CSE 445 / L-1 / 21.05.2025 /

Absent

Attendance - 5%.

Assignment - 20%.

↳ 3 out of 4

i. ML Project base Code

ii. Project Proposal

iii. VIVA on Project progress

iv. Video demonstration of project.

Midterm - 25%.

Final - 25%.

Project - 25%.

L-2 / 26.05.2025

Mean :

— Arithmetic mean:

⇒ direct average of a data set. is found by adding all numbers in the data set and then dividing by the number of values in the set.

⇒ Harmonic Mean:

- defined as the reciprocal of the average of the reciprocals of the data values.

* Median:
  - is the middle value when a data set is ordered from least to greatest.

* Mode:
  - is the number that occurs most often in a data set.

* Conditional Probability:
  - is a measure of the probability of an event occurring, given that another event has already occurred.

* Tom Mitchell:
  - a computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at task in T, as measured by P, improves with experience E.

* Arthur Samuel:
  - Machine learning is the field of study that gives the computer the ability to learn without being explicitly programmed.

* The difference between traditional programming and the ML approach!

Slide - 9, 10

* ML Approach:

- allow fine tuning and make long list of rules.
- can adapt in changing environment.
- provide insights from large amount of data.
- solve complex problem.

* Deep learning:

- subset of ML
- automatic features extraction
- function like human brain.
- more data, better prediction.

*

AI > ML > DL

* Narrow AI:

- train for specific task.

* Parametric knowledge:

- learn from data.

L-3 / 28.05.2025

Supervised Learning:

- trained with labeled data.

⇒ - Linear Regression
- Logistic Regression
- SVM
- Decision Tree
- NN

Unsupervised Learning:

- data has no label.
- model try to make group based on familiar features.

⇒ Used in
- clustering
- anomaly detection
- association mining
- data preprocessing

⇒ - K-means
- SVD
- PCA
- ICA

Semisupervised Learning:

- partially labeled data
- model categorise the data based on familiar feature and then asked for the label to human.

Reinforcement Learning:

- first provide a output, based on the feedback of the output, it try to re-learn its state for the data

Batch Learning:

- learning is not possible after deployment.
- for new data, we need to train from scratch.

Online Learning:

- continue learning after deployment as new data comes.
- use parallel computing, no down~~time~~. time.

Instance Based Learning:

- memorize known data and try to match with these
- KNN

Model-Based Learning:

- more generalize, it divide the data area and build a model, then predict the output.

3 V's of Big Data:

- Volume: amount of data
- Variety: data ~~fo~~ for each possibility.
- Velocity: how the data change over time

Overfitting:

- very good on training data
- very bad on test data.
- model need to be simple respect to the data.
- need to tune the hyperparameters to control the regularization.

Data Split:

- Training
- Validation
- Test

⇒ Good mode: 60-20-20

⇒ an okay model: 70-15-15

⇒ barely acceptable model: 80-10-10

allowed only when we have millions of intances in the dataset.

L-4 / 02.06.2025

- Problem with the housing data:
  - price changes over time.
  - influencing features also change

Might be a good research area.

- Classification problem:
  - predict the class of the data
- Regression problem:
  - predict continuous variable.

⇒ one value ⇒ univariate regression problem

multiple value ⇒ multivariate regression problem.

Performance Measurement:

- RMSE : Root mean squared error

$$= \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(h(x^{(i)}) - y^{(i)}\right)^2}$$

- MAE : Mean absolute error

$$= \frac{1}{m}\sum_{i=1}^{m}\left|h(x^{(i)}) - y^{(i)}\right|$$

RMSE:
- straight line distance
- Euclidean norm or L2 norm

MAE:
- city block distance
- Manhattan norm or L1 norm

} differences available on google

⇒ Describe Code from Colab:

Correlation between attributes:

⇒ 1:
- if x goes up, y goes up too
- both are in same direction

⇒ -1:
- x and y in opposite direction.

⇒ ≈0:
- no relation between them.

L-5/04.06.2025/

Explained Code from colab.

Steps:

(i.) Download the data

(ii) Quick look at Data Structure

- plot histogram

(iii) Data visualization:

- scatter plot
- bar plot

(iv) Preprocess the data

- remove outlier
- remove or manage NaN value
- Transform the data
  - use encoder
  - scaling

(v) Split the data : stratified sampling

(vi) Select model and train.

- validate using validation set

(vii) Test the model:

- using test set.

L-06 / 16.06.2025

- What is one-hot encoder?
- From similar correlation, we need to take only one, as the effect is same.
- During encode, we need to maintain the order.
- Remove outlier before apply normalization.
  - or use standardization.
- What the fit and fit_transform do in the custom transformer method?

L-07 / 18.06.2025

- Confusion Matrix:

TP ⇒ True Positive ⇒ Actual true & Predicted True.

TN ⇒ True Negative ⇒ Actual Negative & Predicted Negative

FP ⇒ False Positive ⇒ Actual Negative & Predicted Positive

FN ⇒ False Negative ⇒ Actual Positive & Predicted Negative

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

$$\text{Precision} = \frac{TP}{TP+FP}$$; How precisely this model can ~~ide~~ identify the positive class.

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

$$\text{F1 Score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

~~Precision & Recall~~ $\frac{1}{}$