Clustered Linear Regression

Clustered linear regression (CLR) is a new machine learning algorithm that improves the accuracy of classical linear regression by partitioning training space into subspaces. CLR makes some assumptions about the domain and the data set. Firstly, target value is assumed to be a function of feature values. Second assumption is that there are some linear approximations for this function in each subspace. Finally, there are enough training instances to determine subspaces and their linear approximations successfully. Tests indicate that if these approximations hold, CLR outperforms all other well-known machine-learning algorithms. Partitioning may continue until linear approximation fits all the instances in the training set — that generally occurs when the number of instances in the subspace is less than or equal to the number of features plus one. In other case, each new subspace will have a better fitting linear approximation. However, this will cause over fitting and gives less accurate results for the test instances. The stopping situation can be determined as no significant decrease or an increase in relative error. CLR uses a small portion of the training instances to determine the number of subspaces. The necessity of high number of training instances makes this algorithm suitable for data mining applications.

Evaluation of the CLR algorithm

CLR is an algorithm based on linear regression, so most of the advantages and disadvantages of linear regression are carried to CLR. For example, curse of dimensionality is one of the biggest problems of linear regression since it increases the complexity. If the number of features increases, accuracy will decrease because possibility and effects of interaction between features will increase.