1. Continuous, quantitative, interval
Discrete,qualitative,ordinal
Continuous, quantitative, ratio(in light year)

2 True. Outliers will distort the mean and variance.
False. Undefined
True.Aggregate can save more information than sampling. There are several motivations for aggregation. First, the smaller data sets resulting from data reduction require less memory and processing time, and hence, aggregation may permit the use of more expensive data mining algorithms. Second, aggregation can act as a change of scope or scale by providing a high-level view of the data instead of a low-level view.
False.$(1, 1, 1, 1)(0, 0, 0, 0)$ Hamming:4 Euclidean:2

3

a

Correlation
The document data is sparse and asymmetric. and the similarity depends on the number of characteristics the document data shared. Only presence should be counted to the similarity measure.
b

1.The scales are different and for euclidean distance, the largest-scaled feature would dominate the others.
2.standardizing each attribute.
3.Mahalanobis distance. the Mahalanobis distance is a generalization of Euclidean distance and it is useful when attributes are correlated, have different ranges of values.It takes into account the covariance between attributes when computing the distance between data instances.
4

ANN
strength:can handle redundant attributes (annual income,income tax filed.) redundant attributes receive similar weights and do not degrade the quality of the classifier

weakness:cannot handle different scales.

SVM

strength: It can also handle redundant attributes by learning similar weights for the duplicate attributes. Furthermore, the ability of SVM to regularize its learning makes it more robust to the presence of a large number of irrelevant and redundant attributes than other classifiers, even in high-dimensional settings.

weakness: naive bayes assumption on conditional independence will fail in the presence of interacting attributes(age,income)

ripper

strength: ripper can handle such a skewed class distribution more effectively than decision tree due to its ability to focus on the rare class.

weakness: decesion tree cannot handle the imbalance class.

5

Model 2

lower validation error rate.Since the validation error is representative of the generalization performance of a model on unseen instances, Model 2 is expected to have better performance.Also note that Model 2 shows similar error for both datasets A and B, highlighting the fact that it is not overfitted to the training set, and thus its error on the validation data is likely to be closer the true generalization error (i.e., one can expect similar error on data set C). In contrast, Model 1 performs much worse on the validation set, indicating that it is overfitted to the data used for its training, and thus is likely to perform much worse on data sets that are different than the one used for training.

no.

No, even though the training error is the lowest and the validation error is also the lowest, by Occam's Razor , given two models of similar generalization errors, one should prefer the simpler model over the more complex model. The validation error for model 3 is much higher than its training error. So, it is less likely to generalize well over unseen instances

6

best:kNN worst:Linear SVM KNN is the best due to the proximity of the examples of the same class to each other. Linear SVM cannot separate the Linear inseparable data.

best:NB worst:Knn Knn cannot handle noisy attribute.

7.

False.Optimistic error is. Pessimistic generalization error adds the complexity of the model.

2

False. We should seek a model that minimizes the overall cost function.
8.

T2 would be worse.The presence of noise can lead to the learning of a highly complex decision tree that has improved performance over noisy instances on training data but poor generalization performance over unseen test instances, which is over-fitting.

9.

best decision tree worst: naive bayes explanation: For Decision tree,the presence of redundant attributes will have no impact on the classification performance, as only one of them will be chosen for splitting at an internal node in the tree. ANN can handle the redundant attributes as well, however, if the number of irrelevant or redundant attributes is large, the learning of the ANN model may suffer from overfitting, leading to poor generalization performance. naive bayes would impacted by the redundant attribtues since all redundant attributes will have a contribution to the joint conditional probabilities.

best:naive bayes worst:ann explanation: Naïve Bayes can handle missing values both during training and testing. Decision trees can handle missing values during training and testing to some extent, but it requires correction

mechanisms like the probabilistic split method.In addition, ANN cannot handle instances with missing values in the training or testing phase.