# AIC? BIC? WTF?

No one gets the perfect model on the first try. Model selection and tuning is an iterative process. But after you have tried every model and combination of hyperparameters, how do you know which one is the “best” model? Measures such as R$^{2}$ and root mean squared error work well for regression tasks, while accuracy and F$\_{1}$ scores are good starts for classification problems. However, all of those measures simply assess how well a model has learned a particular sample of data. Unless compared to a test set of data, those measures give no insight into overfitting or unjustified complexity of the model.

If you are a fan of Bayesian statistics, two measures to consider while answering the “which model is best?” question are the Akaike Information Criterion (AIK) and the Bayesian Information Criterion (BIC). Both of these measures are based on information theory and give a means of estimating a model’s quality in terms of information loss and complexity. In order to understand AIC and BIC fully, we need to take a minute to understand what the modeling process is trying to accomplish and how these two measures fit into that process.

## Purpose of Modeling

The goal of the modeling process is to build an equation that describes the underlying behaviors and trends of a population based on a sample from that population. Even in simplistic terms, it is clear that is a tall order. No model is going to be able to describe an entire population perfectly from a sample. Even if that sample is a good representation of the population, it is likely that small details present in the population are missing from the sample. For this reason, it is better to develop a model that generalizes to the population rather than memorizes the sample. In this way, the model may perform slightly worse when evaluated on the sample used to develop it, but will have a much better performance on a sample it has never seen than if it had just memorized the original sample.

The instance of “memorizing a sample” that was presented above is termed overfitting. It speaks to the power of machine learning/deep learning and their capability to learn a task perfectly (i.e. achieve 100% accuracy) given enough training and leeway with parameters. However, as mentioned in the previous paragraph, this power needs to be tamed in order for the model to be able to generalize to the population. The main strategy to decrease overfitting is to reduce the complexity of the model by restricting how many parameters it can incorporate. This too, though, comes with a trade-off as reducing model complexity leads to a loss of the ability to describe certain aspects of the sample’s underlying behavior and trends. In statistics, the loss of the ability to model certain details about the population is often called “information loss.”

Balancing the how well a model “fits” the sample with its complexity is exactly what the AIC and BIC measures are trying to estimate. They reward a model for how well it describes the underlying trends of a sample by assessing the maximum likelihood function\* of the model and the statistical characteristics of the sample used to develop it. Furthermore, they penalize a model based on the number of parameters it includes. The greater the number of parameters, the larger the value of the penalty. Thus, assessing model quality based on AIC or BIC provides a means of comparing both the goodness of fit and potential of overfitting.

## Akaike Information Criterion

$$2\_k\_ - 2\_ln\_(\_ML\_)$$

The equation for the AIC is shown above. In the equation, \_k\_ represents the number of parameters included in the model and \_ML\_ is the maximum of the likelihood function. We can see that the formula penalizes a model with a value twice the number of its parameters.

## Bayesian Information Criterion

$$ln(x)\_k\_ - 2\_ln\_(\_ML\_)$$

The equation for the BIC is shown above. In the equation, \_k\_ still represents the number of parameters included in the model and \_ML\_ still is the maximum of the likelihood function. However, x is equal to the number of observations in the sample. In this case, as long as there are more than 7 observations in the sample ($ln(8)$ = 2.04) the penalty term is larger than that of the AIC.

## Practical Usage

The AIC and BIC measures are used to compare between a set of models. When using these measures, the best model is the one with the lowest score (negative or positive). From the two formulas displayed above, we can see that a low score can be achieved by either having a small penalty term or a large likelihood term. However, it is important to note that AIC and BIC scores do not actually give a measure of overall model performance. They strictly allow for the selection of the best model from a set of models. If all of the models are poor, the AIC/BIC score will simply indicate which model is the least bad. Thus, it is important to assess the overall performance of the model selected from set with other evaluation metrics like the ones mentioned at the beginning of this discussion.

## Conclusion

Comparison of AIC and BIC scores for a set of candidate models is an efficient method for assessing quality, and thus providing justification for which one model is the ”best.” These measures evaluate the models based on how well they estimate the statistical characteristics of the sample observations and penalize the model based on the number of parameters it includes. When using these measures, the lowest score, negative or positive, indicates the best model. However, AIC and BIC scores are not a measure of overall model performance. Thus, the model selected based on AIC/BIC scores should be vetted further with appropriate evaluation metrics to determine the overall quality of the model.

\* Note: The concept of the likelihood function and its maximum are difficult to grasp within the context of this discussion and is deserving of its own blog post. However, to try to aid in the comprehension of this discussion, it is safe to define the maximum of the likelihood function as selecting the parameters for the model that maximize the model’s ability to reproduce the observed data. For example, if the observed data were from a normal distribution, the maximum of the likelihood function would represent the mean and standard deviation of the observed sample distribution. For a more in-depth explanation of maximizing the likelihood function, check out this discussion (<https://towardsdatascience.com/probability-concepts-explained-maximum-likelihood-estimation-c7b4342fdbb1>)