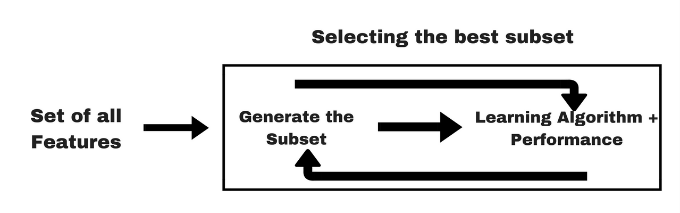
**Suppose you are training a multiple linear regression model to predict a patients' annual hospital expenditure ($) based on 5 health record inputs like BP, BMI etc.**

1. ***What problem is Regularization solving for this business case (what is the value added) and how does it achieve that?***

If we decide to do a regression with the data available to predict the patient’s annual hospital expenditure and optimize the results using such data, we would be making an assumption (whether we are aware of that or not). Such assumption is that the unseen (new upcoming data) will behave the same way as the already seen data (in other words, what we have is all there is). This assumption is (for the most part) invalid, unless we can prove otherwise thru data or statistical analysis. Therefore, it is important to take extra precaution in predicting into the future based on current data (which is probably limited), and work under the more conservative assumption that it is likely that new data points might lead the analysis into a slightly different direction, and this is where “Regularization” comes into play.

The value added for the business by regularization in such regression model is to provide a more conservative model. In other words, to provide a model that might not yield the highest/greatest performance on the training data, but it will potentially improve the performance on unseen data. Regularization is a good way to prevent overfitting, which in this case (prediction of patient’s annual hospital expenditure) if the model is not limited by regularization it might yield to an overestimation of such expenditure and consequently an over extension on their investments and/or expenses.

The way regularization achieve that is by providing a way to penalize the model by adding a small bias (to the training model) to avoid that certain features (that might not be giving a lot of value to the model) take over an lead to a wrong conclusion when expose to new data. Usually there are 2 flavors of such models, Lasso & Ridge and a third one that is the combination of both. Ridge regression is useful in the case of a model that does not have a lot of features and it is desired to keep them in the final model but minimize the effect of such variables that do not contribute a lot to the model. Lasso regularization on the other hand provide a way to eliminate or minimize to zero the contribution of those features that do not contribute to the model and only add complexity to such. For this reason, this model is usually used as a Feature Selection model. The third method is called “Elastic Nets” and in summary is a linear combination of both previous models being modulated by a constant called alpha, which provide the best of both words.



Saurav Kaushik. “Embedded Methods” by Kaushik, Dec 1, 2016. Analytics Vidhya, <https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/>

1. ***How are shrinkage methods different from the other 2 methods in this regard?***

***Please explain your answers (max 1 to 2 pages, with diagrams)***

The other two methods are: wrapping methods & dimension reduction methods. The first provides a way to identify a subset of predictors (features/columns) that are believed to be related to the response (target variable). This can be performed by identifying a particular relational parameter such as: the correlation coefficient or the mutual information coefficient and then selecting those features that provide the greatest values in term of those relational parameters. The second methods project the predictors into a smaller dimensional subspace that offers virtually the most information possible in comparison with the original dataset.

The main difference between the shrinkage method (embedded method) and the other two is that the shrinkage method interacts with the prediction method and modify such from the inside, while the other two just provide either a new (smaller) set of number of features or a transformed (projected) also smaller number of features. Another clear difference between the shrinkage method and the dimension reduction method comes when talking about interpretability. The shrinkage methods have the advantage of being highly interpretable and provide a clear way to identify which variables are really relevant to the model.

1. ***Can regularization be applied for classification? How would the math work? (Pl. research this).***

Yes, regularization can be applied for classification problems, in fact is widely used in Neural Network problems.

The shrinkage methods are a general framework of the following shape:

**Minimization of: <loss function> + λ <some measure of model complexity>**

For a classification model the typical loss function could be the hinge loss function. The noted hinge loss is used for “maximum-margin” classification:

*l(y) = max(0, 1 – t \* y)*

As for the measure of the model complexity, that could be restrictions for smoothness (the number of continuous derivatives over some domain) and vector space norm.