Quiz 1 Solutions

Q1 SQL

Q1.1:

Option 1: We discarded this option, as technically it is possible to represent a player who plays for multiple teams, but it requires redundantly repeating the player name, so is not a good database design.

Option 2: True

Option 3: False; each game includes the teams that played in it

Option 4: False; this could be added to each player, since the player only plays for one team

Q1.2:

You would need to add a mapping table from players to games, e.g.:

GamesPlayers(gp_p_id, gp_g_id)

Q1.3

SELECT t_id FROM mw_players WHERE t_id NOT IN (SELECT t_id from mv_players)

Q2 Pandas

A B C
1 2 3
4 5 6
7 8 9

Q3 Regular Expressions

Regex: `^[1-9]{2},[a-z]{3},(red|blue)$`

Q4 Cosine Similarity

Words: the slug bug eats runs on rug in
v1 = [3, 1, 1, 1, 0, 0, 1, 1]
v2 = [2, 0, 1, 0, 1, 1, 1, 0]
Similarity = v1.v2 / |v1||v2| = 0.7559289460184544

Q5 Clustering

Q5.1 K-Means Clustering 1 Step

We accepted two different solutions, based on different interpretations of “one iteration”. The solution we were originally looking for simply assigns the points to the centroids given, leading to :
C = [O1]
C = [B4]
C = [all other points]

We also accepted the clusters resulting after assigning points and re-calculating centroids (i.e. the clusters after *two* iterations):
C = [O1]
C = [B1, B2, B3, B4]
C = [A1, A2, A3, A4, A5, B5]

Q5.2 K-Means Clustering Final Clustering

C = [O1]
C = [B1, B2, B3, B4, B5]
C = [A1, A2, A3, A4, A5]

Q5.3 K-Means 2 Centroids

It is not guaranteed that the points A1-A5 will be grouped together and separately from the points B1-B5 - if one of the centroids is randomly assigned to O1, all other points will be clustered together.

Q5.4 DB-Scan

DB-scan would mark O1 as an outlier and cluster all other points together in one cluster, since they are all density-reachable from each other for the given parameters.

Q5.5 Other Algorithms

We need a density-based clustering approach to identify the right clusters; Gaussian mixture is therefore the appropriate choice.

Q6 Data Cleaning

Q6.1 Detecting Errors

We accepted various possible answers, as long as more than one reasonable technique was presented and justified. Example techniques include: visualization, analyzing dataset statistics, looking at outliers, counting null/missing values.

Q6.2

Listwise deletion, since it will only keep the two records with the highest values in the revenue column, leading to an average revenue of $2,500M. Mean substitution will impute $1,174.4M, leading to an average revenue of $1,174.4M. Stochastic imputation will at most impute $2,800M, leading to an average revenue of $1,445.3M.

Q7 Evaluation

Q7.1

Accuracy 0.84

With respect to accuracy, the classifier is not good. Just picking the majority class always would lead to an accuracy of 0.94%.

Q7.2

Arithmetic mean of all precision scores = ((10/80)+(10/70)+(820/850))/3 = 0.41
Arithmetic mean of all recall scores = ((10/30)+(10/30)+(820/(820+55+65)))/3 = 0.51

F1 macro: 2/(1/R + 1/P) = 0.46

F1 macro gives each class the same weight and thus is a better metric if imbalance of labels is an issue.

Q7.3

The training error is very good (F1 macro score is 0.9) and much lower compared to the test error (F1 macro is 0.46). Hence, it can be assumed that the classifier suffers from high variance.

Q7.4

Given that we have high variance the following measures help
“Add more training data”
“Increase regularization”

“Remove features”

Q8 Decision Trees

Q8.1

- Remove “Product bought on”, as it leaks information. The feature is only known when the label is known.
- Remove or change “Last contact”. If it is used in its current form, last contact will not generalize to new data. For example, transforming it to a relative time (e.g., time from first contact) would work.
- Transforming the other values (e.g., embeddings for Sales Contact, etc.) might be necessary depending on the model type. For decision trees no modifications are necessary.

Q8.2

Correct answer: Sales Contact

Just from looking at the data, it can be easily seen that only splitting on “Sales Contact” or “Contacted by Phone” would significantly reduce the entropy (make the group more homogeneous afterwards). However, splitting on “Sales Contact” will have a higher information gain as it is also able to better distinguish the three status values (“Subscribed to Plan A”, “Subscribed to Plan B”, “Lost”).

More formally: if we would split on “Contacted by Phone”, the entropy afterwards would be 0.18 vs splitting on “Contacted by Phone” would be 0.26. So the information gain on “Contacted by Phone” is higher.