applied using only labeled/unlabeled dataset.

Problems-

Active learring

Challenges-

Additional assumptions to be made from unlabeled data

Keeping it efficient in lareg amount of unlabeled data

Assignment 7.2

Consider real numbers as our data objects. Given are two data points 1 and 10. Point

1 is labeled as one class (rectangle), point 10 as another class (circle). Use the nearest

neighbor classi\_er for the classi\_cation of a new instance.

1. How would a new data point 7 be labeled?

1 7 10

Since nearest neighbor is 10 hence shall be labeled round

b) Now assume that we have given more unlabeled data. How would point 7 be

labeled after applying self-learning? Show each step of the algorithm.

1 2 3 4 5 6 7 10 11 12 13 14 15 16

Step 1: labeled training data- 10-round, 1-sqaure

Step 2-Learned classifier, NN here

Step3- as per NN, 7 shall be round since 10 is NN

Step 4- now training data is- 10-round, 7-round, 1-sqaure

Step 5- Classify 6 or any other point..

Keept repeating step 2 untill all unlabeled points ar e classified, add most confident classificati to training data

7 shall be round still

But if we take order differently like label 2 first then,

2 shall be square

New data = 1,2,-sqaure, 10-round

Now classify 3

NN of 3 is 2 hence 3= square

New data = 1,2,3-sqaure, 10-round

Now classify, 4, NN of 4 is 3 which is square hence 4 is square

Hence 7 shall be labeled as sqaure

c) Discuss in general which kind of unlabeled data most likely improves classi\_cation

quality and which does not.

Unlabaled data distribution imporves classification quality by shifting the decision boundary i.e. which lies near to decision boundary , some doesn’t not if all data is into same class, i.e. farthest away from the decision boundary.

Assignment 7.3

Consider the following 2D data set and again a nearest neighbor classi\_er:

1 2 3 4 5 6 7

1

2

3

4

5

x

y

a) If point (3,2) is labeled as + and point (7,3.5) is labeled as 􀀀, how would that

a\_ect self-training, if you always would classify the most con\_dent point next?

(Hint: The most con\_dent point in this case is the nearest one.)

In this case it shall label everything as +ve since most confident is NN and from –ve point all pointa are relatively farther than the +ve , depending on the order.

b) What would a better initial labeling be and why? How would it a\_ect the selftraining

process in this case?

Better initial labelling shall result in intial correct predcitions of unlabeled data leading to beter accuracy. Informative i.e. which are close to decision boundary should be given as intital labeling

Hence (2,2) and (3,2), (3,3) seems more reasonable.

Assignment 7.4

What is a generative model? How do they relate to clustering? Give examples for

generative models. What are advantages and disadvantages?

Generatiev models-

A generative models gives the joint probability P(X,Y) , or P(x) if no Y is aviabale.

So, it shall be able to generate new instances, and also preict how likely is a new instacnce.

It has the data distribution.

In realting to clustering, in generative models it assumes that the data given is “generated” from a model with some parameters(mean, variance). In case of clustering, it assumes that the clusters k are generated from a model with k differnet distributaion normally k gissian distributions in GMM and by adjusting the paramerts we increase the probability that the model shall produce the given data.

Hence if for a generated given point mean and variance is know then its known that which distribution it belongs to.

Same way, in given cluatering problem, for given k clusters if the thea is known for a data point it can be know to which cluter it belongs to.

Examples of generative models in semi-supervisedleanirng-

1.Mixture of gussian distribution

Used in text classification, EM algorithm

2. Mixture of multinomial distributions using NB

Used in text categorization

Advantages-

Its well defined probbalistic framework

If the chosen model is correct then works very effeictively

Disadvantage-

Its difficult to verify the correctness of model

Unlabeled data may reduce performance if selected model is wrong, EM can get stuck in local minima