EL Activity: Hand Written Digit Recognition (Machine Learning and Evaluation Metrics)

Presented By:

Dr. Aditi Sharma

Assistant Professor

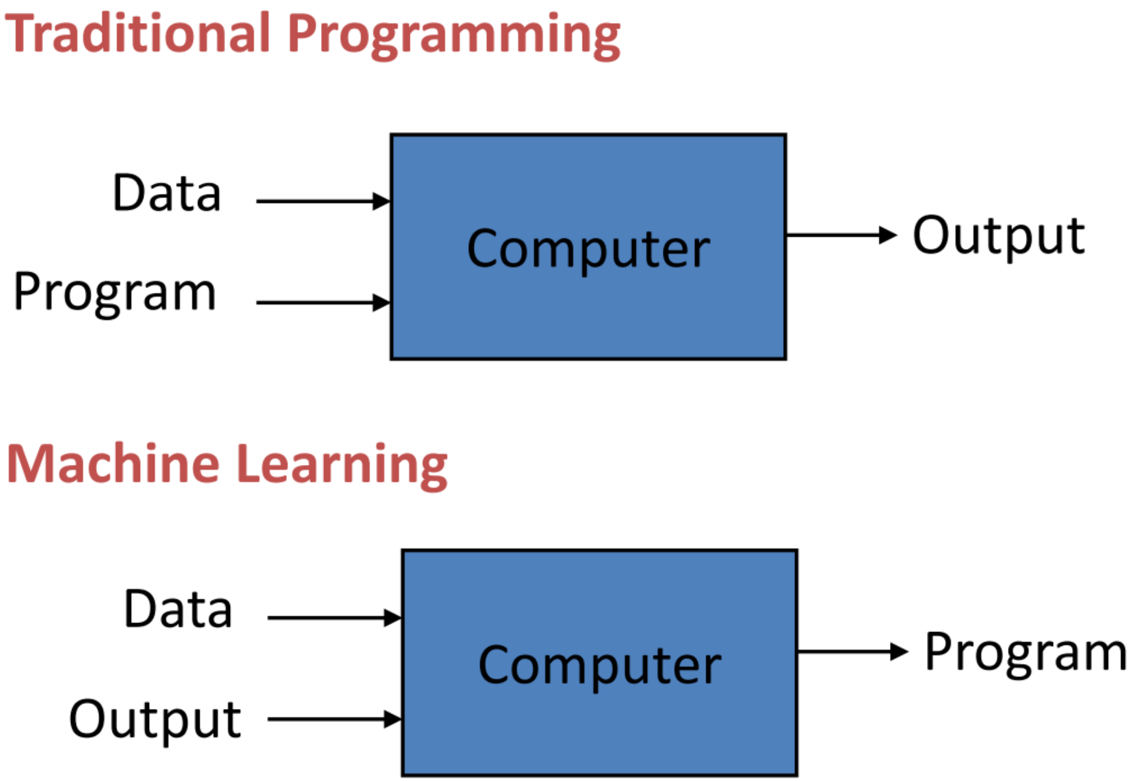
Thapar Institute of Engineering and Technology

Topics Covered

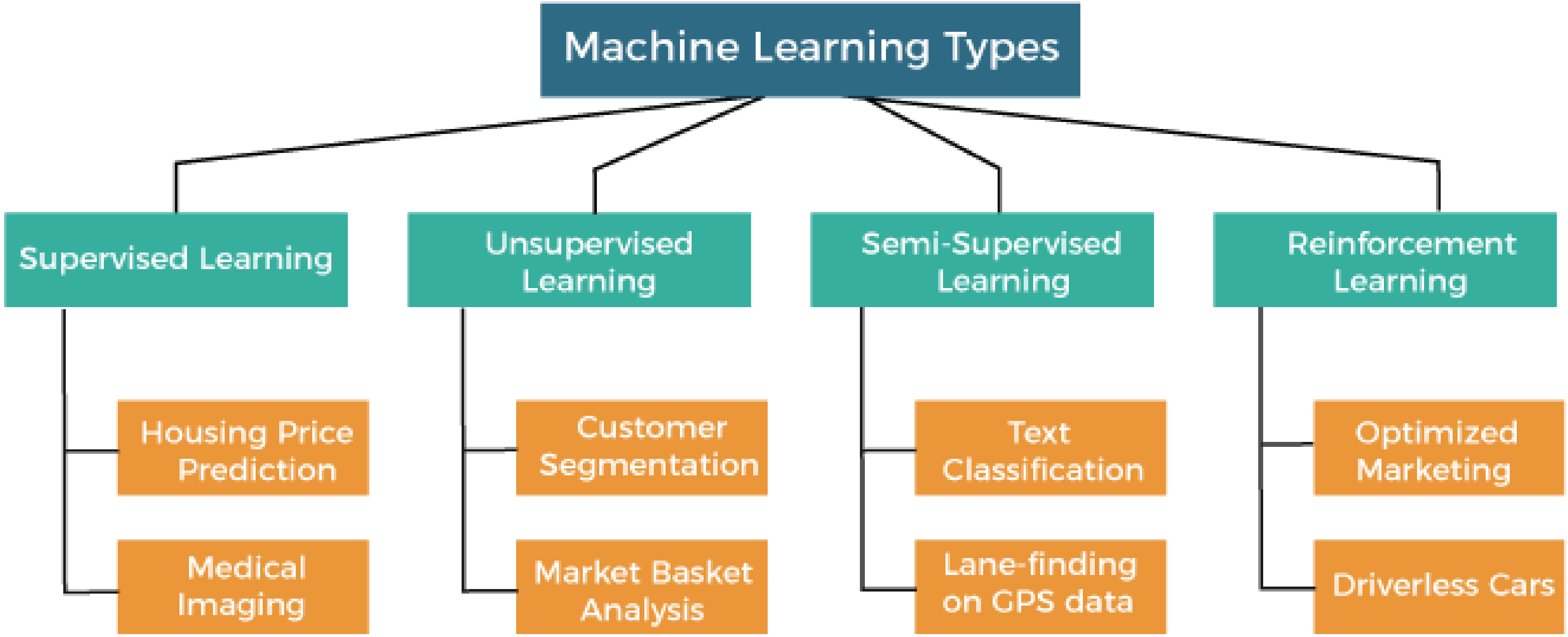
- Machine Learning
- Types of Learning
- Classification Problem
- Handwritten Digit Recognition
- Dataset
- KNN
- Accuracy
- Confusion Matrix
- Precision-Recall
- F-1 Score

MACHINE LEARNING

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 tasks in T, as measured by P, improves with experience E.



Types of Machine Learning

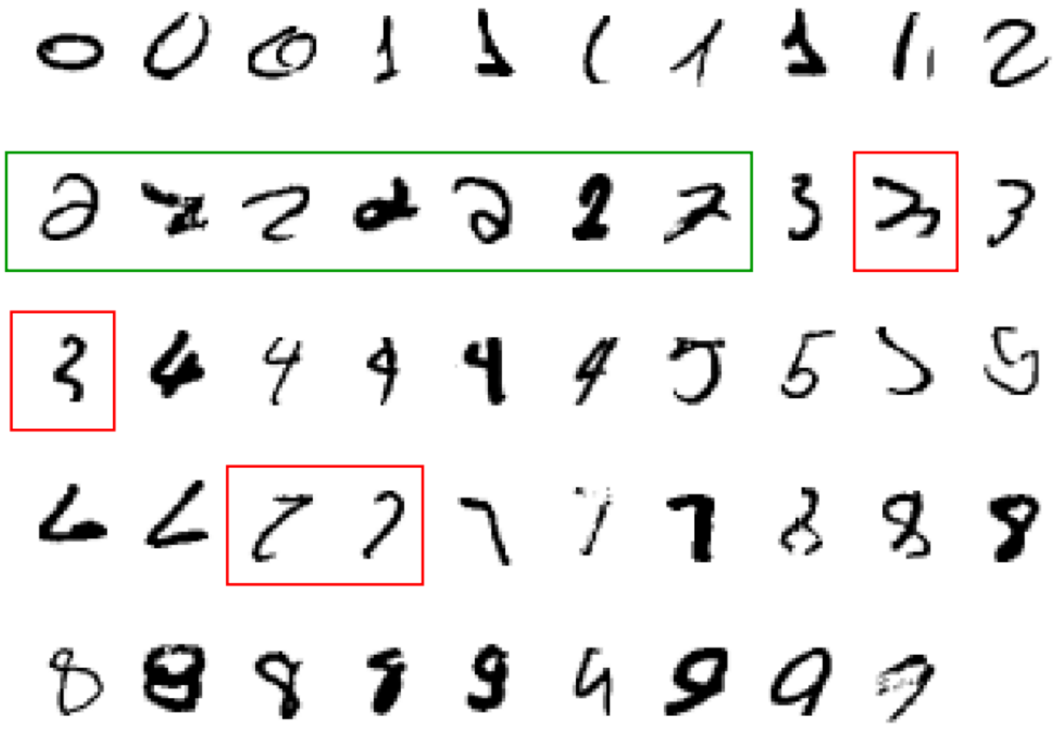


Classification Problem

- Assigns an instance to one of the predefined category or class.
- Model learning on labelled data.
- Unseen data given to learned model.
- Output provided either in terms of class labels or probabilities.
- Performance evaluation metric for a model is chosen based upon problem statement, dataset and type of output.

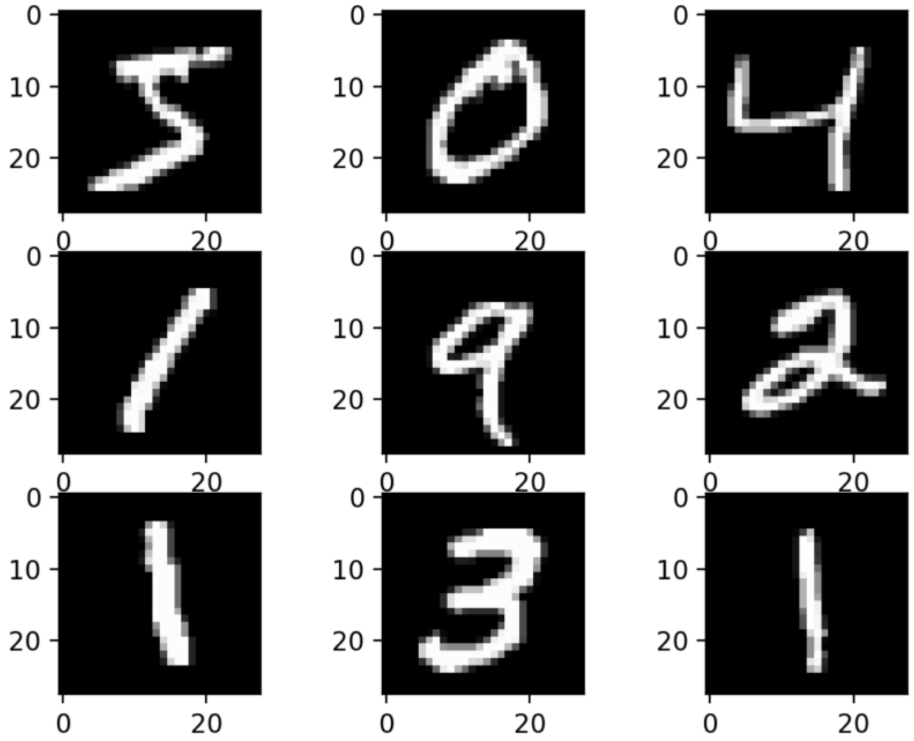
Handwritten Digit Recognition

- To recognize images of Handwritten digits based on classification methods for multivariate data.
- Optical Character Recognition (OCR)
  - Predict the label of each image using the classification function learned from training
- OCR is basically a classification task on multivariate data
  - Pixel Values -> Variables
  - Each type of character -> Class



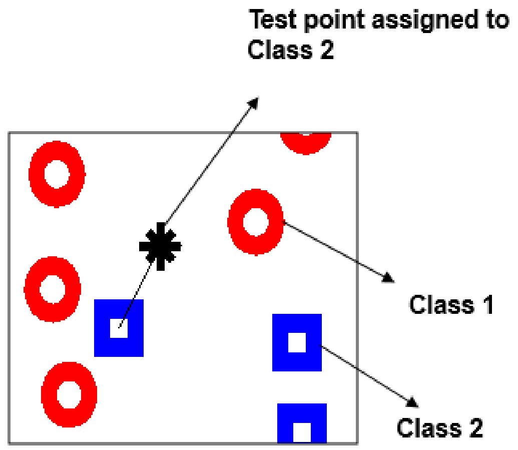
Dataset

- Challenges in processing of images due to differences in:
    - Size
    - Resolution
    - Line Thickness
    - Background and Digit Color
    - Shear etc..
- MINIST dataset:
    - Publicly available processed Dataset.
    - Created in 1994.
    - Originally contained 128*128 binary images, now available as 28*28 grayscale images.
    - Contains around 70000 images of 10 digits.



K-Nearest Neighbor (KNN)

- It is a supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point.
- Finds the nearest neighbors from the training set to test image and assigns its label to test image.
- No Assumption about the data.
- Euclidean Distance to find nearest neighbor.
- Compute the k nearest neighbors and assign the class by majority vote.
- K value can be changed, and the model accuracy varies for that.



Accuracy

- Simplest and Most commonly used approach.
- How many predictions made by the model are correct.

$$Accuracy = \frac{Correct\ Predictions}{Total\ Predictions\ Made}$$

- Error rate of a model is calculated from the accuracy as well.

$$Error\ Rate = 1 - Accuracy$$

- Works best for balanced dataset, i.e., when every class in the dataset is equally important.
- Not a reliable indicator of classifier's effectiveness in times of imbalanced dataset.

Issues of Accuracy metric

- Given an 8-year dataset of Stock market, labelling the days of Bank Nifty as bearish and bullish.
- Out of 2000 instances 1850 days are bullish, and 150days are bearish.
- Train : Test split on 80:20.
- Dumb model learned each day is bullish.
- Testing Model on 400 instances (360 bullish, 40 bearish).
- Accuracy $= \frac{360}{400}$ = 90%
- Error Rate= 10%
- On deploying model fails miserably, when the bear market begins.

Confusion Matrix

- A visual representation of model performance.
- Shows a more detailed breakdown of correct and incorrect classification for each class.
- General idea is to count the number of times instances of one class are classified as other.
- Provides model performance for each class.
- Rows of the matrix represents actual class.
- Columns represents predicted class.
- Call confusion_matrix function on test dataset.

Predicted→ / Actual↓	YES	NO
YES	Correctly Predicted Yes	Incorrectly Predicted No
NO	Incorrectly Predicted Yes	Correctly Predicted No

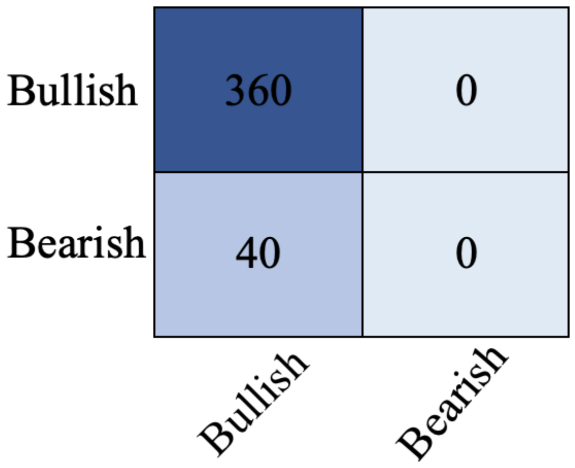
Confusion Matrix

- A confusion matrix for two classes (+, -).
- There are four quadrants in the confusion matrix, which are symbolized as below:
- **True Positive (TP)** : The number of instances that were positive (+) and correctly classified as positive (+v).
- **False Negative (FN):** The number of instances that were positive (+) and incorrectly classified as negative (-). It is also known as **Type 2 Error**.
- **False Positive (FP):** The number of instances that were negative (-) and incorrectly classified as (+). This also known as **Type 1 Error**.
- **True Negative (TN):** The number of instances that were negative (-) and correctly classified as (-).

	C_1	C_2
C_1	True positive	False negative
C_2	False positive	True negative

	+	-
+	++	+-
-	-+	--

Confusion Matrix



- Perfect classifier have only true positives and true negatives.
- Non-diagonal values should be minimum or zero(perfect classifier).
- Accuracy can be calculated from Confusion matrix.

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

- Color coded matrix help visualization of model performance.
- Drawback: not understandable by layman.
- Concise Results serve better for non-technical persons.
- Many concise results can be drawn from confusion matrix.

Precision- Recall

Precision:

- How many predicted positive values are actually positive.
- It is defined as the fraction of the positive examples classified as positive that are really positive.

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

- It is also known as Positive Prediction Value(PPV).

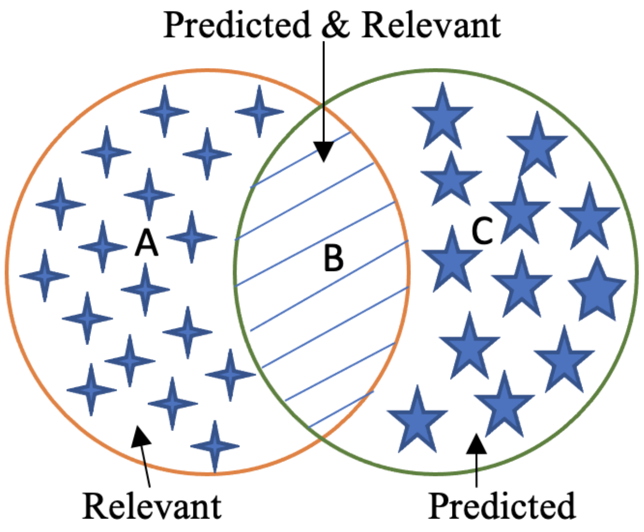
Recall:

- Out of total actual positive values, how many positives did model predict.

$$TPR = \frac{TP}{TP+FN}$$

- Also known as TPR.

Precision- Recall



Precision answers:

"Out of the items that the classifier predicted to be relevant how many are truly relevant?"

Recall Answers:

"Out of all the items that are truly relevant, how many are identified by the classifier?"

F-1 Score

- When both Precision and Recall have importance for the problem statement.
- Harmonic Mean of Precision and Recall.

$$\text{F-1 Score} = \frac{2}{\frac{1}{Precision}+\frac{1}{Recall}} = 2 * \frac{Precision * Recall}{Precision+Recall} = \frac{2*TP}{2*TP+FP+FN}$$

- Regular mean treats all the value equally.
- Harmonic mean gives more weightage to low values.
- A classifier will get a high F-1 Score if both Recall and Precision are high.

THANK YOU