CS 422 – Data Mining

Spring 2021 – All Sections

Midterm Exam

Part I – Short Answer (Show Points/Results) – 5 points each, 40 points total

1. Given the following feature vector $x = [4.4, 5.1, -3.7, 2.1, -1.9]$, what would a categorical representation of this feature vector be if we assumed 3 discrete categories with values $x \leq -2.5$ as A, $-2.5 < x < 2.5$ as B, and $x \geq 2.5$ as C?

2. Given a binary classification problem with classes $\{C_1, C_2\}$, draw a Confusion Matrix showing result counts $(f_{11}, f_{10}, f_{1+}, f_{0+}, \ldots)$ in terms of Predicted and Actual class. Provide calculations for Accuracy and Error Rate, highlighting False Positives and False Negatives (FP, FN) as functions of these result counts.

3. For frequent itemsets $\{\{A, B\}, \{C\}\}$, show the difference between the Confidence c vs the Interest Factor (Lift) for the Association Rule $\{A, B\} \implies \{C\}$. What value does Lift take into account that Confidence does not?

4. Given a dataset with n observations, what is the size of the training set if we choose to hold out $\frac{n}{k}$ records as a test set? If we allow for $k \rightarrow n$, what does the corresponding training set size approach?

5. With a data set containing $d = 15$ features and $N = 12{,}000$ observations, what is the dimensionality of the covariance matrix of the predictors? If we were to represent the predictors with a multivariate normal (Gaussian) distribution, how many distribution parameters would need to be estimated from the feature data?

6. Given the following point observations: $x_1 = [3,4]$ and $x_2 = [5,12]$, what would the length of each vector in terms of the Manhattan and Euclidean norms (L_1, L_2) be defined as? Would the distance between the two points be larger under the L_1 or L_2 norm?

7. Draw the 2-way contingency table for a binary association rule $\{A\} \implies \{B\}$, containing presence/absence counts $(f_{11}, f_{10}, f_{1+}, f_{0+}, \ldots)$. Interest Factor (Lift) can be interpreted as a conditional probability $\dfrac{P(A,B)}{P(A)P(B)}$, show this probability in terms of these counts.

8. For a binary association rule $\{a\} \implies \{b\}$, show that the ϕ coefficient for the rule's correlation measure is not invariant under null addition (unchanged with added unrelated data) in terms of changes to the relevant counts $(f_{11}, f_{10}, f_{1+}, f_{0+}, \ldots)$.

Part II – Long Answer (Show Reasoning/Calculations) – 10 points each, 40 points total

1. Show the cosine similarity of the two vectors $x = [3,4,5]$ and $y = [5,12,13]$. Results can be kept in formula form in terms of the component values of x and y (calculation of final value not required).

2. Given a classifier with True Positives/Negatives (TP, TN) and False Positives/Negatives (FP, FN), what is the highest Recall r value that a model can achieve? Define the Recall measure via (TP, TN, FP, FN). How can one design a simple model which achieves the maximum value for Recall?

3. Given the following transactions: $\{a, b, c\}, \{a, c\}, \{b, c\}, \{a\}, \{b\}, \{c\}$, with $minsup = 60\%$, what itemsets would be frequent? What would be the support s of the association rule: $\{a\} \implies \{c\}$ be? What would the confidence c of this rule be? Given the $minsup$ value, would this be a valid rule that is extracted via the Apriori Algorithm?

4. Given a data matrix D with $d = 5$ features/columns with a total variance of 100, an analyst performs a PCA via eigenvalue decomposition, with the resulting eigenvalues as $[35,25,20,15,5]$. If the analyst wishes to reduce dimensionality with 80% of variance explained, how many dimensions would the analyst be able to reduce their selection to? What would be the standard deviations σ_i of the data for each these selected dimensions?

Part III – Essay Question (Show Argument/Proof) – 20 points each, 20 points total

1. Given a decision tree node containing 10 records, half of which belong to Class C_A and the other half which belong to Class C_B, show the impurity I of the node under the Entropy, Gini, and Misclassification Error measures. What would be the value of these measures be for the child nodes, assuming an optimal split is performed? (**Hint:** Assume $0 \log_2 0 = 0$).

Lucky 7 – Bonus Questions (Industry News, AI/ML Topics) – 1 point each, 7 points total

1. What model recently released by DeepMind allows for accurate prediction of 3-dimensional shape of a protein molecule given input amino acids?

2. Which firm recently fired its head of AI ethics, shortly after the controversial departure of one of its senior researchers?

3. What family of algorithms were recently developed which are able to solve classic treasure hunting video games such as Pitfall on Atari?

4. What disease was IBM able to predict the onset of based on changes in writing/ language via the use of machine learning models?

5. What category of modified videos did a consortium led by Facebook/Microsoft/ Cornell/MIT recently introduce a detection challenge for?

6. Which firm recently released a new image recognition algorithm that was trained on over 1 billion images, but did not require manual labels?

7. What quantum computing goal was recently achieved by Google which was revealed to the public via NASA?