Gaussian Modell Pruning for Linear Regression

Raúl Rojas

February 20, 2015

Abstract

Given a dat set of input-output pairs of the form (x_i, y_i), with n- and one-dimensional entries, respectively, we can compute the vector of coefficients β for a multivariate linear regression. Simpler models are always preferred, since they can usually achieve better performance on new data. Therefore, we would like to "prune" unnecessary coefficients for the vector β. One way of doing this is by looking at the confidence interval for each coefficient.

1 Noisy linear regression

If we are given a data set of n-dimensional data vectors x_i associated with a one-dimensional output y_i, for $i = 1, \ldots, N$, we can compute the vector of coefficients for the linear regression $y = x\beta + \epsilon$, where ϵ represents an error term.

If the y values are sampled with noise (we write them as $\hat{y}$), we obtain from the linear regression a vector $\hat{\beta}$ which is also a noisy version of the optimal linear regression vector β. Since in a linear regression

$$\hat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\hat{y}$$

the covariance matrix of the coefficients is given by $E[(\beta-\hat{\beta})(\beta-\hat{\beta})^T]$. That is

$$\mathrm{Var}(\beta) = E[(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}(y-\hat{y})(y-\hat{y})^{\mathrm{T}}X(X^{\mathrm{T}}X)^{-1}]$$

If the covariance matrix of the y values is $\sigma^2 I$, where I is the identity matrix, the expression above reduces to just

$$B = \mathrm{Var}(\beta) = E[(X^{\mathrm{T}}X)^{-1}\sigma^2].$$

If the matrix X is constant, we just have $B = (X^{\mathrm{T}}X)^{-1}\sigma^2$.

2 Confidence Intervals for the coefficients

Notice that the vector of coefficients has a covariance matrix which not necessarily is an identity, that is, there could be correlations between the individual coefficients. If we want to look at the variance of the k-th coefficient in the $(n+1)\times(n+1)$-dimensional matrix, in the direction of the k-th axis, we project the covariance in the direction u_k, which is a vector full of zeros, except at the k-th entry. The variance of the k-th coefficient along the k-th axis is then $u_k^{\mathrm{T}}Bu_k$, that is, it corresponds to the k-th diagonal entry in the matrix B.

Having an estimation of the variance for each coefficient in $\hat{\beta}$, we can associate confidence bars with every coefficient. We can look at the error bars at two standard deviations, for example, as shown in Fig. 1, where the coefficients have been plotted according to their numerical value (vertical axis). The lines stretch for two standard deviations above and below the value of each coefficient. The Figure shows that the first two coefficients are significantly different from zero (at two standard deviations). The last two coefficients are not significantly different from zero, but β_3 is more “compromised” than β_2. If we decide to prune one coefficient, we would prune β_3. If we prune two coefficients, we would prune both β_2 and β_3. The size of each coefficient relative to its standard deviation provides us with a ranking for keeping or eliminating the coefficient.

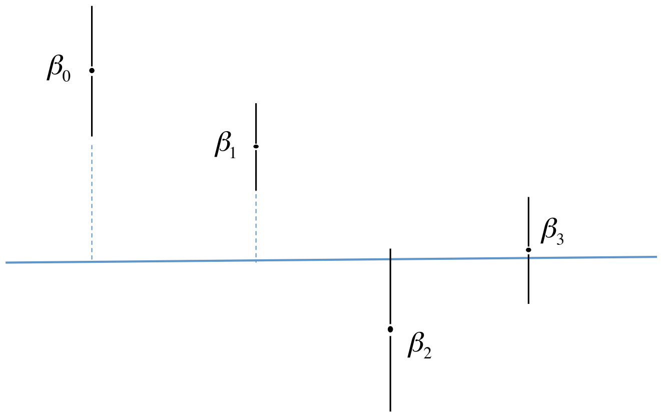

Figure 1: The numerical value of each regression coefficient (vertical axis) compared with two standard deviations of their associated noise

3 Uncorrelated data

One disadvantage of the approach developed in the previous section is that the coefficients can exhibit a certain correlation, that is, the covariance matrix B does not need to be a diagonal matrix. We would like to look at each coefficient independently of each other and that can be done if we first diagonalize the covariance matrix of the x-data. First we can just center the data, subtracting from each vector x_i their mean value μ (as computed from the training data). We can then reduce the matrix X by one column (the column of ones), since we do not need a constant term in the regression function.

Given the new matrix $(X^{\mathrm{T}}X)$ we can find its principal components. Assume that the matrix is of full rank, and that the matrix U contains all the eigenvectors of length one as columns. We can then rotate the data set using the matrix U. The covariance matrix of the linear regression is now

$$B = E[(U^{\mathrm{T}}X^{\mathrm{T}}XU)^{-1}\sigma^2].$$

Since U consists of the eigenvectors of $X^{\mathrm{T}}X$, the result is

$$B = D\sigma^2$$

where D is a diagonal matrix with diagonal elements of the form $1/\lambda$, for $i = 1, \ldots, n$, where λ_i represents the i-th eigenvalue of the matrix $X^{\mathrm{T}}X$. Now the coefficients of β are decorrelated and we can compute their confidence intervals with more precision.

One open question is how to compute the covariance matrix of the original data $\hat{y}$, since we do not know the real values y. One way of doing this is just by taking the covariance matrix of the approximation error of the linear regression, that is of $\hat{y} - x\hat{\beta}$. If the errors are decorrelated, the covariance matrix will be a diagonal matrix.

Notice also that we have used a Gaussian model for the noise in the linear regression coefficients. If the noise in the $\hat{y}$'s is Gaussian, we expect Gaussian noise in any linear transformation of the y values (and the vector $\hat{\beta}$ is computed from a linear transformation of $\hat{y}$).

References

[1] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning*, Springer-Verlag, New York, 2001.