Review for Final Exam

COSC 6335

Dec. 2, 2022

**1) Decision Trees/Classification**

Compute the GINI-gain[[1]](#footnote-1) for the following decision tree split (just giving the formula is fine!)[3]:

(12,4,6) (3,3,0)

(9,1,0)

 (0,0,6)

 G(6/11,2/11,3/11) – (6/22\*G(0.5,0.5,0) + 10/22\* G(0.9,0,1,0) + 0)

**G(6/11,2/11.3/11)= 1-(6/11\*\*2 + 2/11\*\*2+3/11\*\*2)= 1- ((36+4+9)/121)= 72/121=0.595**

**The worst Gini for 3 classes is: G(1/3.1/3.1/3)=0.667**

1. Assume there are 3 classes and 50% of the examples belong to class1, and 25% of the examples belong to class2 and class3, respectively. Compute the entropy of this class distribution, giving the exact number not only the formula! [2]

H(1/2,1/4,1/4)= ½\*log2(2)+ 2\*1/4log2(4)=3/2=1.5

1. The decision tree learning algorithms is a greedy algorithm—what does this mean? [2]

**Seeks the shortest path from the current state to a goal state/makes local decisions[1]; does not backtrack[1]; frequently, does not find the optimal solution[1].**

1. Assume you learn a decision tree for a dataset that only contains numerical attributes (except the class attribute). What can be said about the decision boundaries that decision trees use to separate the classes? [1]

**Axis parallel lines/hyperplanes (of the form att≥x where att is one attribute of the dataset and value is a floating point number) where each axis corresponds to the line were the value of att is x; other answer: axis-parallel rectangeles.**

2. 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

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

3. 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

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

**4) 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)

**5. Non-Parametric Density Estimation [10]**

a. What are the key ideas of non-parametric density estimation? Verbally describe how do they compute the density in a query point—we expect an answer in words; not a formula! [3]

Density of a query points is computed by adding up the influences of the points in the dataset on the query point [2]; influences decrease of dataset objects with distance to the query point [1].

b. We use Gaussian Kernel Density estimation for the following dataset

O={(0,0), (2,2), (0,1)}.

We additionally, assume σ=2 and moreover assume Manhattan distance[[2]](#footnote-2) is used as the distance function and assume that (1,2) is the query point whose density needs to be computed. Compute fGauss ((1,2))! [7]

fGauss((1,2))= e-9/8 + e-4/8 + e-1/8

one error: up to 4.5 points; 2 errors at most 1 point. Might just subtract a single point if the “correct” sum of influences is multiplied with the “wrong” constant.

**6) Expect one more essay-style question in the final exam**

**Important: this is an essay: write complete sentences!**

**e.g What skills are important to be hired as a Data Scientist / A job which requires data mining?**

(see slides that discuss this topic)

* Should know R and/or Phyton
* Should have sound software development skills
* Should have some sound knowledge of Statistics
* Should have sound knowledge of the different data analysis tasks; e.g. clustering, classification, similarity assessment
* Should be knowledgeable in data visualization
* Data scientists are involved with gathering data, massaging it into a tractable form, making it tell its story, and presenting that story to others.”
* The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that’s going to be a hugely important skill in the next decades."
* But what’s even harder is finding people who have those skills *and* are good at communicating the story behind the data.”

Remark: Other final exams topics have been discussed in the GHC presentations given in the last weeks.

1. (GINI before the split) minus (GINI after the split) [↑](#footnote-ref-1)
2. [↑](#footnote-ref-2)