Regularization

- Regularization is a technique used in regression analysis to prevent overfitting and to improve the generalization of a model. In the context of regression, overfitting occurs when a model is too complex and fits the training data too closely, capturing noise and fluctuations that might not be representative of the underlying patterns in the data. Regularization introduces a penalty term to the loss function that the model is trying to minimize, discouraging overly complex models and promoting simpler ones.

- Regularization seeks to solve a few common model issues by:
  - Minimizing model complexity
  - Penalizing the loss function
  - Reducing model overfitting (add mode bias to reduce model variance)

- There are two common types of regularization used in regression:

1. **L1 Regularization (Lasso):**
   - In L1 regularization, a penalty term is added to the loss function proportional to the absolute values of the model's coefficients.
   - The regularization term is the sum of the absolute values of the coefficients multiplied by a regularization parameter (lambda or alpha).
   - The L1 regularization encourages sparsity in the model, meaning it tends to force some of the coefficients to be exactly zero. This can be useful for feature selection.
   - The regularized cost function for linear regression with L1 regularization is given by:

   $$J(\theta) = \mathrm{MSE} + \lambda \sum_{i=1}^{n} |\theta_i|$$

   where MSE is the Mean Squared Error, θ_i are the model coefficients, and λ is the regularization parameter.

2. **L2 Regularization (Ridge):**
   - In L2 regularization, a penalty term is added to the loss function proportional to the squared values of the model's coefficients.
   - The regularization term is the sum of the squared values of the coefficients multiplied by a regularization parameter.

- L2 regularization tends to shrink the coefficients towards zero without causing them to be exactly zero, promoting a more stable model.
- The regularized cost function for linear regression with L2 regularization is given by:

$$J(\theta) = \mathrm{MSE} + \lambda \sum_{i=1}^{n} \theta_i^2$$

where MSE is the Mean Squared Error, θ_i are the model coefficients, and λ is the regularization parameter.

- The choice of regularization parameter (λ) is important and is usually determined through techniques like cross-validation.
- Regularization helps prevent overfitting by penalizing overly complex models and promotes models that generalize well to new, unseen data.
- The appropriate type and strength of regularization depend on the specific characteristics of the dataset and the model.
- How to decide the choosing between L1 and L2 regularization:
- L1 Regularization (Lasso):
  - Suitable for situations where you suspect that many features are irrelevant or contribute little to the overall predictive power.
  - Can be effective for feature selection because it tends to force some coefficients to be exactly zero, effectively eliminating certain features from the model.
  - Useful when you want a sparse model with a smaller number of important features.
- L2 Regularization (Ridge):
  - Suitable when you have a high-dimensional dataset with many features that might be correlated.
  - Generally, L2 regularization is less prone to causing coefficients to be exactly zero, making it suitable when you don't want to completely eliminate any features.
  - Tends to distribute the regularization penalty more evenly across all features.
- In practice, a combination of L1 and L2 regularization, known as **Elastic Net regularization**, can also be used.
- Elastic Net introduces a mixing parameter that allows you to control the balance between L1 and L2 regularization.