SNS COLLEGE OF ENGINEERING

Kurumbapalayam(Po), Coimbatore – 641 107
Accredited by NAAC-UGC with ‘A’ Grade
Approved by AICTE, Recognized by UGC & Affiliated to Anna University, Chennai

Department of AI &DS

Course Name – 19AD602 DEEP LEARNING

III Year / VI Semester

Unit 2-DEEP NETWORKS

Topic: REGULARIZATION

Case Study

A deep learning model used for detecting fraudulent transactions was overfitting due to a small training dataset. By applying L2 regularization, the model improved its generalization, reducing false positives by 20% and increasing accuracy on unseen data.

Sum of Squares $= \sum (y-\hat{y})^2$ + penalty

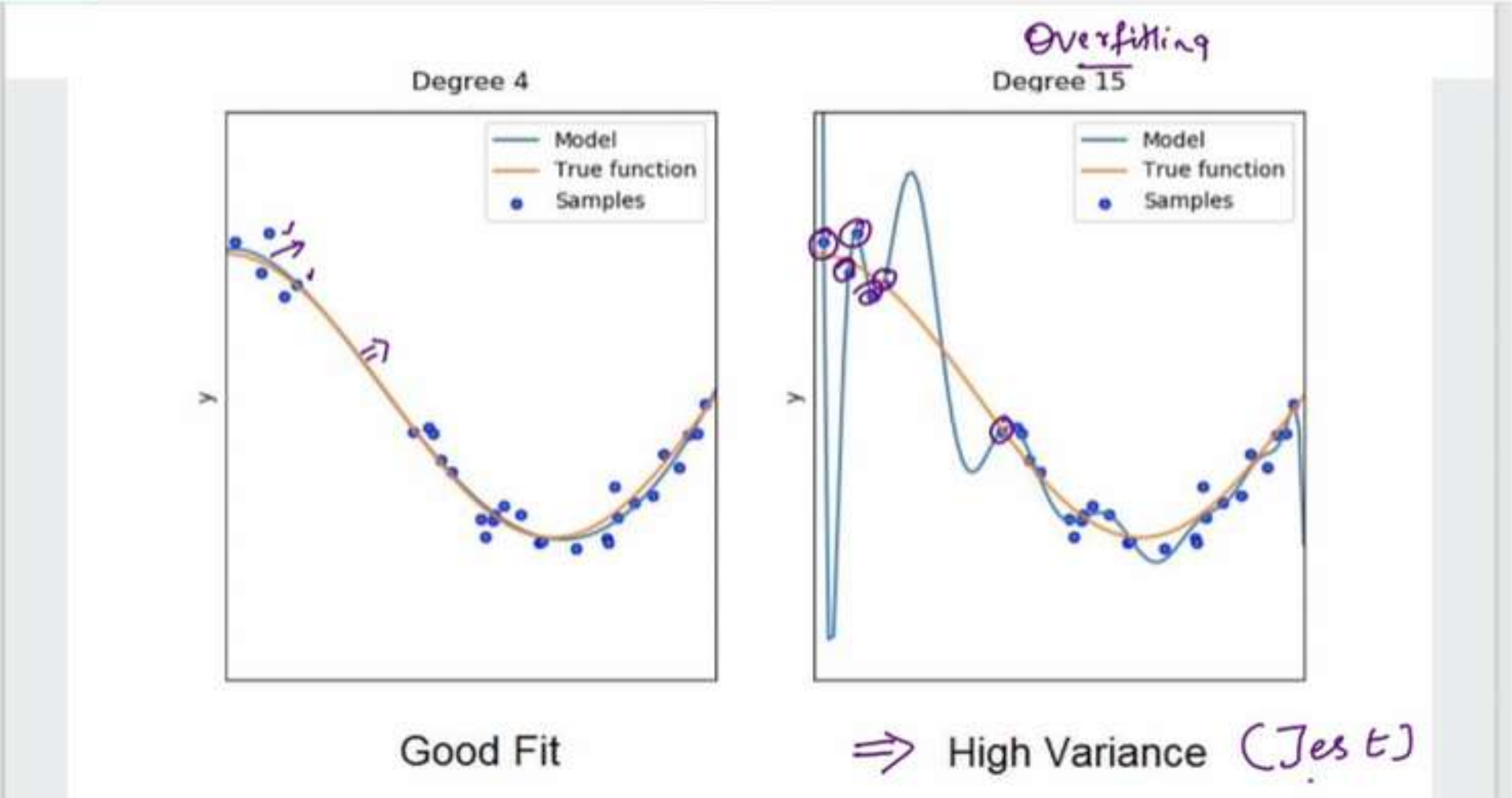
What Is Regularization?

Regularization means restricting a model to avoid overfitting by shrinking the coefficient estimates to zero. When a model suffers from overfitting, we should control the model's complexity. Technically, regularizatior avoids overfitting by adding a penalty to the model's loss function:

$$\text{Regularization} = \text{Loss Function} + \text{Penalty}$$

There are three commonly used regularization techniques to control the complexity of machine learning models, as follows:

- L2 regularization
- L1 regularization
- Elastic Net


Overfitting
Degree 4
Degree 15
Model
True function
Samples
y
Good Fit
⇒ High Variance
(Test)

Need of Regularization

- Overfitting refers to the phenomenon where a neural network models the training data very well but fails when it sees new data from the same problem domain.
- Overfitting is caused by noise in the training data that the neural network picks up during training and learns it as an underlying concept of the data.
- This learned noise, however, is unique to each training set. As soon as the model sees new data from the same problem domain, but that does not contain this noise, the performance of the neural network gets much worse.
- The reason for this is that the complexity of this network is too high.
- The model with a higher complexity is able to pick up and learn patterns (noise) in the data that are just caused by some random fluctuation or error.
- Less complex neural networks are less susceptible to overfitting. To prevent overfitting or a high variance we must use something that is called regularization.

L2 Regularization

A linear regression that uses the L2 regularization technique is called *ridge* regression. In other words, in ridge regression, a regularization term is added to the cost function of the linear regression, which keeps the magnitude of the model's weights (coefficients) as small as possible. The L2 regularization technique tries to keep the model's weights close to zero, but not zero, which means each feature should have a low impact on the output while the model's accuracy should be as high as possible.

$$\text{Ridge Regression Cost Function} = \text{Loss Function} + \frac{1}{2}\lambda\sum_{j=1}^{m} w_j^2$$

Where λ controls the strength of regularization, and w_j are the model's weights (coefficients).

By increasing λ, the model becomes flattered and underfit. On the other hand, by decreasing λ, the model becomes more overfit, and with $\lambda = 0$, the regularization term will be eliminated.

L1 Regularization

Least Absolute Shrinkage and Selection Operator (*lasso*) regression is an alternative to ridge for regularizing linear regression. Lasso regression also adds a penalty term to the cost function, but slightly different, called L1 regularization. L1 regularization makes some coefficients zero, meaning the model will ignore those features. Ignoring the least important features helps emphasize the model's essential features.

w_j^2

$$\text{Lasso Regression Cost Function} = \text{Loss Function} + \lambda \sum_{j=1}^{m} |w_j|$$

Where λ controls the strength of regularization, and w_j are the model's weights (coefficients).

Lasso regression automatically performs feature selection by eliminating the least important features.

Elastic Net Regularization

$w^2 + |w|$

The third type of regularization,(you may have guessed by now) uses both L1 and L2 regularizations to produce most optimized output.
In addition to setting and choosing a lambda value elastic net also allows us to tune the alpha parameter where α = 0 corresponds to ridge and α = 1 to lasso. Simply put, if you plug in 0 for alpha, the penalty function reduces to the L1 (ridge) term and if we set alpha to 1 we get the L2 (lasso) term.

$\Rightarrow L_1$ $\Rightarrow L_2$

Cost function of Elastic Net Regularization

$$\Rightarrow J(\beta_1, \beta_2, \ldots\ldots, \beta_m) = \sum_{i=1}^{n}(y_i - \underbrace{\sum_{j=1}^{m} x_{ij}\beta_j}_{w})^2 + \lambda\underbrace{(\alpha\sum_{j=1}^{m}|\beta_j| + \frac{1-\alpha}{2}\sum_{j=1}^{m}\beta_j^2)}$$

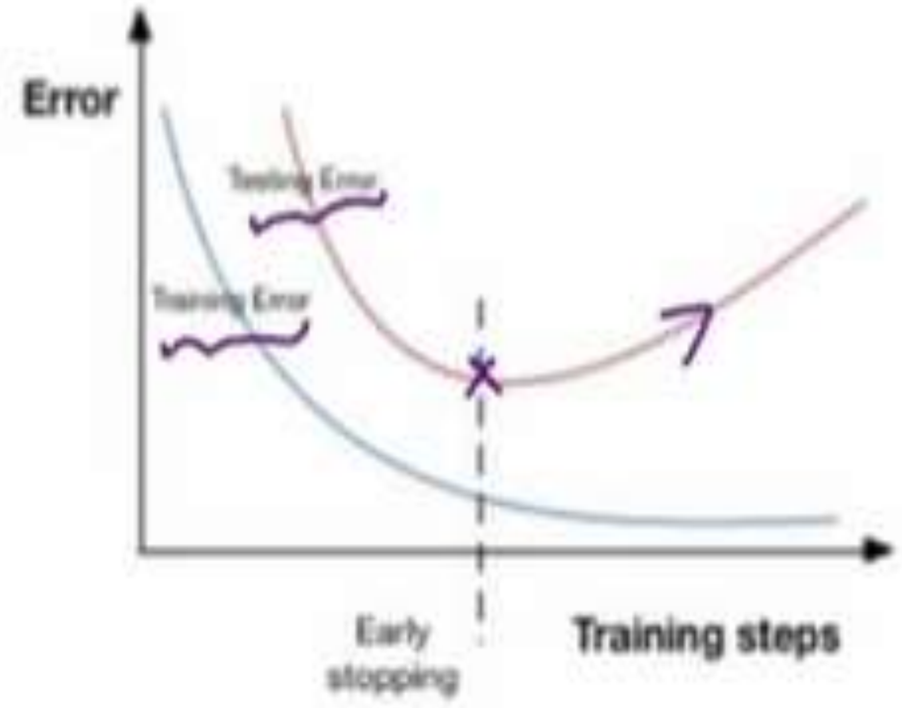
Therefore we can choose an alpha value between 0 and 1 to optimize the elastic net(here we can adjust the weightage of each regularization,thus giving the name elastic). Effectively this will shrink some coefficients and set some to 0 for sparse selection.

$\alpha \to 0 \Rightarrow L_2$ $\alpha \to 0.5$ $0.5 L_2$

$\alpha \to 1 \Rightarrow L_1$

Early stopping

Early stopping is a kind of cross-validation strategy where we keep one part of the training set as the validation set. When we see that the performance on the validation set is getting worse, we immediately stop the training on the model. This is known as early stopping.



Activity

Train a neural network on a small dataset and observe overfitting, then apply L2 regularization to reduce the overfitting and improve the model's performance on a validation set.

THANK YOU