Review for Nov. 26, 2024 2:30p Exam

COSC 6335

On Nov. 21, 2024

1. EM

a) What cluster models does EM use

Each cluster is described by:

a. a mean value

b. a covariance matrix

c. a cluster prior/weight (weights of the k clusters have to add up to one)

[Gaussian Mixture Models — PyPR v0.1rc3 documentation (sourceforge.net)](https://pypr.sourceforge.net/mog.html)

b) How does EM determine if a point i belongs to a cluster j

$$p\left(x\_{i}\right)=\frac{p\left(C\_{j}\right)p(C\_{j})}{\sum\_{l=1}^{k}p\left(C\_{l}\right)p(C\_{l})}$$

2. Fuzzy C-Means (FCM)

a. How is FCM different from K-means?

FCM uses soft cluster memberships expressed in weight wij which can be interpreted as probability of object i belonging to cluster j; that is, objects have to belong to exactly one cluster, as it is the case with k-means. FCM uses weight based computations to determine the centroid.

b. How does FCM update the weights in its iterations

Let us assume we run FCM for K=2 and the centroids are cluster 1=(1,1) and cluster 2=(2,3) and hyper parameter p is 2 and we use Manhattan distance; furthermore point i is: (1,4) in this case;

Wi1= 1/3\*\*2/(1/9+1/4)=0.309

Wi2= 1/2\*\*2)/(1/9+1/4)=0.692

$$w\_{ij}=(1/dist(x\_{i},c\_{j})^{2})^{\frac{1}{p-1}}/\sum\_{q=1}^{k}(1/dist(x\_{i},c\_{q})^{2})^{\frac{1}{p-1}}$$

c) When does Fuzzy C-Means terminate?

When its k×n weight matrix does not change between iterations / changes very very little between interactions

where

n number of objects in the dataset

k number of clusters

**3) Miscellaneous Questions [19]**

a. What are the characteristics of overfitting when learning decision trees? What can be done to deal with overfitting? [3]

1. reduce the size of the tree [1.5]

2. increase the size of the training set by adding new training examples [1.5]

b) A confusion Matrix of a classification model for distinguishing dogs, cats and rabbits is given below:

What is the accuracy of the classification model; what is its precision for class rabbit? What is its recall for class rabbit? It is okay to represent your answers as fractions; e.g. 17/36! [3]



Accuracy=23+29+24/(total number of examples)

Precision rabbit: 24/(24+10+4)=24/38

Precision rabbit: 24/(24+13+7)

**4) Autoencoders**

a) What role do Kullback–Leibler (KL) divergences play in Variational Autoencoders (VAEs)?



i. KL-divergences measure the distance between two distributions e.g. how close are N(0.3,1.3) and N(5,13) to N(0,1); obviously,

dKL(N(0.3,1.3),N(0,1))<< dKL(N(5,13),N(0,1))

ii. KL-divergences are used in VAE loss functions to create a penalty that is proportional to how much the latent vector deviates from an assumed prior (e.g. from N(0,1) or from a covariance matrix which has 1 in the diagonal and 0 everywhere else[[1]](#footnote-1)); this accomplish some regularization of the latent space and also can be used to enforce independence of the different latent variables)

b) How can autoencoders be used for outlier detection?

Steps:

i. Learn an Autoencoder Y for your Dataset D

ii. Feed all example e∈D into Y obtaining ē and add the reconstruction loss d(e, ē)—e.g. the Euclidian distance from e to ē—as an outlier score to example e.

Remark: the larger d(e,ē)is the more likely x is an outlier

c) If I decrease the dimensionality of the latent vector h what will be the consequences?

The reconstruction loss will increase.

**5) Association Rule Mining**

a) Assume we have the following Transaction Database

T1: {A,B,C,D}

T2: {A,C,D,E}

T3: {C,D,E,F}

T4: {B,C,D,E}

T5: {A,D,E}

What is the support and confidence the following association rule:

IF (C and D) THEN E? [3]

Support = 3/5 [1.5]

Confidence=3/4 [1.5]

b) Assume the APRIORI algorithm identified the following five 4-item sets that satisfy a user given support threshold: **abcd, acde, acdf, acdg adfg;** what initial candidate 5-itemsets are created by the APRIORI algorithm; which of those survive subset pruning? [4]

acdef, acdeg, acdfg [3] One error: at most one point!

None survives pruning [1]

c) Why are association rule mining systems interested in finding rules with high support? [2]

Rules with high support are more likely to predict the occurrence of an item based on the occurrences of other items in the transaction accurately; it is hard to learn accurate rules from just a few examples.

d) Assume an association rule if smoke then cancer has a confidence of 86% and a high lift of 5.4. What does this tell you about the relationship of smoking and cancer? [2]

Con = 86% 🡪 86% people who smoke tend to get cancer; that is P(Cancer|Smoke)=0.86

Lift = 5.4 🡪 Smoking increases the probability of getting cancer by a factor of 5.4; that is, P(Cancer|Smoke)/P(Cancer)=5.4

**6) Neural Networks [8]**

a) Take a look at the sub neural network consisting of nodes A, B, C, and D in the figure below; give a formula that computes the associated error ΔA for a node A. Assume the used activation function is g and its derivative is denoted by g’, and the activation of a node X is denoted by aX and the linear input of a node X is denoted by zX. First provide a general formula; then, replace general variables in the formula by their actual known values, for those which are known! [4]

 wA,B=0.5

 ΔB=0.2

 A B

 wC**,A**=1 wD,A=0.2

 C D

Formula: ΔA=g’(za)\*wAB\*ΔB=g’(zA)\*0.2\*0.4=g’(0.24)\*0.5\*0.2

If not correct at most 1 point partial credit

b) The step width in NN gradient descent search varies, depending on various factors including depending on the derivative of the activation function and the activation of the input node. Explain why each of those two variations is desirable? [4]

If the gradient of the activation function is low small steps are used to not overshoot the local minimum which is characterized by a gradient of zero, [2]

If the activation of the input is low the link has very little influence on the activation of the output node; consequently, the weight of the connection is only changed a little or not at all. [2]

7. CLIQUE

Assume we apply the CLIQUE algorithm to a numerical dataset with attributes A, B, C, D and E. What is the main difference between CLIQUE and traditional Clustering algorithms such as K-means with respect to the clusters CLIQUE finds? How does CLIQUE take advantage of the APRIORI principle? How does CLIQUE form clusters? [6]

*Find clusters in the subspace rather in the complete space A-B-C-D-E-F [2]*

*K+1 dimensional grid-cell candidates are computed from K-dimensional grid-cell which are dense (the number of points they contain is above the density threshold. [2]*

*Clusters in subspaces are formed by a growing algorithm which starts with a seed grid-cell and adds neighboring grid-cells [2]*

Other answers might deserve full or partial credit!

8) K-means

a) Assume the following dataset is given: (1,1), (2,2) (4,4), (5,5), (4,6), (6,4) . K-Means is used with k=2 to cluster the dataset. Moreover, Manhattan distance is used as the distance function (formula below) to compute distances between centroids and objects in the dataset. Moreover, K-Means’s initial clusters C1 and C2 as follows:

C1: {(1,1), (3,3), (4,4), (6,6)}

C2: {(6,4), (4,6)}

Now K-means is run for a single iteration; what are the new clusters you obtain[[2]](#footnote-2) [4]

**d((x1,x2),(x1’,x2’))= |x1-x1’| + |x2-x2’| Manhattan Distance**

centroid C1= (3.5,3.5}

centroid C2= {5,5}

New Clusters

C1={(1,1), (3,3), (4,4)}

C2={(6,6},(4,6), (6,4)}

1. No correlation between variables. [↑](#footnote-ref-1)
2. If there are any ties, break them whatever way you want! [↑](#footnote-ref-2)