Birla Institute of Technology and Science-Pilani, Hyderabad Campus

Second Semester, 2021-2022

Mid Semester Examination

CS F415 – Data Mining

Total Marks: 75 Time: 1.5 hours

**Answer All Questions**

**Short Answer Types (30 marks)**

1. **(a)**What are the distinguishing factors of heterogeneous data from other types of data?

**(b)**What are the different types of heterogeneity?

**(c)**Explain each types in a sentence or two. [2 + 4 + 4]

1. **(a)**Name five data preprocessing techniques.

**(b)**Give an example to each of these techniques. [5+5]

1. **(a)**Define curse of dimensionality.

**(b)**Name two techniques for dimensionality reduction with brief explanations of each of these techniques. [2+3]

1. Name four different approaches for feature subset selection techniques. [5]

**Long Answer Types (45 marks)**

1. **(a)**Based on the maximum variance formulation, what is the goal in PCA?

**(b)**Write the steps for normalizing the given dataset.

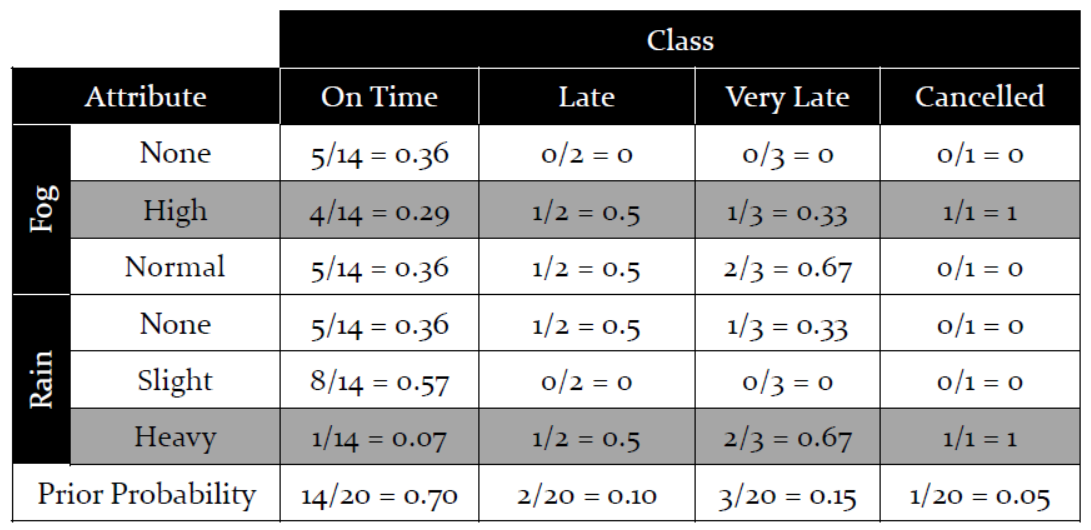
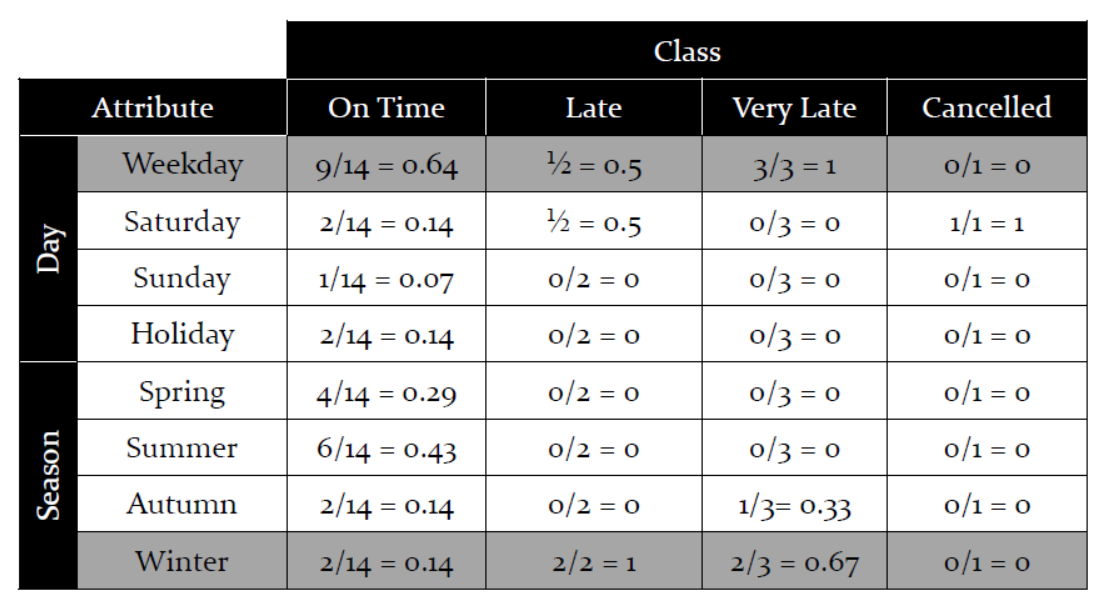
**(c)**Write the data covariance matrix where we find top-k eigenvectors u1, u2, …, uk as principal components.

**(d)**If x(i) is the original data point and y(i) is the projection of x(i) along the k-principal components, write the relationship between x(i) and y(i). [3+6+3+3]

1. **(a)**When can we say that two events A and B are mutually independent? Define in terms of probability.

**(b)**When can we say that two events A and B are conditionally independent on another event C? Define in terms of probability.

**(c)** Given the two tables of class conditional probabilities, decide the class using naïve bayes classifier When **Day: Week Day**, **Season: Winter**, **Fog: None**, **Rain: Heavy** [3+4+8]



1. (a) Write the optimization problem formulation that needs to be solved in SVM.

(b) What are slack variables in soft margin classification?

(c) Write the optimization problem formulation for soft margin classification.

(d) For the Kernel K(x,y) = (xTy + c)2, derive the feature map (x).

(e) show that the kernel K is symmetric. [3+1+3+6+2]