Probabilistic Model Selection with AIC/BIC in Python

Shachi Kaul · Follow

Published in Analytics Vidhya

6 min read · Jun 1, 2020

Blog Milestones

Dear learning souls..sit in a comfortable posture, set your focus, and let's kick-off this dilemma of selecting your best machine learning model.

Presenting a secret of how to choose the best model. Model selection plays a very vital role in building a machine learning model. There can be multiple eligible algorithmic models, treated as candidate models but only one with optimized parameters can be chosen as best performed robust model. The selection of best from candidates is what we call a Model Selection.

Brief about Model Selection

Model Selection is like choosing either a model with different hyper-parameters or best among different candidate models. Normally, the selection of any model shouldn't rely only on its performance but instead, also on its complexity.

Usually, we categorize the techniques of model selection as follows:

- Random Train/Test Split
- Resampling Techniques
- Probabilistic Model Selection Techniques

This blog will only discuss the probabilistic model selection in depth because random_train/test and resampling techniques are covered in Model Selection in ML/AI with Python and Deeply Explained Cross-Validation in ML/AI.

> *The best-fit model according to AIC is the one that explains the greatest amount of variation using the fewest possible independent variables.*

Probabilistic model selection

- The best model is chosen with the help of **probability framework of log-likelihood under Maximum Likelihood Estimation.**
- The quality of statistical methods can be measured by **Information Criteria** (IC) with some score. So, it refers to model selection methods **based on likelihood functions**. Lowest the score, best the model. This has come from the Information Theory of Statistics.
- Takes into account of **model performance and complexity** while another model selection technique of resampling checks only model performance.
- Model is chosen by a scoring method where scores are based on:
  - **Performance** on train data is evaluated using **log-likelihood** which comes from the concept of MLE so as to optimize model parameters. It says about how well your model is fitted with your data. It provides an indication of total error.
  - **Model Complexity** is evaluated using number of parameters (or degrees of freedom) in model.

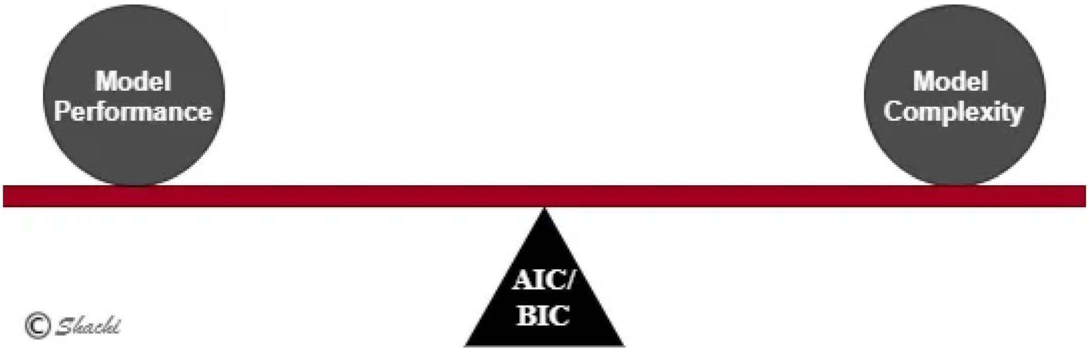
> *Score rewards models that achieve high goodness-of-fit and penalize them if they become over-complex*

- Common probabilistic methods are:
  ~ AIC (Akaike Information Criterion) from *frequentist probability*
  ~ BIC(Bayesian Information Criterion) from *bayesian probability*

Let's know more about AIC and BIC techniques.

What are AIC/BIC criteria

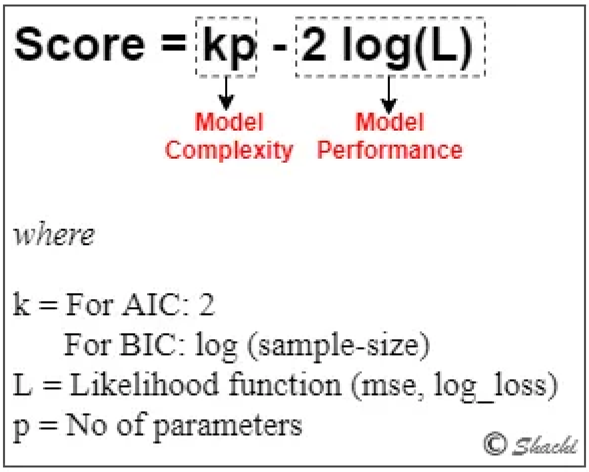
- These are IC methods coming from the field of **frequentist and bayesian probability**. Any selection method scoring lowest means less information is lost and hence the best model. This is a crux of information theory.
- **Calculated using *Log-likelihood:*** includes *mean squared error (regression)* and log_loss such as cross_entropy (classification).
- **Penalize parameters to combat over-fitting:** It is advised to maximize the likelihood by adding more parameters which may end up in making model more complex and over-fitted. Thus, AIC/BIC add a penalty for additional parameters. That's how it's maintaining balance.
  Concluded as keep a balance between fitting data (log-likelihood) and model complexity(penalty to estimate model parameters for the sample).



- Estimates the quality of model among candidate models.

> *How well your model fits data without over-fitting it*

- Formula is a form of penalized likelihood (penalty term+negative likelihood)



$$\text{Score} = \underbrace{kp}_{\text{Model Complexity}} - \underbrace{2\log(L)}_{\text{Model Performance}}$$

where

k = For AIC: 2
For BIC: log (sample-size)
L = Likelihood function (mse, log_loss)
p = No of parameters

© Shachi

Also, let’s clear a common confusion.
Eligible for both linear and non-linear models?
Answer: Yes, since AIC/BIC are based on log-likelihood function for a model which you can have for both linear and non-linear models.

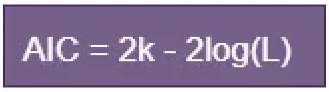
AIC (aka Akaike information criterion)

- **Birth of AIC**

Simple Answer. Information Theory in Statistics.

Any model (let’s say linear regression) doesn’t exhibit whole truth of study, it’s all about approximate. We accept that there is always some loss of information. Now what? We have to select best model which is closest to the reality of truth or should say minimize the loss of information. Kullback and Leibler coined KL information which is a measure of information loss. Later, Japanese statistician, Hirotugu Akaike addressed the relation between maximum likelihood and KL information. He developed IC to estimate KL information, termed as Akaike’s Information Criterion (AIC). Thus, K-L distance is a measure of information lost in terms of distance, or discrepancy between two models.

- *Formula:*



$$AIC = 2k - 2\log(L)$$

k= number of independent variables to build model
L= maximum likelihood estimate of model

For maximizing likelihood log(L), more variables supposed to be added in model, leads to overfitting. Hence, “2k” penalty term introduced which doesn’t remove overfitting completely. Because of weak penalty included. Also, formula doesn’t take observations into account instead of only model parameters.

- In case of **small samples**, most chances of having over-fitting as AIC will end up selecting many parameters. Thus, **AIC corrected** was introduced to address this issue. For small sample size,

$$AICc = AIC + \frac{2k(k+1)}{n-k-1}$$

Smaller the value, lesser information be lost and best model fit.

- **Information loss**(Δi) can be measured as when use fitted model (gi) rather than best model (gmin):

$$\Delta_i = AIC_i - AIC_{min}$$

BIC (aka Schwarz information criterion)

Before jumping with the concept, one obvious question pops in my mind.
“Why is BIC called bayesian?”

Most of the references quoted below.
Though “bayesian” is included in its name but no prior information needed. BIC computation doesn’t require Bayesian knowledge. It is only derived in framework of Bayesian theory to maximize posterior probability of model.
Since it ignores prior distribution, the new IC method came called prior based BIC (PBIC). Refer to this paper for more details.

- BIC comes under **parsimonious models** that explain good enough even with minimum no of parameters.
- Formula:

$$BIC = \log(n)\,k - 2\log(L)$$

k= number of independent variables to build model
L= maximum likelihood estimate of model
n = sample size (#observations)
log-base = e(natural log)

For maximizing likelihood log(L), more variables supposed to be added in model, lead to over-fitting. BIC tackles this issue by including a strong penalty of "log(n)k" which instead may fall you in under-fitting i.e. too simple models. Simple models aren't capable of catching variations in data.

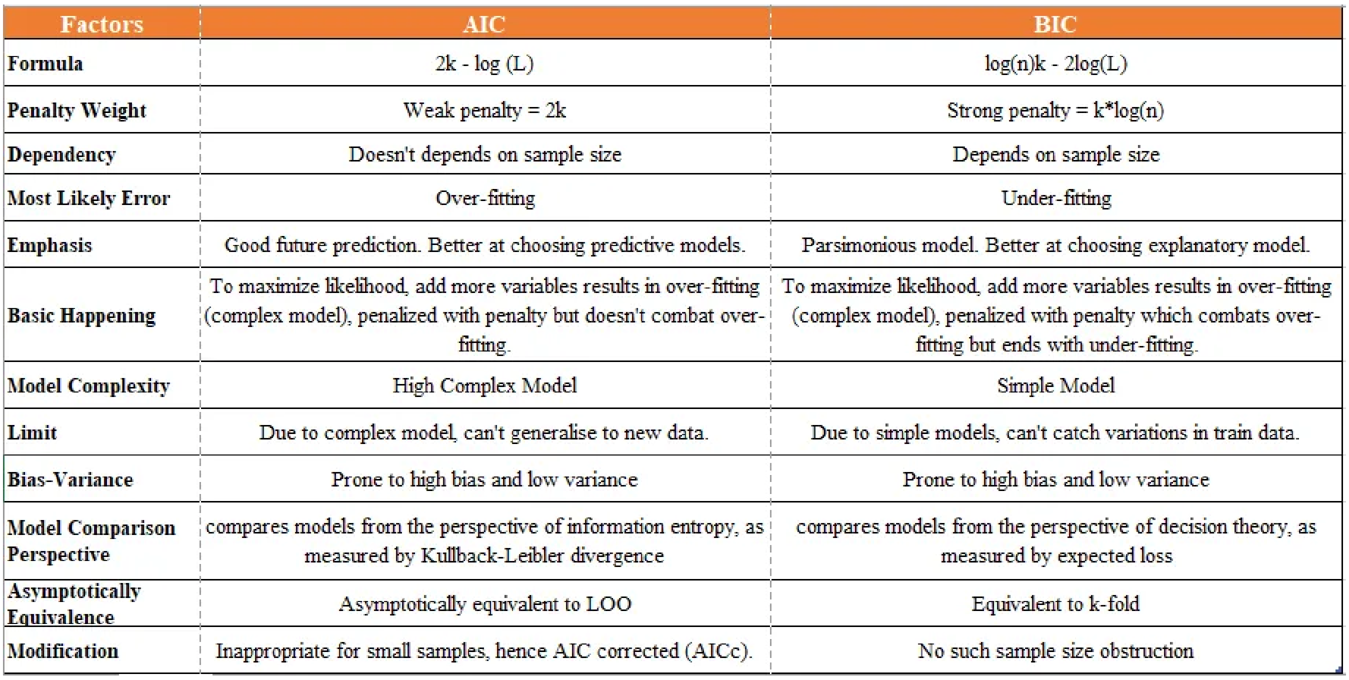
- Interesting stuff now…
  From the paper, during model selection, we can see how much suspended candidate models differ from the best-chosen model.
  Δ = BIC(M1|D) − BIC(M2|D)
  If Δ is positive, then M2 is better than M1 but how much better?

$$|\Delta| = \begin{cases} 0-1 & \text{insignificant} \\ 1-3 & \text{meaningful} \\ 3-5 & \text{strong} \\ 5+ & \text{very strong} \end{cases}$$

Source

Quick Analogy

Here's quick-glance learning of AIC/BIC instead of lazy reading the content.



Factors	AIC	BIC
Formula	2k - log (L)	log(n)k - 2log(L)
Penalty Weight	Weak penalty = 2k	Strong penalty = k*log(n)
Dependency	Doesn't depends on sample size	Depends on sample size
Most Likely Error	Over-fitting	Under-fitting
Emphasis	Good future prediction. Better at choosing predictive models.	Parsimonious model. Better at choosing explanatory model.
Basic Happening	To maximize likelihood, add more variables results in over-fitting (complex model), penalized with penalty but doesn't combat over-fitting.	To maximize likelihood, add more variables results in over-fitting (complex model), penalized with penalty which combats over-fitting but ends with under-fitting.
Model Complexity	High Complex Model	Simple Model
Limit	Due to complex model, can't generalise to new data.	Due to simple models, can't catch variations in train data.
Bias-Variance	Prone to high bias and low variance	Prone to high bias and low variance
Model Comparison Perspective	compares models from the perspective of information entropy, as measured by Kullback-Leibler divergence	compares models from the perspective of decision theory, as measured by expected loss
Asymptotically Equivalence	Asymptotically equivalent to LOO	Equivalent to k-fold
Modification	Inappropriate for small samples, hence AIC corrected (AICc).	No such sample size obstruction

© Shachi

Applications

- **Feature Selection**: Compare the models with adding/removing features with scores
- **Regularization parameter**: AIC/BIC select this parameter in Ridge/Lasso models

Implementation

AIC and BIC techniques can be implemented in either of the following ways:

- **statsmodel library:** In Python, a statistical library, **statsmodels.formula.api** provides a direct approach to compute aic/bic.
- **scikit-learn**: Sklearn library also provides the AIC/BIC score with LassoLarsIC estimator which limits only linear models. hence, it's not of much use when it comes to non-linear models.
- **Manual Computation**: It's best to compute these scores by implementing their formulae directly.

Let's switch over to Code Implementation of AIC/BIC in Python to get hands-on.

References

- AIC related explanation: http://www.cef-cfr.ca/uploads/reference/mazerolle_2006.pdf
- Information Theory: https://arxiv.org/pdf/1511.00860.pdf
- Strongly recommendation http://www.cs.toronto.edu/~mbrubake/teaching/C11/Handouts/BIC.pd
- https://peerj.com/preprints/1103.pdf

You are free to follow this author if you liked the blog because this author assures to back again with more interesting ML/AI techs.
Thanks,
Happy Reading! :)

Can get in touch via LinkedIn.