Regularization in Machine Learning

Before we get into **Lasso** and **Ridge**, let's first understand the **need for regularization**.

When building machine learning models, especially linear models like **Linear Regression**, the goal is to minimize the loss function (like **Mean Squared Error (MSE)**) to fit the data well. However, in many cases, especially with high-dimensional or noisy data, the model can **overfit**. Overfitting occurs when the model learns the noise in the training data instead of the underlying pattern, leading to poor performance on new, unseen data.

> **Note**
>
> **Regularization** helps address this issue by adding a **penalty** to the model's complexity, discouraging large weights (coefficients) that may lead to overfitting.

Lasso and Ridge Regularization

Two common types of regularization are **Lasso** (Least Absolute Shrinkage and Selection Operator) and **Ridge** regression. Both methods add a penalty to the linear regression objective, but they differ in the type of penalty they impose.

---

1. Ridge Regression (L2 Regularization)

How Ridge Works:

Ridge Regression, also known as **L2 regularization**, adds a **penalty term** to the ordinary least squares (OLS) objective function. The penalty is proportional to the **sum of the squares** of the coefficients.

Mathematical Formula:

In regular linear regression, we minimize the following **cost function** (the residual sum of squares or RSS):

```
m
```

J(θ)=∑(yi-yi)^2
i=1
Where:

- yi is the true value,
- y^i is the predicted value,
- m is the number of data points.

Now, in Ridge Regression, we modify this cost function by adding the **L2 penalty term**:

Ridge Regression Formula:

The objective function in Ridge regression is the sum of the ordinary least squares (OLS) loss and the L2 penalty:

$$J(\theta) = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{n} \theta_j^2$$

Where:

- J(θ) is the cost function to minimize.
- m is the number of training examples.
- yi is the actual target value for the i-th training example.
- y^i is the predicted value for the i-th training example.
- θj are the parameters (coefficients) of the model, where j ranges over the features (including the intercept if applicable).
- λ is the regularization parameter, controlling the strength of the regularization.

Explanation of Notations:

- θj: The weight coefficients (parameters) of the model.
- λ: The regularization parameter that controls how strongly the regularization affects the model. A higher λ increases the penalty on the size of the coefficients.
- θ2j: Squared values of the coefficients, which adds a penalty for large weights. This encourages the model to prefer smaller coefficients.

Where to Use Ridge:

- Ridge regression is useful when you have many features, and you want to prevent overfitting by constraining the model complexity.
- It is particularly helpful when multicollinearity exists (features are highly correlated).

---

2. Lasso Regression (L1 Regularization)

Lasso Regression, or **L1 regularization**, also adds a penalty to the loss function, but this time the penalty is proportional to the **sum of the absolute values** of the coefficients.

Lasso Regression Formula:

The objective function in Lasso regression is the sum of the ordinary least squares (OLS) loss and the L1 penalty:

$$J(\theta) = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{n} |\theta_j|$$

Where:

- J(θ)is the cost function to minimize.
- m is the number of training examples.
- yi is the actual target value for the i-th training example.
- y^i is the predicted value for the i-th training example.
- θ are the parameters (coefficients) of the model.
- λ is the regularization parameter.

Explanation of Notations:

- θj: The weight coefficients (parameters) of the model.
- λ: The regularization parameter that controls the penalty. Larger λ\lambdaλ values lead to more regularization.
- |θj|: Absolute value of the coefficients. This results in some coefficients potentially becoming exactly zero, which helps in feature selection.

Where to Use Lasso:

- Lasso regression is useful when you suspect that many features are irrelevant and you would like to perform **automatic feature selection** by setting some coefficients to zero.
- It is ideal when you expect that only a small subset of features are important for the model.

Why Regularization is Needed

Regularization is necessary to:

1. **Prevent Overfitting**: Regularization adds a penalty to the model complexity, discouraging the model from fitting the noise in the training data, which helps it generalize better to new, unseen data.
2. **Control Model Complexity**: By adding a regularization term, you prevent the model from learning overly complex patterns that do not represent the true underlying relationship in the data.
3. **Improve Generalization**: Regularization helps improve the model's performance on test data by making it less sensitive to the variations in the training data.
4. **Handle Multicollinearity**: In cases where features are highly correlated, regularization helps by shrinking the coefficients, thus stabilizing the model and making it more reliable.
5. **Feature Selection**: In the case of Lasso, it can drive some feature coefficients to zero, effectively performing feature selection and reducing the dimensionality of the model.

Choosing Between Lasso and Ridge

- **Ridge regression** is preferred when you believe that most features are relevant but you want to reduce their magnitude and prevent overfitting.
- **Lasso regression** is preferred when you believe only a few features are important and you want to reduce the number of features used by the model (since Lasso can set coefficients to zero).

Key Differences Between Lasso and Ridge

Aspect	Lasso (L1)	Ridge (L2)
Penalty	L1 norm (sum of absolute values of coefficients)	L2 norm (sum of squared values of coefficients)
Feature Selection	Can drive some coefficients to exactly zero	Does not drive coefficients to zero, only shrinks them
Use Case	Useful for feature selection and sparse models	Useful when features are highly correlated or when you want to shrink coefficients but not eliminate them
Behavior	Sparse models (some features may be excluded)	Coefficients are generally small but not exactly zero

---

Conclusion

Regularization techniques such as **Ridge** and **Lasso** are essential tools in machine learning for improving model performance, especially in preventing overfitting and controlling model complexity. Ridge is preferred when all features are believed to have some impact, while Lasso is better suited for feature selection when many features are expected to be irrelevant.

> **Note**
>
> Please refer to the .ipynb file attached to this lecture for an example to help understand