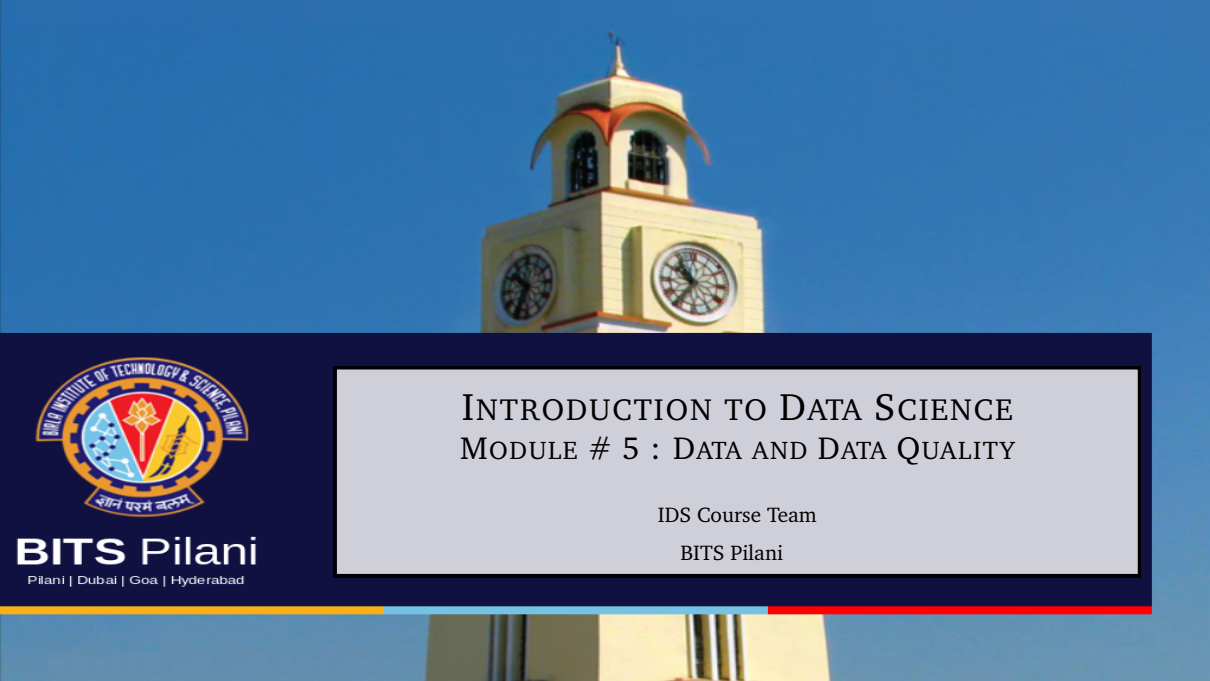


BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Introduction to Data Science
Module # 5 : Data and Data Quality

IDS Course Team

BITS Pilani

The instructor is gratefully acknowledging the authors who made their course materials freely available online.

Table of Contents

Data Models

- **Model is something we construct to help us understand the real world.**
- A statistical model serves two key purposes in a data analysis,
  - quantitative summary of data.
  - impose a specific structure on the population from which the data were sampled.

Data Model - Case Study

- Conduct a survey of 20 people to ask them how much they'd be willing to spend on a product you're developing.
- The survey response

$$25, 20, 15, 5, 30, 7, 5, 10, 12, 40, 30, 30, 10, 25, 10, 20, 10, 10, 25, 5$$

- What do the data say?

Statistical Model

- The first key element of a statistical model is data reduction.
- Take the original set of numbers consisting of your dataset and transform them into a smaller set of numbers.
- The process of data reduction typically ends up with a **statistic**.
- **A statistic is any summary of the data.**
- The sample mean, median, the standard deviation, the maximum, the minimum, and the range are statistic.

Modelling and Evaluation

- Data Model Development and experiment framework setup
  - Data Modelling based on training sets – At its core, a statistical model provides a description of how the world works and how the data were generated.
  - Framework to feed in new data and test the models and change training data and retrain model based on new data sets as sliding window.
  - 3 main tasks involved
    - ★ Feature Engineering: Create data features from the raw data to facilitate model training.
    - ★ Model Training: Find the model that answers the question most accurately by comparing their success metrics.
    - ★ Determine if your model is suitable for production.
- Data Model Evaluation and KPI Checks
  - Read papers, research material to finalize the algorithmic approaches.

Developing a Benchmark Model

- The goal is to develop a benchmark model that serves us as a baseline, upon we'll measure the performance of a better and more attuned algorithm.
- Benchmarking requires experiments to be comparable, measurable, and reproducible.
- Models
  - Null Model
  - Bayes rate model
  - Normal Model

Null Model

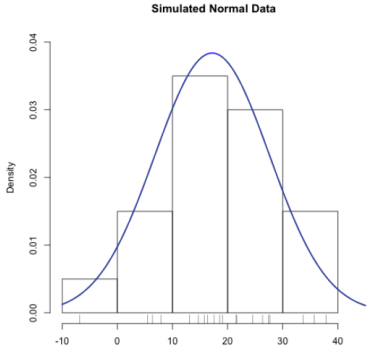
- A null model is the best model of a very simple form you are trying to outperform.
- Two types
  - Model that is a single constant (returns the same answer for all situations)
  - Model that is independent (doesn't record any important relation or interaction between inputs and outputs).
- We use null models to **lower-bound desired performance**, so we usually compare to a best null model.

Bayes Rate Model

- The Bayes Optimal Classifier is a probabilistic model that makes the most probable prediction for a new example.
- A Bayes rate model (also called a saturated model) is a best possible model given the data at hand.
- Bayes rate model is the perfect model and it only makes mistakes when there are multiple examples with the exact same set of known facts (same Xs) but different outcomes (different Ys).
- It isn't always practical to construct the Bayes rate model, but we invoke it as an **upper bound** on a model evaluation score.

Normal Model

- Normal model says that the randomness in a set of data can be explained by the Normal distribution, or a bell-shaped curve.
- The Normal distribution is fully specified by two parameters – the mean and the standard deviation.



Models as Expectations

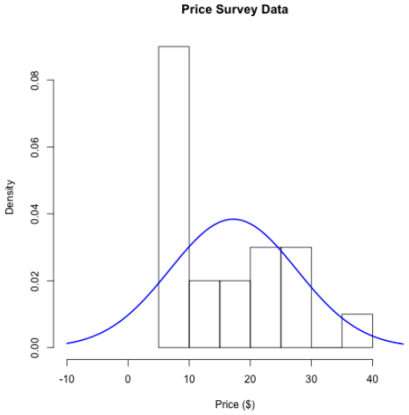
- A statistical model must impose some structure on the data.
- **A statistical model provides a description of how the world works and how the data were generated.**
- A statistical model allows for some randomness in generating the data.

Normal Model

- Most popular statistical model
- The randomness in a set of data can be explained by the Normal distribution, or a bell-shaped curve.
- The Normal distribution is fully specified by two parameters-the mean and the standard deviation.
- Use the Normal distribution to setup the shape of the distribution that we expect the data to follow.

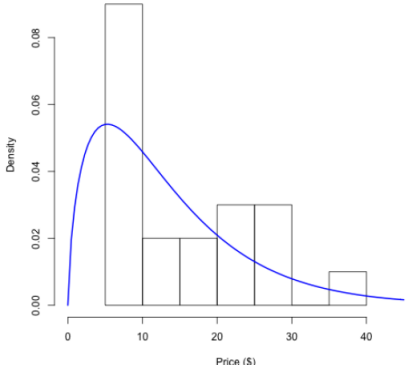
Comparing Model Expectations with Reality

- Given the parameters, our expectation under the Normal model is that the distribution of prices that people are willing to pay looks like a bell-shaped curve.
- E.g. Normal curve on top of the histogram of the 20 data points of the amount people say they are willing to pay. The histogram has a large spike around 10.
- Normal distribution allows for negative values on the left-hand side of the plot, but there are no data points in that region of the plot.



Refining Our Expectations

- When the model and the data don't match very well.
  - Get a different model.
  - Get different data.
  - Do both.
- E.g. Choose a different statistical model to represent the population, the Gamma distribution, which has the feature that it only allows positive.



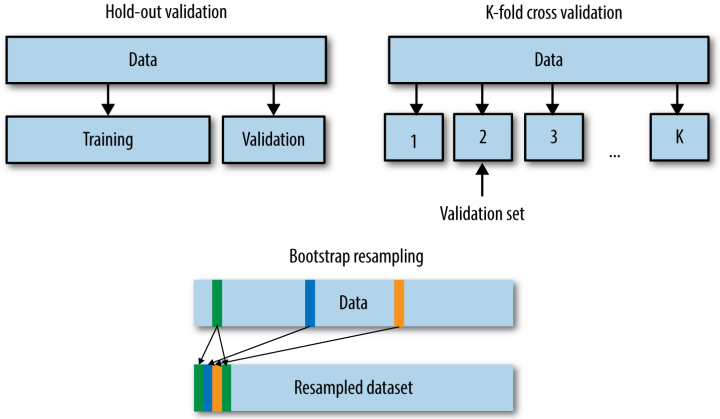
Model Evaluation Metrics

- Performance Metrics vary based on type of models i.e. Classification Models, Clustering Models, Regression Models.
  - Regression Models
    - Root mean squared error (RMSE)
  - Classification Models
    - Confusion Matrix
    - Precision
    - Recall
    - F1-score
  - Clustering Models
    - BCubed Precision
    - BCubed Recall
    - Silhouette Coefficient
    - F-score

Cross Validation Techniques

- Exhaustive
  - Leave p-out
  - Leave 1-out
- Non-Exhaustive
  - K-fold
  - Holdout
  - Repeated random sampling

Cross Validation Techniques

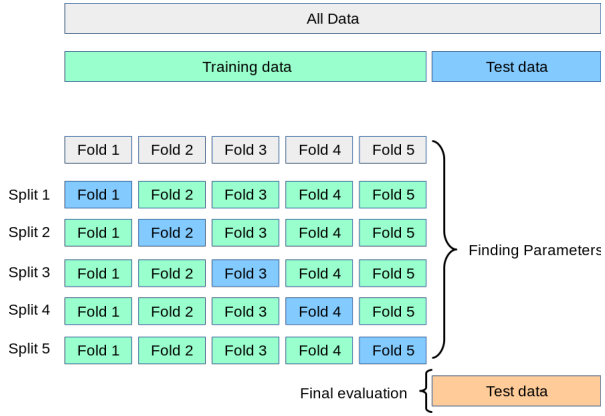


https://moredvikas.wordpress.com/2018/10/10/machine-learning-model-validation-techniques/

Iterated K-Fold Validation with Shuffling

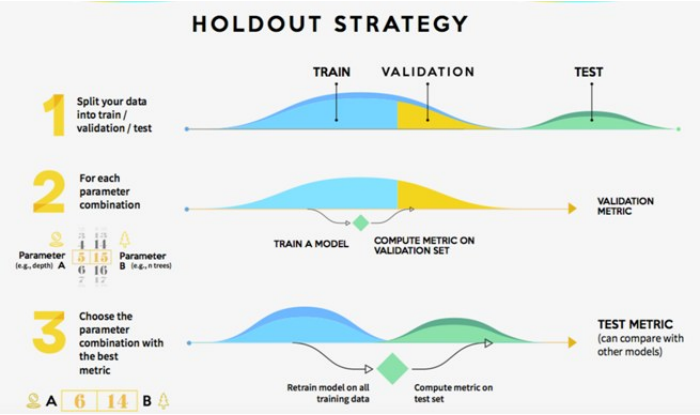
- It consist on applying K-Fold validation several times and shuffling the data every time before splitting it into K partitions.
- The final score is the average of the scores obtained at the end of each run of K-Fold validation.
- This method can be very computationally expensive, as the number of trained and evaluating models would be $I \times K$ times; I the number of iterations and K the number of partitions.

K-Fold Cross Validation



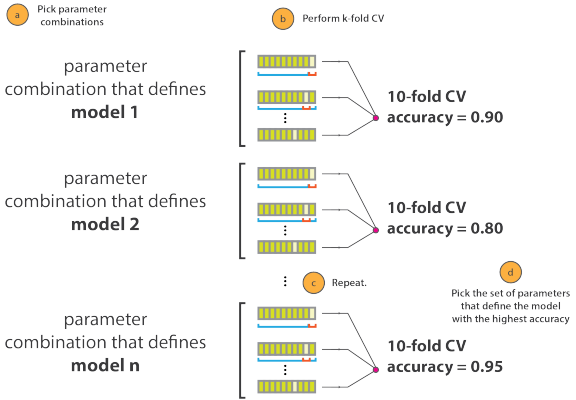
scikit-learn.org

Holdout Strategy

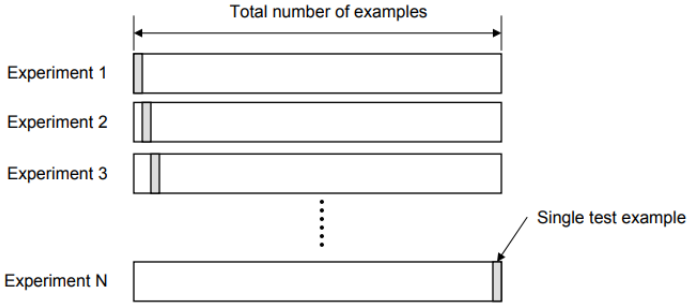


www.kdnuggets.com

Choosing Parameters

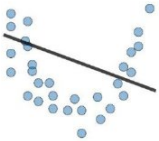
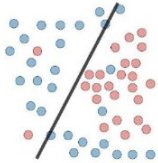
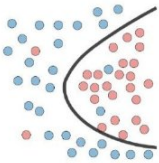
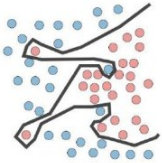


Leave-One-Out Cross-Validation (LOOCV)

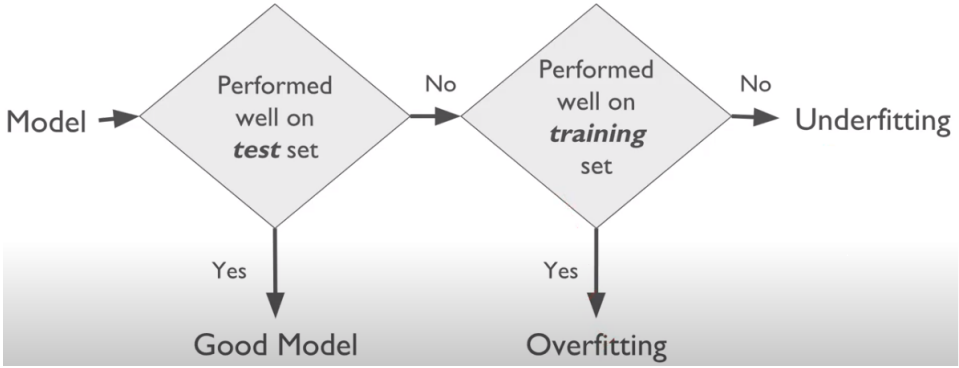


https://moredvikas.wordpress.com/2018/10/10/machine-learning-model-validation-techniques/

Validating Models

	Underfitting	Just right	Overfitting
Symptoms	- High training error - Training error close to test error - High bias	- Training error slightly lower than test error	- Low training error - Training error much lower than test error - High variance
Regression			
Classification			

Validating Models



https://datascience.foundation/sciencewhitepaper/underfitting-and-overfitting-in-machine-learning

Balancing Bias and Variance to Control Errors

- A general rule is that, as a statistical method tries to match datapoints more closely or when a more flexible method is used, the bias reduces, but variance increases.
- In order to minimize the expected test error, we need to select a statistical learning method that simultaneously achieves low variance and low bias.

https://towardsdatascience.com/balancing-bias-and-variance-to-control-errors-in-machine-learning-16ced95724db

- Introduction to Data Mining, by Tan, Steinbach and Vipin Kumar (T3)
- The Art of Data Science by Roger D Peng and Elizabeth Matsui (R1)
- Introducing Data Science by Cielen, Meysman and Ali
- https://www.deltapartnersgroup.com/managing-data-quality-optimize-value-extraction
- http://www.dataintegration.ninja/relationship-between-data-quality-and-master-data-management/

THANK YOU