Om Shete
THADOMAL SHAHANI
TSEC
ENGINEERING COLLEGE
Group 7

Assignment I

Q.1 What is learning? Explain different types of learning with example.

⇒

- Machine learning is a subset of A.I, which enables the machine to automatically learn from data, improve performance from past experiences and make predictions.
- Learning contains a set of algorithms that work on a huge amount of data.
- Data is fed to these algorithms to train them, and on the basis of training, they build the model and perform a specific task.
- Types of Learning :-

Types of Learning → Supervised Learning, Unsupervised Learning, Semi-Supervised Learning, Reinforcement Learning

1) Supervised Learning

⇒ Supervised learning is defined as when a model gets trained on a "Labelled Dataset".

- Labelled Dataset have both inputs and outputs parameters.
- In supervised learning algorithms learn to map points between inputs and correct outputs.
- It has both training and validation datasets labelled.

- For e.g., consider the scenario where you have to build an image classifier to differentiate between cats and dogs.
- If you feed the datasets of dogs and cats labelled images to the algorithm, the machine will learn to classify between a dog or a cat from these labeled images.
- When we input new dog or cat images that it has never seen before, it will use the learned algorithms and predict whether it is a dog or cat.
- This is how supervised learning works, and this is particularly an image classification.

2) Unsupervised Learning

=> Unsupervised learning is a type of machine learning technique in which an algorithm discovers patterns and relationships using unlabeled data.
- Unlike supervised learning, unsupervised learning doesn't involve providing the algorithm with labeled target outputs.
- The primary goal of unsupervised learning ~~doesn't involve providing the~~ is often to discover hidden patterns, similarities or clusters within the data, which can be then used for various purposes, such as data exploration, visualization, dimensionality reduction and more.

THADOMAL SHAHANI TSEC ENGINEERING COLLEGE

3) Semi-supervised Learning

⇒ Semi-supervised is a machine learning algorithm that works between supervised and unsupervised learning so it uses both labelled and unlabelled data.

- It's particularly useful when obtaining labeled data is costly, time-consuming, or resource-intensive.
- This approach is useful when the dataset is expensive and time-consuming.
- Semi-supervised learning is chosen when labeled data require skills and relevant resources in order to train or learn from it.

4) Reinforcement Learning

⇒ Reinforcement learning algorithm is a learning method that interacts with the environment by producing actions and discovering errors.

- Trial, error and delay are the most relevant characterstics of reinforcement learning.
- For e.g., consider that you are training an AI agent to play a game like chess.
- The agent explores different moves and receives positive or negative feedback based on the outcome.
- Reinforcement learning also finds applications in the which they learn to perform tasks by interacting with their surroundings.

THADOMAL SHAHANI
TSEC
ENGINEERING COLLEGE

Q.2 Define overfitting and underfitting. How evaluate a ML model for overfitting or underfitting explain using diagram. What measures need to be taken in case of overfitting and underfitting?

⇒

- Overfitting

⇒ A statistical model is said to be overfitted when the model does not make accurate predictions on testing data.
- When a model gets trained with so much data it starts learning from the noise and inaccurate data entries in our data set.
- And when testing with test data results in High variance. Then the model does not categorize the data correctly, because of too many details and noise.

- Underfitting

⇒ A statistical model or machine learning algorithm is said to have underfitting when a model is too simple to capture data complexities.
- It represents the inability of the model to learn the training data effectively result in poor performance both on the training and testing data.

THADOMAL SHAHANI
TSEC
ENGINEERING COLLEGE

- In simple terms, an underfit model's are inaccurate, especially when applied to new, unseen examples.

- Techniques to evaluate Overfitting & Underfitting.

1) Cross-validation

⇒ Use k-fold cross-validation to assess the model's performance. The dataset is divided into k subsets, and the model is trained and validated k times, each time using a different subset as the validation set and the remaining as a training set.

- Average the performance metrics across the k trials.

2) Learning Curves

⇒ Plot learning curves for both training and test errors. Learning curves shows the model's performance on the training set and validation set as function of the number of training examples or training epochs.

- Measures to Address Overfitting:

1) Increase Training Data

⇒ Gather more training data to help the model generalize better.

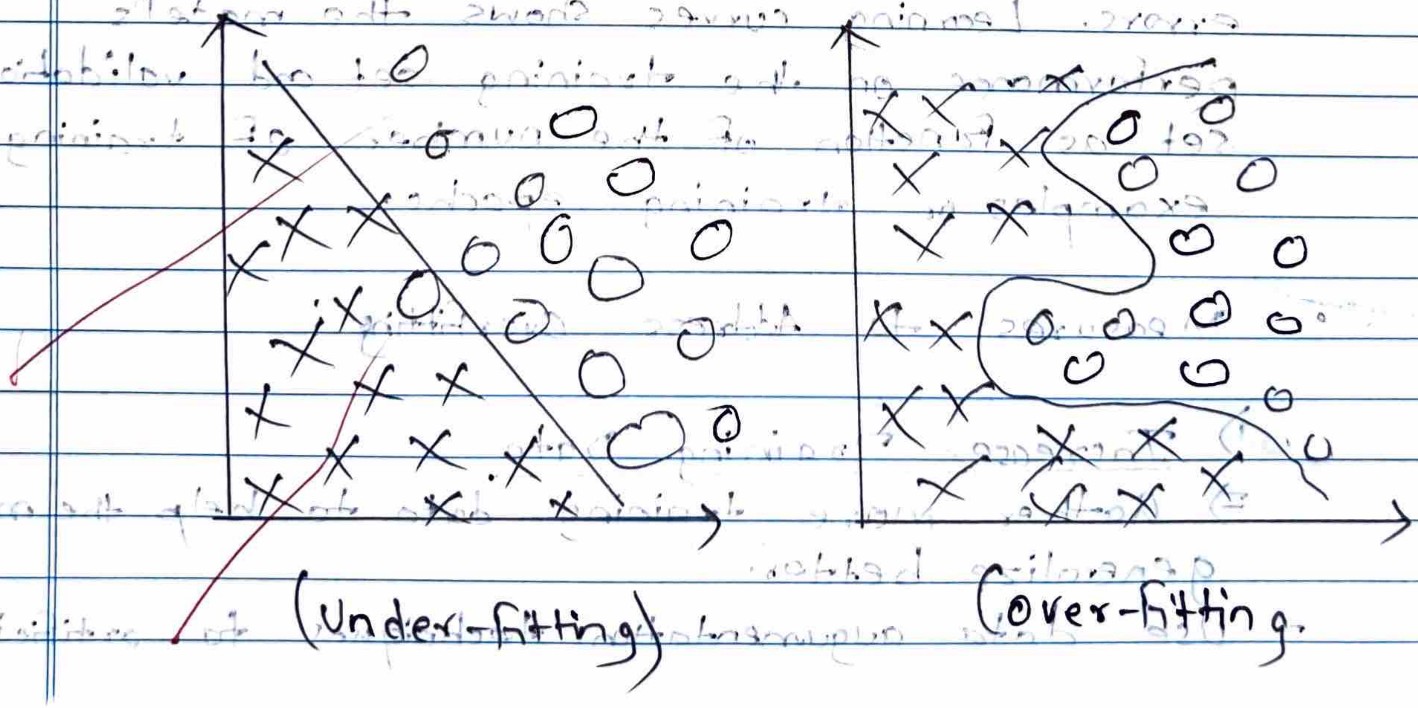
- Use data augumentation techniques to artificially increase the size of the training dataset.

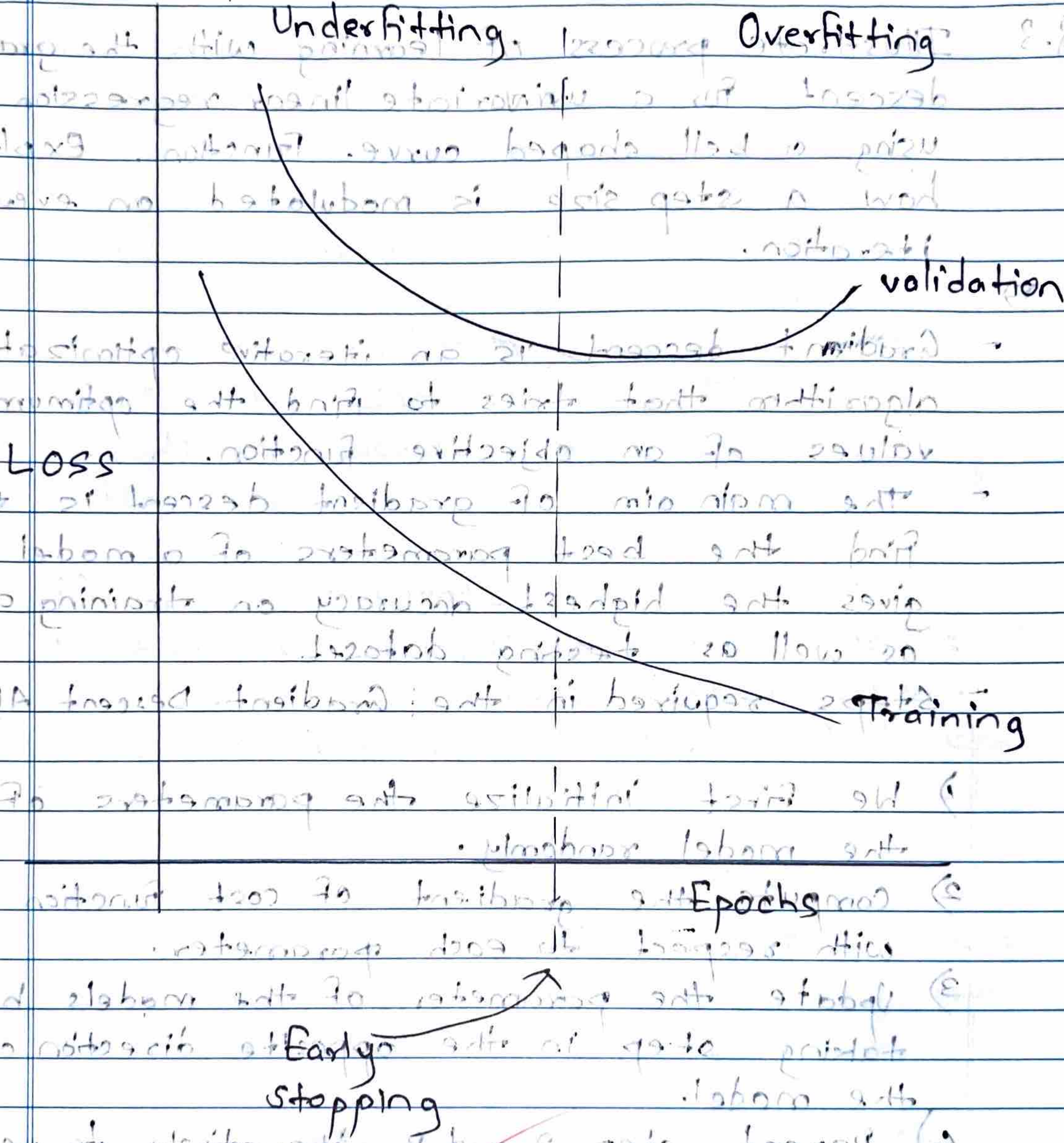
2) Pruning
⇒ Remove the parts of tree that provide little power to the predict target variables to reduce the model complexity.

- Measures to address Underfitting :-

1) Increase Model complexity
⇒ Use more complex models or algorithms that can capture more intricate patterns in the data.
- Increase the number of features or parameters.

2) Train Longer
⇒ Train the model for more epochs to allow it to learn more from the data.




Underfitting
Overfitting
validation
Loss
Training
Epochs
Early stopping

THADOMAL SHAHANI
TSEC
ENGINEERING COLLEGE

Q.3 Illustrate process of learning with the gradient descent for a univariate linear regression, using a bell shaped curve. Function. Explain how a step size is modulated on every iteration.

⇒

- Gradient descent is an iterative optimization algorithm that tries to find the optimum values of an objective function.
- The main aim of gradient descent is to find the best parameters of a model which gives the highest accuracy on training data as well as testing dataset.
- Steps required in the Gradient Descent Algorithm:-

1) We first initialize the parameters of the model randomly.
2) Compute the gradient of cost function with respect to each parameter.
3) Update the parameter of the models by taking step in the opposite direction of the model.
4) Repeat step 2 and 3 iteratively to get the best parameter for defined model.

- To apply gradient-descent using programming language we make use of four functions:-

THADOMAL SHAHANI
TSEC
ENGINEERING COLLEGE

i) gradient_descent
⇒ We make predictions on dataset and compute the difference between predicted and actual target value and accordingly update the parameter to return it.

ii) compute_prediction
⇒ In this function, we will compute the prediction using the parameters at each iterations.

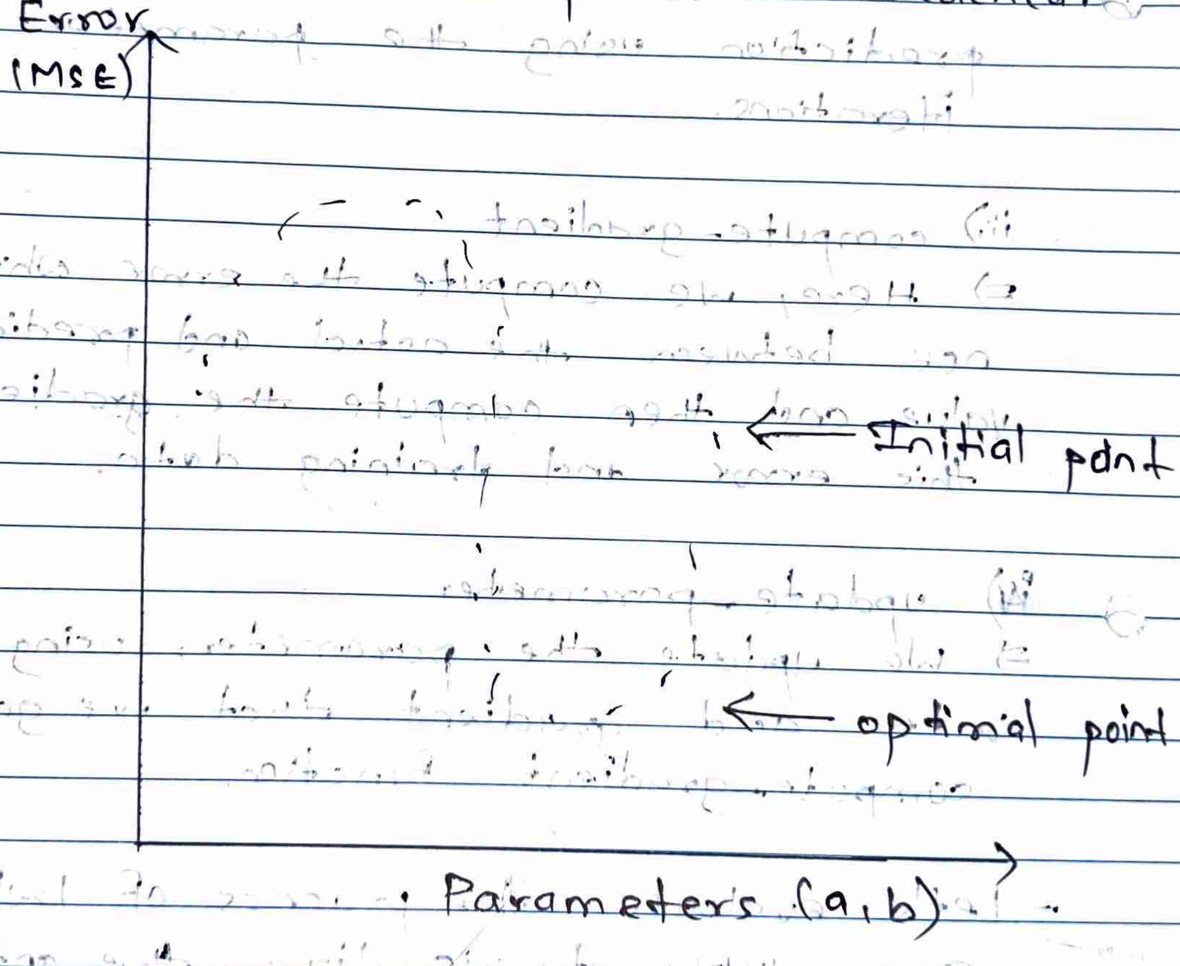
iii) compute_gradient
⇒ Here, we compute the error which is difference between the actual and predicted target value and then compute the gradient using this error and training data.

iv) update_parameter
⇒ We update the parameters using learning rate and gradient that we got from compute_gradient function.

- Let's evaluate the process of bell-shaped error curve to visualize the optimization.

- ⇒ Imagine plotting the curve error (MSE) as a function of the model parameter a and b

- The error surface might look like a bell-shaped curve.
- The minimum point on this curve represents the optimal values of a and b where the error is minimised.
- Gradient descent moves the parameters the iteratively downhill on this surface until the minimum point is reached.



THADOMAL SHAHANI
TSEC
ENGINEERING COLLEGE

Q.4 Explain the following performance evaluation parameters with the help of confusion matrix. Illustrate using appropriate example:

a) Accuracy

⇒

- A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known.
- It consists of four outcomes :-

i) True Positive (TP) : The model correctly predicts the positive value/class.

ii) True Negative (TN) : The model correctly predicts the negative class.

iii) False Positive (FP) : The model incorrectly predicts the positive class.

iv) False Negative (FN) : The model incorrectly predicts the negative class.

- Accuracy is the ratio of correctly predicted observations to the total observations.

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

THADOMAL SHAHANI
TSEC
ENGINEERING COLLEGE

b) Precision
⇒
- Precision measures the proportion of true positive predictions in the total predicted positives.
- It indicates how many of the positively classified instances were actually positive.

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

c) Recall
⇒
- Recall measures the proportion of actual positives that were correctly identified by the model.
- It is also known as Sensitivity or True positive rate.
-

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

d) F1-Score
⇒
- The F1-score is the harmonic mean of Precision and recall.
- The F1-score takes both false positive and false negative into account.

THADOMAL SHAHANI
TSEC
ENGINEERING COLLEGE

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

e) Specificity

⇒

- Specificity measure the proportion of actual negatives that were correctly identified by the model.
- It is also known as the True Negative Rate.

$$Specificity = \frac{TN}{TN + FP}$$

f) ROC-AUC Curve

⇒

- The ROC curve is a graphical representation of model's performance across all classification thresholds.
- It plots the True Positive Recall (Rate) against the False positive Rate at various threshold settings.
- The AUC provides an aggregate measure of performance across all possible classification threshold.

- True positive Rate = $\frac{TP}{TP+FN}$

- False positive Rate = $\frac{FP}{FP+TN}$

- The higher the AUC, the better the model's ability to distinguish between positives and negative classes.

A

22/9/24