Hiring Manager Session

Tamojit Maiti
Masters in Applied Statistics and Operations Research, ISI Kolkata
Data Scientist at Sixt R&D, previously at Rapido & AB InBev

Agenda

- Linear Regression
- Logistic Regression
- Decision Trees
- Random Forest
- Clustering
- Model Selection, Regularization & Feature Engineering

Appendix

Model Information Criteria

- Given a dataset and a family of parametric models that is the output of a supervised learning approach, the information criteria tells us which is the best model
- It does NOT tell us which is the universally best model of all family of models
- Akaike Information Criteria

$$AIC = -\frac{2}{N}LL + \frac{p}{N}$$

- Bayesian Information Criteria

$$BIC = -2\,LL + p\log N$$

- AIC picks up more complex models than BIC (Why?)

Binary Cross Entropy

- Loss function for logistic regression

$$BCE = -\frac{1}{N}\sum(y_i \log p_i + (1 - y_i) \log(1 - p_i))$$

Measures of Purity

- Gini Impurity

$$GI = 1 - \sum p_i^2$$

- Entropy

$$\mathrm{En} = -\sum p_i \log p_i$$

- Information Gain

$$\mathrm{IG}(Y) = \mathrm{En}(X) - \mathrm{En}(X, Y)$$