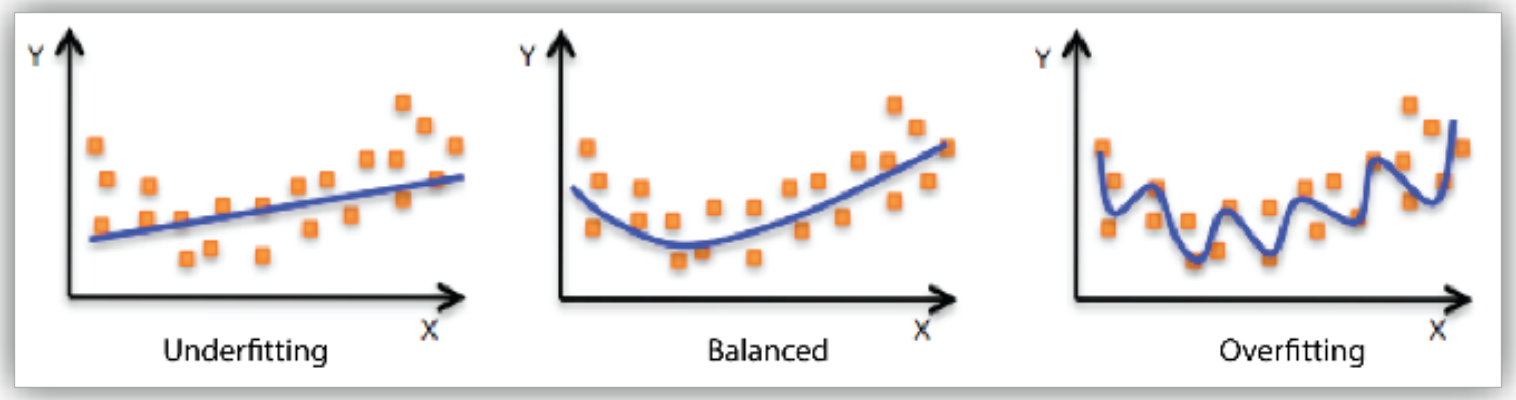
METHODS OF MODEL SELECTION

By Madeleine Armer

TODAYS LESSON:

- How do we compare models?
- What are these acronyms:
  - AIC
  - BIC
  - MDL
- A live performance!

THE NEXT TOP MODEL – HOW ARE WE FIT TO JUDGE?



- How do you check how well your model is doing?
  - R^2 Values
  - Cross Validation with Loss Function or Prediction accuracy
- What if you do not have out of sample data?
  - … are there other ways you can compare your models?

https://docs.aws.amazon.com/machine-learning/latest/dg/images/mlconcepts_image5.png
https://statisticsbyjim.com/regression/r-squared-invalid-nonlinear-regression/

AIC - AKAIKE INFORMATION CRITERION

$$AIC = -2\ln(L(\theta)) + 2k$$

- $\ln L(\theta)$ is loglikelihood
- k is number of parameters included in the model
- Penalizes adding parameters
- Keep it on the LOW
- R application – does model selection for you!

https://machinelearningmastery.com/probabilistic-model-selection-measures/
https://www.scribbr.com/statistics/akaike-information-criterion/

BIC – BAYESIAN INFORMATION CRITERION

$$BIC = k\ ln(\mathrm{n}) - 2\ln(L(\theta))$$

- where $\ln L(\theta)$ is loglikelihood
- k is number of parameters included in the model
- n is sample size
- Similar application to AIC
  - Do not use with smaller sample sizes

https://machinelearningmastery.com/probabilistic-model-selection-measures/

MDL – MINIMUM DESCRIPTION LENGTH

$$MDL = L(h) + L(D \mid h)$$

- Goal is shortest description for the information found in the observed data
- cost in of the model VS the cost of the data given the model
- Normal Maximum Likelihood (NML) and Fischer's Information Approx. (FIA)
- Machine learning, data mining and other black box modeling

Hansen, Mark H., and Bin Yu. "Minimum Description Length Model Selection Criteria for Generalized Linear Models." *Institute of Mathematical Statistics Lecture Notes - Monograph Series*, 2003, pp. 145–163., doi:10.1214/lnms/1215091140.
Zou, Yuan, and Teemu Roos. "On Model Selection, Bayesian Networks, and the Fisher Information Integral." *Advanced Methodologies for Bayesian Networks*, 2015, pp. 122–135., doi:10.1007/978-3-319-28379-1_9.

LIVE PERFORMANCE

- Iris data set in R
- Multiple linear regression to predict Sepal Length based on other flower factors
- Use AIC and BIC to pick the “best” model

https://www.scribbr.com/statistics/akaike-information-criterion/

THANK YOU FOR LISTENING!

Questions??

Preparing people to lead extraordinary lives