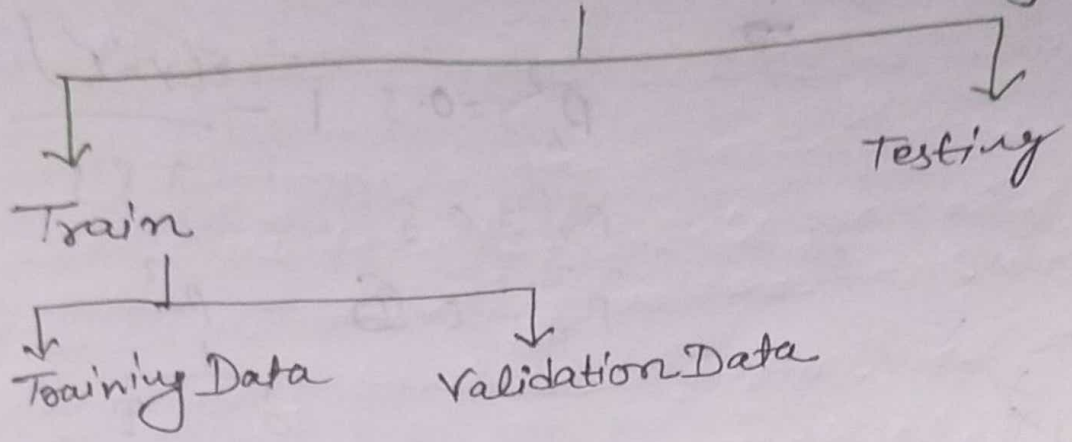
Data in Machine Learning



⊙ Training Data - The part of data we use to train our model. This is the data that your model actually sees (both input and output) and learns from.

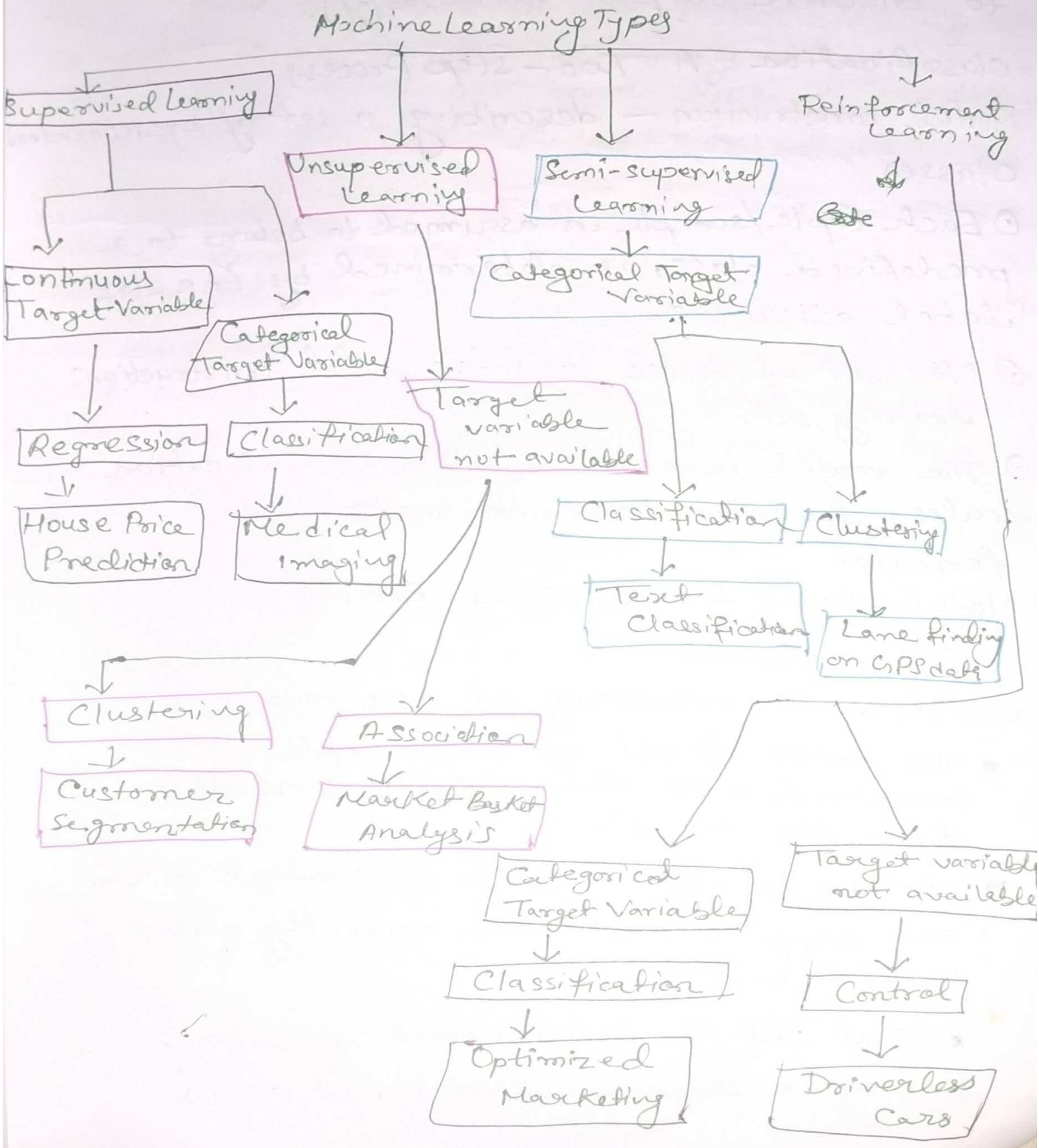
⊙ Validation Data: The part of data that is used to do a frequent evaluation of the model, fit on the training dataset along with improvement involved hyperparameters (initially set parameters before the model begins learning). This data plays its part when the model is actually training.

⊙ Testing Data - Once our model is completely trained, testing data provides an unbiased evaluation. When we feed in the inputs of Testing data, our model will predict some values (without seeing actual output). After prediction we evaluate our model by comparing it with the actual output present in the testing data. This is how we evaluate and see how much our model has learned from the experiences feed in as training data, set at the time of training.

⊙ First split their dataset into 2 - Train and Test.

⊙ After this, they keep aside the Test set, and randomly choose X% of their Train

dataset to be the actual Train set and the remaining (100 − X)% to be the Validation set, where X is a fixed number (say 80%), the model is then iteratively trained and validated on these different sets.



What is classification?

Classification is the process of predicting the class of given data points. Classes are sometim called as targets / labels or categories. Classif predictive modeling is the task of approximat a mapping function (f) from input variables to discrete output variables (y).

Classification - A Two-step Process

Model construction - describing a set of predete classes

- Each tuple/sample is assumed to belong to a predefined class, as determined by the class label attribute.
- The set of tuples used for model construction training set.
- The model is represented as classification rules, decision trees, or mathematical formula.

Model usage: for classifying future or unknown objects

- Estimate accuracy of the model.
  - The known label of test sample is compared with the classified result from the model.
  - Accuracy rate is the percentage of test set samples that are correctly classified by the model
  - Test set is independent of training set otherwise over-fitting will occur

There are two types of learners in classification as lazy learners and eager learners.

① Lazy learners

Lazy learners simply store the training data and wait until a testing data appear. When it does, classification is conducted based on the most related data in the stored training data. Compared to eager learners, lazy learners have less training time but more time in predicting

Ex. K-nearest neighbor, Case-based reasoning.

② Eager learners – Eager learners construct a classification model based on the given training data before receiving data for classification. It must be able to commit to a single hypothesis that covers the entire instance space. Due to the model construction, eager learners take a long time for train and less time to predict.

⊙ Ex. Decision Tree, Naive Bayes, Artificial Neural Networks.