Lesson-3 (Data Import & Transformation)

Data wrangling :-

↳ process of cleaning and transforming data into much more suitable format for ML algorithms

↳ Iterative process (Trial & error)

---

① Explore raw data & check general quality

② Transform raw data

③ Validate & publish the data.

Manage data :-

- Data stores (hidden)
- Datasets (specific data files)

Sets of files that contain test, validation, training data

Datastore :- Securily connect to my data

→ abstraction of azure data store

→ Compute location independence. (Several computes can access data stores

Datasets:- Explore, transform & manage data

Dataset versioning:-

* when new data is available
* when different feature engineering

Feature Engineering:-

↳ Derive new features from existing features

Curse of dimensionality :-

Feature Engineering

* data source (Relational database)
* within python environment
* ~~big~~ streaming data like spark
* during training the model.

Feature Engineering tasks:-

* Aggregation (mean, median)
* part-of (part of data)
* Binning (group of entities into bins. apply aggregation on bins)
* Flagging (derive boolean conditions)
* Frequency-based
* Embedding
* Deriving by Example

most widely- used data

↳ num (Tabular format — .
↳ text
↳ img

Text Embedding :-

Text frequency - inverse document freque
Tf - idf
word Embedding

img :- RGB - need to be translated

500 × 400 × colordepth (3).

Grey scale - (1).
Simple vector →
Multi dimension matrix.

---

feature engineering naturally happens in hiddenlayers of neural network

Convolution neural networks — dedicate to learn complex patterns in image data.

Feature selection :-

↳ choosing useful features for a model.

(1) Elimination of irrelevant, redundant, highly correlated data.

(2) Reduction of dimensionality

dimensionality reduction.

Commonly used techniques are.

* PCA (principle component analysis)
* t-SNE (t- Distributed stochastic Neighboring Entities)
* Feature Embedding.

↳ linear technique (statistical)

↳ probablistic approach (2 or 3 dimensions are resulted)

↳ used for visualization of data.
(scatter plot)

↳ multi-dimension x. x-dimensionality

↳ Encode large no of features to few features

Filter Based Feature selection:-

Permutation Feature Importance.

data drift:- causes degradation in model's performance.

↳ sensor info

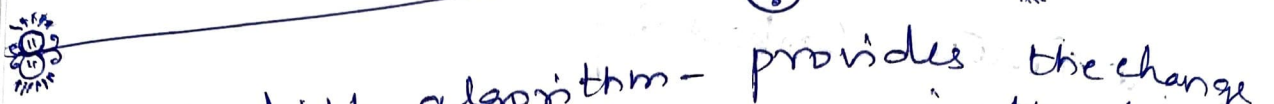
↳ sensor breaks (data quality)

↳ data changes over time
(Customer behavior not guaranteed)

↳ changes in relations b/w features
(degree of correlation)

dataset monitors — detects drift in time.

models to maintain the level of performance in time.



data drift algorithm - provides the change in the data.

baseline dataset - training dataset

target dataset - input data for the model

---

model training :

predict a value of a feature - model aim

What is the problem?
classification? (categorical)
Regression? (numerical value).

↳ choice of algorithm
↳ various approaches to get to the result.

* scale | encode data
* split data ← Training dataset
validation dataset
Test dataset.

Selecting hyperparameters - train the model - Evaluate the model.

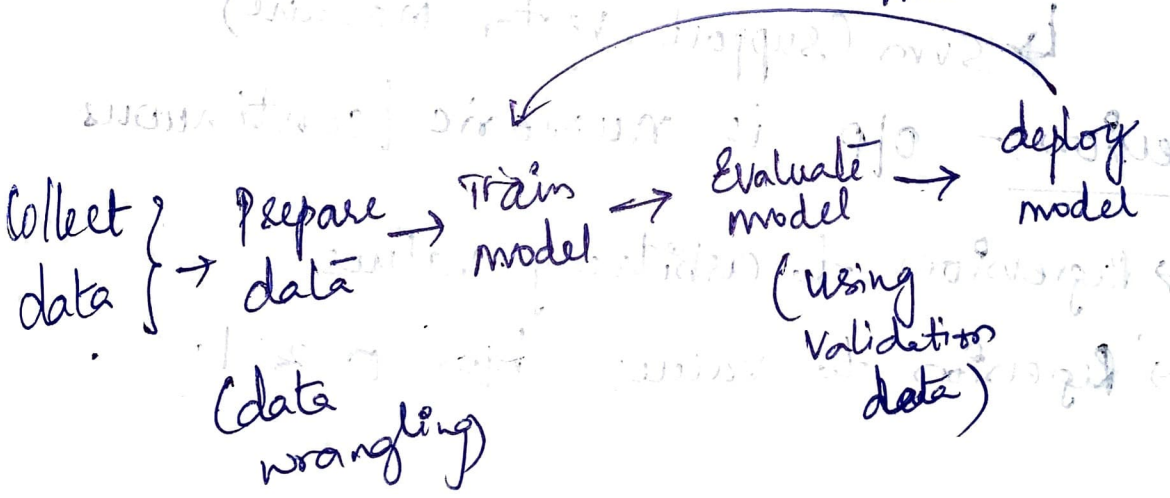
hyperparameters — values are not set before training.

Ex:- no. of layers in deep neural network
no. of clusters in a k-means
learning rate of a model.

Splitting data

- Training dataset — used to learn parameters
- validation dataset — data that checks model's performance
- Test dataset — To finally check the model performance.

↳ Tune the hyperparameters until it performs better on validation data.



Classification:-

Outputs are categorical.

Binary classification

Eg: Fraud detection,
anamoly detection

Multi class single lable. classification
↳ o/p can belong to a single clas.

Multi class multiple lable classification
↳ o/p can belong to multiple classes

Eg:- several tags to texts

classification algorithms.

↳ logistic regression.

↳ svm (support vector machine)

Regression:- o/p is numeric | continuous

↳ Regression to arbitrary values

↳ Regression to values b/w 0 & 1.

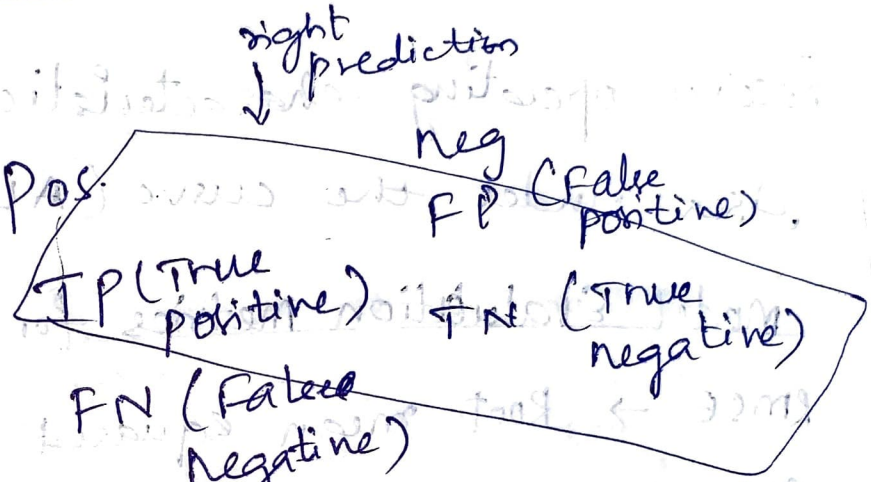
* Linear Regression

* Decision forest Regressor

Evaluating model performance:-

test dataset is a portion of labeled data that is split off and reserve for model evaluation

Confusion matrices:-



$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + FP}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1 score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

⇒ Always used in pairs rather than

Receiver operating characteristics (ROC)

Area under the curve (AUC)

Model Evaluation Metrics for Regression:-

RMSE → Root mean squared error.

MAE → mean absolute Error.

R-squared → how close the values are to the regression line.

Spearman → strength of the correlation

Histogram of residuals. does bias.

regressor

↳ decent level of performance

Train and Evaluate a model:-

→ Ensemble learning → multiple models

→ Automated Machine Learning → Automate

train multiple models, and collect the results

Ensemble learning :- combines multiple models to produce one predictive model.

Three main types:-

- Bagging or Bootstrap
  - Reduces variance
  - Homogeneous learners
  - Random sampling
  - Equally weighed average
- Boosting
  - weak, homoge
  - Reduce bias
  - sequential learning
  - weighted avg predictio
- Stacking
  - heterogeneous

Le-4

Supervised & Unsupervised learning:

Types of classification:-

- classification on tabular data
- classification on image (or) sound data
- classification on text data.