LR 1

7-Minute Summary of Linear Regression Lecture (Part 1)

Intuition and Mathematics Behind Linear Regression

Linear regression is one of the simplest and most widely used statistical techniques for predictive modeling. It's primarily used to find a linear relationship between the independent variable(s) and a dependent variable. The intuition behind linear regression is to draw a line (or a hyperplane in cases of multiple independent variables) that best fits the distribution of data points.

Intuition

The goal of linear regression is to model the relationship between one or more predictors (independent variables) and a response variable (dependent variable). The "line of best fit" is calculated such that the differences (errors) between the observed values and the values predicted by the linear model are minimized.

For instance, if you want to predict the price of a house (dependent variable) based on its size (independent variable), linear regression will allow you to plot a line through the data points that best predicts house price based on house size.

Why We Need Train-Test Split & The Process

Why Train-Test Split Is Needed:

- The train-test split is crucial for assessing the performance of a regression model. By dividing the dataset into a training set and a testing set, you can train your model on one portion of the data and then test it on unseen data (mimicking real world data) to evaluate how well it generalizes ON UNSEEN DATA

Process:

- Typically, the dataset is randomly split into two parts: usually around 70-80% of the data for training and the remaining 20-30% for testing. The key is that the split should maintain a representative distribution of the key variables and outcomes in both datasets.

Univariate vs. Multivariate Regression

Univariate Regression:

- This involves a single independent variable to predict the dependent variable. For example, predicting house prices based solely on house size.

Multivariate Regression:

- Involves multiple independent variables. For example, predicting house prices using size, age, number of rooms, and location. This type is more complex but can provide a more accurate prediction as it considers multiple influencing factors.

Examples to Understand Structure

Univariate Example:

- Predicting the salary based on years of experience only. You could fit a simple linear regression line such as $\text{Salary}=b0+b1\times\text{Years of Experience}$.

  Salary=w0+w1×Years of Experience

Multivariate Example:

- Predicting house prices based on size, age, and number of rooms. The regression model would then look something like $\text{Price}=b0+b1\times\text{Size}+b2\times\text{Age}+b3\times\text{Rooms}$.

  Price=w0+w1×Size+w2×Age+w3×Rooms

MSE, MAE, and Other Metrics

Mean Squared Error (MSE):

- Measures the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value. MSE is very sensitive to outliers as it squares the errors before summing them.

Mean Absolute Error (MAE):

- Measures the average magnitude of the errors in a set of predictions, without considering their direction (i.e., it takes the absolute value of each error). This makes it less sensitive to outliers compared to MSE.

R-Squared

R-Squared (Coefficient of Determination):

- Measures the proportion of the variance in the dependent variable that is predictable from the independent variables. It provides an indication of goodness of fit and therefore a measure of how well unseen samples are likely to be predicted by the model. An $R2$ value of 1 indicates that the regression predictions perfectly fit the data.

Categorical Variable Handling in Regression

Categorical variables represent types of data which may be divided into groups. In regression analysis, which typically handles numerical input, categorical variables need to be appropriately transformed. Here are two common techniques:

1. **Target Variable Encoding**: This method involves replacing a categorical value with a number, typically the mean of the target variable for that category. For example, if you have a categorical feature "Color" with categories Red, Blue, and Green, and you're predicting price, then each color value could be replaced with the average price of items with that color. This method can be particularly useful when the categorical variable has many levels, as it avoids adding many new variables to the model, which could lead to dimensionality issues. However, it risks introducing target leakage, where information from the target variable is used inappropriately during training, potentially leading to overfitting.
2. **One-Hot Encoding**: This approach converts each category value into a new binary column and assigns a 1 or 0 (True/False). For example, if "Color" is a feature with three categories (Red, Blue, Green), one-hot encoding would create three new features ("Color_Red", "Color_Blue", "Color_Green"), each of which would be a binary column reflecting the presence or absence of the respective category. This method can significantly increase the data dimension but is very effective because it does not impose any ordinal relationship between the categories.