1. What exactly is a feature? Give an example to illustrate your point.

Answer: A feature is a measurable property of the object you’re trying to analyze.

Example: In character recognition, features may include histograms counting the number of black pixels along horizontal and vertical directions, number of internal holes, stroke detection and many others.

2. What are the various circumstances in which feature construction is required?

Answer: Feature construction involves transforming a given set of input features to generate a new set of more powerful features which are then used for prediction. This may be done either to compress the dataset by reducing the number of features or to improve the prediction performance.

Feature construction methods may be applied to pursue two distinct goals: reducing data dimensionality and improving prediction performance.

The task of constructing appropriate features is often highly application specific and labour intensive. Thus, building automated feature construction methods that require minimal user effort is challenging. We want methods that:

1. Generate a set of features that help improve prediction accuracy.
2. Are computationally efficient.
3. Are generalizable to different classifiers.
4. Allow for easy addition of domain knowledge.

3. Describe how nominal variables are encoded.

Answer: A column with nominal data has values that cannot be ordered in any meaningful way. Nominal data is most often one-hot (aka dummy) encoded, but there are many options that might perform better for machine learning. Use Category Encoders to improve model performance when you have nominal or ordinal data that may provide value.

For nominal columns try OneHot, Hashing, LeaveOneOut, and Target encoding. Avoid OneHot for high cardinality columns and decision tree-based algorithms.

Classic Encoders:

OneHot — one column for each value to compare vs. all other values. Nominal, ordinal.

Hashing — Like OneHot but fewer dimensions, some info loss due to collisions. Nominal, ordinal.

4. Describe how numeric features are converted to categorical features.

Answer: from sklearn.preprocessing import LabelEncoder: you can convert numeric into categoric.

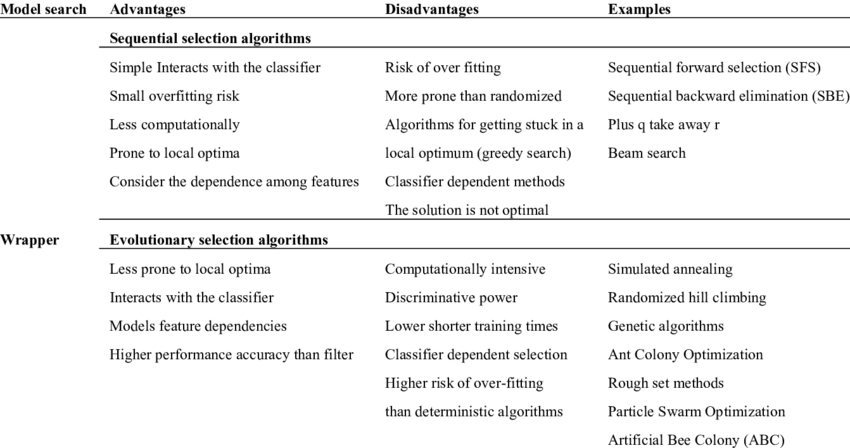
5. Describe the feature selection wrapper approach. State the advantages and disadvantages of this approach?

Answer: In wrapper methods, the feature selection process is based on a specific machine learning algorithm that we are trying to fit on a given dataset.

It follows a greedy search approach by evaluating all the possible combinations of features against the evaluation criterion. The evaluation criterion is simply the performance measure which depends on the type of problem, for e.g. For regression evaluation criterion can be p-values, R-squared, Adjusted R-squared, similarly for classification the evaluation criterion can be accuracy, precision, recall, f1-score, etc. Finally, it selects the combination of features that gives the optimal results for the specified machine learning algorithm.

Most commonly used techniques under wrapper methods are:

* Forward selection
* Backward elimination
* Bi-directional elimination (Stepwise Selection)



6. When is a feature considered irrelevant? What can be said to quantify it?

Answer: Weak relevance implies that the feature can sometimes contribute to prediction accuracy. Features are relevant if they are either strongly or weakly relevant and are irrelevant otherwise. Irrelevant features can never contribute to prediction accuracy. Feature considered irrelevant if they

do not add any information

7. When is a function considered redundant? What criteria are used to identify features that could be redundant?

Answer: Redundant features add no relevant information to your other features, because they are correlated or because they can be obtained by [linear] combination of other features. Having them on your set will not add anything, but it won't hurt either, information-wise.

For example, if two features {X1, X2} are highly correlated, then the two features become redundant features since they have same information in terms of correlation measure. In other words, the correlation measure provides statistical association between any given a pair of features.

8. What are the various distance measurements used to determine feature similarity?

Answer: Four of the most commonly used distance measures in machine learning are as follows:

* Hamming Distance
* Euclidean Distance
* Manhattan Distance
* Minkowski Distance

9. State difference between Euclidean and Manhattan distances?

Answer: Euclidean distance is the shortest path between source and destination which is a straight line but Manhattan distance is sum of all the real distances between source(s) and destination(d) and each distance are always the straight lines.

10. Distinguish between feature transformation and feature selection.

Answer:

Feature transformation is simply a function that transforms features from one representation to another. Reason for transformation is:

* Data types are not suitable to be fed into a machine learning algorithm, e.g. text, categories
* Feature values may cause problems during the learning process, e.g. data represented in different scales.
* To reduce the number of features to plot and visualize data, speed up training or improve the accuracy of a specific model.

Three main transformation techniques:

* Handling categorical variables
* Feature scaling
* Principal Component Analysis

Feature selection is the process of selecting specific features, from a features pool. This helps in simplification, regularization and shortening training time.

11. Make brief notes on any two of the following:

1.SVD (Standard Variable Diameter)

2. Collection of features using a hybrid approach

3. The width of the silhouette

4. Receiver operating characteristic curve

Answer:

1. The singular value decomposition (SVD) is among the most important matrix factorizations of the computational era, providing a foundation for nearly all of the data methods discussed in this book. In particular, the SVD provides a numerically stable matrix decomposition that can be used for a variety of purposes. We will use the SVD to obtain low-rank approximations to matrices and to perform pseudo-inverses of non-square matrices to find the least-squares and minimum norm solutions of a matrix system of equations Ax = b. Another important use of the SVD is in the principal components analysis (PCA), whereby a large, high-dimensional data set is decomposed into its most statistically descriptive factors.
2. Hybrid approach for the feature selection can be done by using artificial bee colony (ABC) and particular swarm optimization. The proposed technique is simulated using the WEKA and results show the better performance of the proposed technique.
3. One of the most commonly applied methods for assessing cluster validity is silhouette width, which encompasses two clustering criteria: separation (i.e., average distance to the closest other cluster) and compactness (i.e., average within-cluster distance). The Average Silhouette Width (ASW) is a popular cluster validation index to estimate the number of clusters.
4. A receiver operating characteristic curve, or ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The true-positive rate is also known as sensitivity, recall or probability of detection in machine learning.