Introduction to Maximum Likelihood Estimators and Ridge Regression

Eli Mecinas Cruz

September 2024

1 Introduction

We often try to predict things based on educated guesses. These educated guesses are based on past experiences, our understanding of how things should work, and sometimes, data. Imagine you are at a carnival, given the chance to win a prize if you can guess how many candies are inside a jar. To win, you will make smart, educated guesses based on the size of the jar, the size of the candies, and maybe even a little math. Even though you can't know the exact number, some guesses will be more likely than others. Maximum Likelihood Estimation works in a similar way: it uses all the available to narrow down to the most likely answer. It's a tool that we use to make the best possible prediction based on what we know whether we're making simple predictions like in a linear regression or more complex predictions ones in advanced machine learning models.

When building predictive models, one of the first challenges we encounter is finding the best set of parameters to explain the relationship between the inputs and outputs. A powerful way of tackling this problem is through the Maximum Likelihood Estimator (MLE), an approach for estimating the parameters of a model. MLE focuses on identifying the parameters that maximize the likelihood of the observed data under the model, effectively providing us with the parameters that make our data most probable. We will begin by using it in the simplest form of regression: linear regression. By applying MLE to Ordinary Least Squares, we will establish a foundation that allows us to explain more complex forms of regression. When simple linear regression isn't enough to predict our response variables, we must turn to Ridge and Lasso regression—both forms of regularization for linear regression. For these cases, MLE is adapted to incorporate constraints and penalties, helping address problems like overfitting or multicollinearity.

Whether dealing with a simple linear regression problem or navigating high-dimensional data where regularization is key, understanding how MLE works across different regression models is essential. By the end of this post, you'll have a better understanding of how MLE guides the estimation process in these models and why it is a central statistical modeling in machine learning.

2 Maximum Likelihood Estimator (MLE)

To further understand what MLE is, we will start by breaking down its key components. The term "Maximum" refers to the highest value or peak of a function. In MLE, we are looking for the parameter values that maximize the likelihood function. In other words, we are looking for parameters that make the observed data most probable.

Next, "Likelihood" refers to how likely it is that a given set of parameters explains the observed data. In the context of MLE, the likelihood function represents the probability of observing the given data based on the model's parameters. Although, likelihood is not the same as probability; rather it is a function of the parameters and not the data itself. The higher the likelihood, the better the parameters explain the observed data, which is why we want to maximize the likelihood function.

The term "Estimator" refers to a rule or method for calculating an estimate of a parameter based on observed data. In MLE, the estimator is the formula or approach that we use to calculate the best parameter values that will maximize the likelihood function. This process provides an estimate, or approximation, of the true underlying parameters of the model based on the data. Remember, the goal is to estimate something that we don't already know, so MLE will give us our best guess for understanding our data.

We can now introduce some mathematical notation. If we let θ represent the parameters of our model, and X represent the data, the likelihood function is defined as $L(\theta|X)$. We want to find the parameter values θ that maximize the function, therefore finding the parameters that make the observed data X most probable. We can do this by solving the following:

$$\hat{\theta} = \arg\max_{\theta} L(\theta|X)$$

Where $\hat{\theta}$ represents the parameter values that maximize the likelihood. It is very common to work with the log-likelihood instead of the likelihood, since it simplifies the calculations. There are benefits when using the log-likelihood function. It's maximum occurs at the same point the likelihood function occurs. The logarithmic function allows us to convert exponents into factors and products into sums, further simplifying our complex functions. To compute the maximum in our MLE, it is easy to take the derivative of a log function as opposed to our original likelihood function, which we will see later on. So now we have the following:

$$\hat{\theta} = \arg\max_{\theta} \log L(\theta|X)$$

To see Maximum Likelihood Estimators in action, we will apply them to linear regression, ridge regression, and lasso regression. It is important to note that in regression models, MLE is closely related to minimizing errors. For example, in linear regression, the MLE solution coincides with minimizing the sum of squared residuals, which is the basis for Ordinary Least Squares (OLS). For ridge and lasso regression, MLE is adapted to account for penalties and/or

constraints to ensure that the model fits the data well without overfitting or multicollinearity issues.

3 MLE in Ridge Regression

When applying Ridge Regression to Multiple Linear Regression (MLR), the MLE is modified to include a regularization term that penalizes the size of the coefficients. This regularization helps prevent overfitting, especially when the predictor variables are highly correlated (multicollinearity) or when there are more predictors than observations. Without ridge regression, the models tend to produce models with large, unstable coefficients that fit the training data well but generalize poorly to new data. By introducing a penalty on the size of the coefficients, ridge regression shrinks them towards zero, making the model more robust. This regularization creates a balance between fitting the data well and also keeping the model simple, improving its performance on unseen data. The multiple regression optimization problem

$$\min_{\beta} \|Y - X\beta\|^2$$

becomes

$$\min_{\beta} \left(\|Y - X\beta\|^2 + \lambda\|\beta\|^2\right)$$

Where:

- Y is the $n \times 1$ vector of observed values (dependent variable).
- X is the $n \times (p+1)$ matrix of input features (independent variables), where n is the number of observations and p is the number of predictors. The first column of X is typically all ones, corresponding to the intercept β_0.
- β is the $(p + 1) \times 1$ vector of coefficients (parameters) that we want to estimate, including the intercept β_0.
- $\| \cdot \|^2$ is the squared Euclidean norm.

The second term $\lambda\|\beta\|^2$ is the regularization term that penalizes large values of β. This λ term acts as a trade-off between minimizing the residual sum of squares and controlling the magnitude of the coefficients. The ridge regression optimization problem combines the least squares loss with the regularization penalty:

$$\min_{\beta} \left[(Y - X\beta)^T(Y - X\beta) + \lambda\beta^T\beta\right]$$

Expanding this:

$$\min_{\beta} \left[Y^TY - 2\beta^TX^TY + \beta^TX^TX\beta + \lambda\beta^T\beta\right]$$

We have to take the derivative with respect to β and set the derivative to zero. We can differentiate each term individually:

1. $\frac{\partial}{\partial \beta} Y^T Y$: This is a constant with respect to β, so its derivative is zero.
2. $\frac{\partial}{\partial \beta}(-2\beta^T X^T Y)$: The derivative of this term with respect to β is: $-2X^T Y$
3. $\frac{\partial}{\partial \beta}\beta^T X^T X\beta$: This is a quadratic form, and the derivative with respect to β is: $2X^T X\beta$
4. $\frac{\partial}{\partial \beta}\lambda\beta^T\beta$: This is the regularization term, and the derivative with respect to β is: $2\lambda\beta$

Combining the results and setting the derivative to zero:

$$-2X^T Y + 2X^T X\beta + 2\lambda\beta = 0$$

Solving for β by simplifying:

$$X^T Y + X^T X\beta + \lambda\beta = 0$$

$$X^T X\beta + \lambda\beta = X^T Y$$

$$(X^T X + \lambda I)\beta = X^T Y$$

$$\hat{\beta} = (X^T X + \lambda I)^{-1} X^T Y$$

The I is the identity matrix of size $(p+1) \times (p+1)$, and λI is the regularization term. Ridge regression provides a powerful extension to linear regression, particularly when dealing with issues with multicollinearity or high-dimensional datasets. By incorporating a regularization term into the MLE framework, it balances model complexity and predictive accuracy. The MLE approach allows us to estimate parameters that maximize the likelihood of the observed data while also shrinking the coefficients to avoid overfitting. This creates a more stable and more interpretable model that we can generalize to new data, making the model more robust.