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

Ans: In machine learning and pattern recognition, a feature is an individual measurable property or characteristic of a phenomenon. The concept of "feature" is related to that of explanatory variable used in statistical techniques such as linear regression. In pattern recognition and machine learning, a feature vector is an n-dimensional vector of numerical features that represent some object.

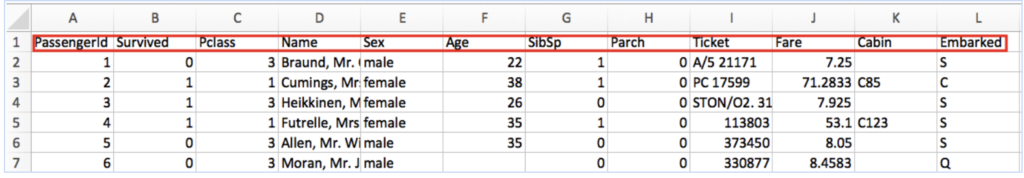
When representing images, the feature values might correspond to the pixels of an image, while when representing texts the features might be the frequencies of occurrence of textual terms.

Feature vectors are often combined with weights using a dot product in order to construct a linear predictor function that is used to determine a score for making a prediction.

Higher-level features can be obtained from already available features and added to the feature vector; for example, for the study of diseases the feature 'Age' is useful and is defined as Age = 'Year of death' minus 'Year of birth' . This process is referred to as feature construction.

The vector space associated with these vectors is often called the feature space. In order to reduce the dimensionality of the feature space, a number of dimensionality reduction techniques can be employed.

A feature is **a measurable property of the object you're trying to analyze**. In datasets, features appear as columns:



The image above contains a snippet of data from [a public dataset](https://github.com/awesomedata/awesome-public-datasets/blob/master/Datasets/titanic.csv.zip) with information about passengers on the ill-fated Titanic maiden voyage. Each feature, or column, represents a measurable piece of data that can be used for analysis: Name, Age, Sex, Fare, and so on. Features are also sometimes referred to as “variables” or “attributes.” Depending on what you’re trying to analyse, the features you include in your dataset can vary widely.

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

Ans: Higher-level features can be obtained from already available features and added to the feature vector; for example, for the study of diseases the feature 'Age' is useful and is defined as Age = 'Year of death' minus 'Year of birth' . This process is referred to as feature construction.

Regardless of the data or architecture, a terrible feature will have a direct impact on your model.

ex: Below are the prices of properties in x city. It shows the area of the house and total price.

|  |  |
| --- | --- |
| Sq ft | AMOUNT |
| 2400 | 5 Million |
| 3200 | 8 Million |
| 2100 | 3 Million |

Now data might have some errors or might be incorrect. To begin , add a new column to display the cost per square foot.

|  |  |  |
| --- | --- | --- |
| Sq ft | AMOUNT | Cost per sq ft |
| 2400 | 5 Million | 2083 |
| 3200 | 8 Million | 5500 |
| 2100 | 3 Million | 1428 |

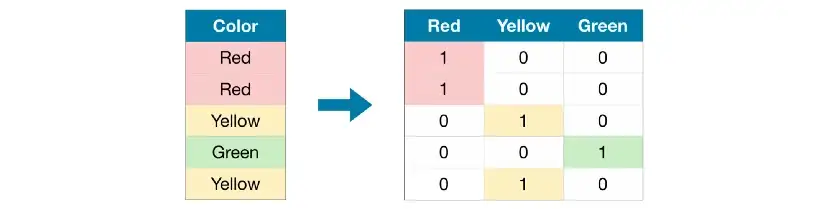
This new feature will help us understand a lot about our data. So, we have a new column which shows cost per square ft.

**3. Describe how nominal variables are encoded.**

Ans: When we have a feature where variables are just names and there is no order or rank to this variable's feature. All the variables in the respective feature are equal. We can't give them any orders or ranks. Those features are called Nominal features.

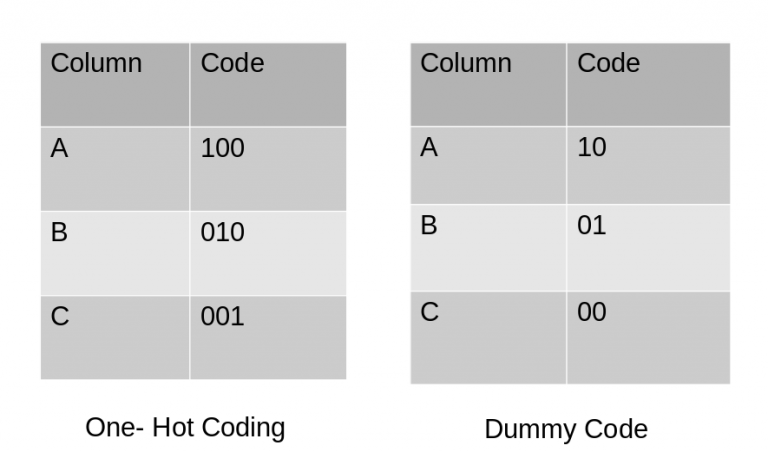
While encoding Nominal data, we must consider the presence or absence of a feature. In such a case, no notion of order is present. For example, the city a person lives in. For the data, it is important to retain where a person lives. Here, we do not have any order or sequence. It is equal if a person lives in Delhi or Bangalore.

**Using one hot encoding :**

ex. 

**Dummy encoding** : Uses dummy (binary) variables. instead of creating a number of dummy variables that is equal to the number of categories (k) in the variable, dummy encoding used k-1 dummy variables

To understand this better let’s see the image below. Here we are coding the same data using both one-hot encoding and dummy encoding techniques. While one-hot uses 3 variables to represent the data whereas dummy encoding uses 2 variables to code 3 categories.



**Effect Encoding**:

This encoding technique is also known as Deviation Encoding or Sum Encoding. Effect encoding is almost similar to dummy encoding, with a little difference. In dummy coding, we use 0 and 1 to represent the data but in effect encoding, we use three values i.e. 1,0, and -1.

The row containing only 0s in dummy encoding is encoded as -1 in effect encoding.

**Hash Encoder:** To understand Hash encoding it is necessary to know about hashing. Hashing is the transformation of arbitrary size input in the form of a fixed-size value. We use hashing algorithms to perform hashing operations i.e to generate the hash value of an input. Further, hashing is a one-way process, in other words, one can not generate original input from the hash representation.

Just like one-hot encoding, the Hash encoder represents categorical features using the new dimensions. Here, the user can fix the number of dimensions after transformation

Since Hashing transforms the data in lesser dimensions, it may lead to loss of information. Another issue faced by hashing encoders is the collision. Since here, a large number of features are depicted into lesser dimensions, hence multiple values can be represented by the same hash value, this is known as a collision.

**Binary Encoding:** Binary encoding is a combination of Hash encoding and one-hot encoding. In this encoding scheme, the categorical feature is first converted into numerical using an ordinal encoder. Then the numbers are transformed in the binary number. After that binary value is split into different columns.

Binary encoding works really well when there are a high number of categories.Binary encoding is a memory-efficient encoding scheme as it uses fewer features than one-hot encoding. Further, It reduces the curse of dimensionality for data with high cardinality.

**Base N Encoding:** Before diving into BaseN encoding let’s first try to understand what is Base here?

In the numeral system, the Base or the radix is the number of digits or a combination of digits and letters used to represent the numbers. The most common base we use in our life is 10 or decimal system as here we use 10 unique digits i.e 0 to 9 to represent all the numbers. Another widely used system is binary i.e. the base is 2. It uses 0 and 1 i.e 2 digits to express all the numbers.

For Binary encoding, the Base is 2 which means it converts the numerical values of a category into its respective Binary form. If you want to change the Base of encoding scheme you may use Base N encoder. In the case when categories are more and binary encoding is not able to handle the dimensionality then we can use a larger base such as 4 or 8.

Hence BaseN encoding technique further reduces the number of features required to efficiently represent the data and improving memory usage. The default Base for Base N is 2 which is equivalent to Binary Encoding.

**Target Encoding:**

Target encoding is a Baysian encoding technique.

Bayesian encoders use information from dependent/target variables to encode the categorical data.

In target encoding, we calculate the mean of the target variable for each category and replace the category variable with the mean value. In the case of the categorical target variables, the posterior probability of the target replaces each category.

We perform Target encoding for train data only and code the test data using results obtained from the training dataset. Although, a very efficient coding system, it has the following issues responsible for deteriorating the model performance-

1. It can lead to target leakage or overfitting. To address overfitting we can use different techniques.
   1. In the leave one out encoding, the current target value is reduced from the overall mean of the target to avoid leakage.
   2. In another method, we may introduce some Gaussian noise in the target statistics. The value of this noise is hyperparameter to the model.
2. The second issue we may face is the improper distribution of categories in train and test data. In such a case, the categories may assume extreme values. Therefore the target means for the category are mixed with the marginal mean of the target.

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

Ans: Three ways to bin numeric features:

1. Equal widths: Takes the range of numerical variables and divides it into equal-sized intervals. The group is identical, but the count of observations in each bin can vary widely.
2. Equal frequency bins: create roughly equal counts in each bin.

Then these bins can be named as categories.

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

Ans: Evaluates the feature set on a specific machine learning algorithm to find optimal features.

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 upon 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.

Techniques:

1. Forward selection: start with a null model and then start fitting the model with individual features one at a time and select the feature with the minimum p-value. Now fit a model with two features by trying combinations of the earliest selected feature with all other remaining features.

steps for the forward selection technique are as follows :

* 1. Choose a significance level (e.g. SL = 0.05 with a 95% confidence).
  2. Fit all possible simple regression models by considering one feature at a time. Total ’n’ models are possible. Select the feature with the lowest p-value.
  3. Fit all possible models with one extra feature added to the previously selected feature(s).
  4. Again, select the feature with a minimum p-value. if p\_value < significance level then go to Step 3, otherwise terminate the process.

1. Backward elimination.: start with the full model(including the independent variables) and then remove the insignificant feature with the highest p-value(>significance level).repeat the process again and again until the final set of significant features.

steps involved in backward elimination are as follows:

* 1. Choose a significance level (e.g. SL = 0.05 with a 95% confidence).
  2. Fit a full model including all the features.
  3. Consider the feature with the highest p-value. If the p-value > significance level then go to Step 4, otherwise terminate the process.
  4. Remove the feature which is under consideration.
  5. Fit a model without this feature. Repeat the entire process from Step 3.

1. Bi-directional elimination (stepwise selection)

It is like forward selection but the difference is while adding a new feature it also checks the significance of already added features and if it finds any of the already selected features insignificant then it simply removes that particular feature through backward elimination.

Hence, It is a combination of forward selection and backward elimination.

In short, the steps involved in bi-directional elimination are as follows:

* 1. Choose a significance level to enter and exit the model (e.g. SL\_in = 0.05 and SL\_out = 0.05 with 95% confidence).
  2. Perform the next step of forward selection (newly added feature must have p-value < SL\_in to enter).
  3. Perform all steps of backward elimination (any previously added feature with p-value>SL\_out is ready to exit the model).
  4. Repeat steps 2 and 3 until we get a final optimal set of features.

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

Ans: Function is considered redundant when is duplicated the results or do not improve the performance of the system.

Redundant features are those that are correlated with other features and not relevant in the sense that they do not improve the discriminatory ability of a set of features.

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

Ans: Various distance measurements used to determine feature similarity:

1. Euclidean
2. Manhattan
3. Minkowski
4. Canberra
5. Cosine
6. pearson
7. Spearman
8. Chi-Square

**9. State difference between Euclidean and Manhattan distances?**

Ans:

|  |  |
| --- | --- |
| Euclidean Distance | Manhattan Distance |
| Shortest path between source and destination | Is a sum of all the real distances between source(s) and destination |
| Formula  Which is Norm2 | he Manhattan distance as the sum of absolute differences  ManhattanDistance [{a, b, c}, {x, y, z}]  Abs [a − x] + Abs [b − y] + Abs [c − z]  which is Norm1 |

10. Distinguish between feature transformation and feature selection.

|  |  |
| --- | --- |
| feature selection. | feature transformation |
| Reduces the feature set by discarding features | refers to building a new feature space from original variables. therefore called as feature extraction |
| methods: filter, wrapper, online(EMBEDDED) | methods: PCAm LDA, ICA,GPCA  Log, reciprocal square,square root transformation, power transformation |
| enables ML algorithm to train faster  reduce complexity of a model and makes it easier to interpret  improves accuracy  reduce overfitting | reduce repetition  improve performance  data integrty |
| Choosing a subset of the original pool of features. | Getting useful features from existing data. |
| Below are some real-life examples of feature selection:  Mammographic image analysis  Criminal behavior modeling  Genomic data analysis  Plat monitoring  Mechanical integrity assessment  Text clustering  Hyperspectral image classification  Sequence analysis |  |

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

1.SVD (Standard Variable Diameter Diameter)

2. Collection of features using a hybrid approach

Ans: The advantages of this approach include the ability to accommodate multiple feature

selection criteria and find small subsets of features that perform well for the target algorithm

Procedure of Hybrid algorithm

1 Begin

2 Initialize: Randomly initialize population of feature subset, initialize E and others;

3 While (stop if condition is not satisfied)

4 Evaluate fitness of all feature subset encoded in the population;

5 Find E best feature subset in the population and put them into elite pop;

6 For (each subset in elite pop)

7 Perform local search and replace it with new feature subset;

8 End For

9 Evaluate fitness of new solutions which is generated by local search;

10 Select the best solution based on fitness function as global optimum;

11 Perform evolutionary operators, i.e. selection, crossover, mutation;

12 End While

13 End

4. Receiver operating characteristic curve

A receiver operating characteristic curve, or ROC curve, is a [graphical plot](https://en.wikipedia.org/wiki/Graph_of_a_function) that illustrates the diagnostic ability of a [binary classifier](https://en.wikipedia.org/wiki/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.[10] The false-positive rate is also known as probability of false alarm[10] and can be calculated as (1 − specificity). The ROC can also be thought of as a plot of the power as a function of the Type I Error of the decision rule (when the performance is calculated from just a sample of the population, it can be thought of as estimators of these quantities). The ROC curve is thus the sensitivity or recall as a function of fall-out. In general, if the probability distributions for both detection and false alarm are known, the ROC curve can be generated by plotting the cumulative distribution function (area under the probability distribution from to the discrimination threshold) of the detection probability in the y-axis versus the cumulative distribution function of the false-alarm probability on the x-axis.