Least Squares and Error Distributions

Dror Baron

This supplement motivates the least squares approach by discussing the implied error distribution, and connects to minimum description length (MDL).

Least squares: Consider a linear model, $\widehat{Y} = X^T\widehat{\beta}$, where X are features, Y are outcomes, $\widehat{Y}$ are predicted outcomes based on our linear model, and $\widehat{\beta}$ are weights that seem to fit the data well. We could also interpret this model as a function, $\widehat{Y} = f(X) = X^T\widehat{\beta}$, which emphasizes that we are predicting outcomes from features.

Gauss proposed a *least squares* approach to determine $\widehat{\beta}$. His approach is to minimize the sum of square errors,

$$\text{Error}(\beta) = \sum_{n=1}^{N}(y_n - x_n^T\beta)^2, \tag{1}$$

where x_n is column n of the features matrix X, and N is the number of outcomes in Y. To minimize the error term Error(β) (1), we note that summing over squared entries of a vector can be obtained by computing the inner product between a transposed row vector version of our vector and itself,

$$\text{Error}(\beta) = (Y - X^T\beta)^T(Y - X^T\beta).$$

To minimize this term, the derivative (more accurately the gradient) with respect to β must be zero,

$$X(Y - X^T\beta) = 0.$$

This should hold for the optimal vector $\widehat{\beta}$, meaning that $XY = XX^T\widehat{\beta}$. We can multiply each side of this latter expression by $(XX^T)^{-1}$, yielding

$$(XX^T)^{-1}XY = (XX^T)^{-1}XX^T\widehat{\beta} = \widehat{\beta},$$

which is a closed form expression for $\widehat{\beta}$.

Why it makes sense: While the above derivation is not too complicated, it relies on the assumption that minimizing Error(β) (1) is sensible. To see why this is the case, it helps to realize that Gauss was implicitly assuming that the prediction errors,

$$\text{Error}(\beta, n) = y_n - x_n^T\beta,$$

are independent and identically distributed (i.i.d.), and obey a zero mean Gaussian distribution.

More formally, a zero mean Gaussian distribution is a probability distribution function (pdf) such that $\text{Error}(\beta, n)$ satisfies

$$f(\text{Error}(\beta, n)) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(\text{Error}_{(\beta,n))^2/2\sigma^2}}, \tag{2}$$

where σ^2 is the variance of the random variable. We can simplify this pdf (2) as $f(\text{Error}(\beta, n)) = c_1 e^{-c_2(\text{Error}_{(\beta,n))^2}}$, where c_1 and c_2 are constants related to σ. Next, express the log likelihood,

$$\log(f(\text{Error}(\beta, n))) = c_3 + c_4(\text{Error}(\beta, n))^2.$$

Finally, because we assume that the various error terms are all independent, the joint pdf of all of them is the product of individual pdf's, and so

$$\begin{aligned} \log(f(\text{Error}(\beta, n = 1), \ldots, \text{Error}(\beta, n = N))) &= \sum_{n=1}^{N} [c_3 + c_4(\text{Error}(\beta, n))^2] \\ &= N c_3 + c_4 \sum_{n=1}^{N} (\text{Error}(\beta, n))^2 .. \end{aligned} \tag{3}$$

The reader can see that this expression (3) involves the sum of least squares, which appears in our definition of $\text{Error}(\beta)$ (1). Minimizing $\text{Error}(\beta)$ corresponds to maximizing the log of the joint likelihood (3), meaning that we choose $\widehat{\beta}$ such that the residual errors have the greatest probability. This can be interpreted as a maximum likelihood approach to parameter estimation, where β are the parameters, and $\widehat{\beta}$ are optimal parameters.

It is important to point out that the assumption that the various error terms are i.i.d. and Gaussian is questionable, but this might be an appropriate approximation in some cases. First, when we have an error composed of a sum of many small errors, the central limit theorem informs us that this error will often have a Gaussian distribution. Second, because this so-called error is actually un-modeled parts of the data due to effects we have not modeled, which may not be present in our features, an independence assumption seems plausible. Indeed, possible strong dependence of the errors would suggest that our model can be easily improved. *The bottom line here is that the assumption of i.i.d. Gaussian error terms results in the least squares formulation being maximum likelihood parameter estimation.*

Connection to MDL: Recall from our past discussions of the minimum description length (MDL) principle that we are adding the coding length for the model with the coding length for the data given the model. Now that we have a Gaussian i.i.d. model for errors, we can encode the data Y using a two part code, where its first part encodes the least squares solution, $\widehat{\beta}$, and the second part encodes the error terms, $\text{Error}(\widehat{\beta}, n), n \in \{1, \ldots, N\}$.

While the Gaussian i.i.d. assumption for error terms may be plausible and well-justified in some applications, in others we may believe that the error terms obey some other distribution. For example, if the error terms ave occasional large values (these are often called outliers), a Laplace distribution has heavier tails that allow outliers to appear with greater probability. In these cases where we have some other distribution for the error terms, least squares in no longer optimal, and we may want to apply a maximum likelihood approach to this modified joint pdf for the errors.