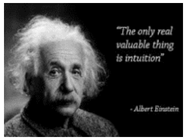
Week 5

(1) Machine Learning
(2) Variable Selection

KOGOD SCHOOL of BUSINESS
AMERICAN UNIVERSITY • WASHINGTON, DC

Machine Learning Overview



Machine Learning: Intuition

- **Machine learning** is a very broad field in computer science that evolved from areas like **pattern recognition**, **computational modeling** and **artificial intelligence**.
- It was originally defined as the ***"field of study that gives computers the ability to learn without being explicitly programmed"*** (Samuel 1959).
- More recently, the field of machine learning has moved closer to **computational statistics**.
- For the purposes of this class, we view machine learning as the **development of models that can learn patterns** from part of the data (i.e., **training**) and evaluated with other parts of the data (i.e., **testing**), to find the most accurate models for decision making.
- Essentially, when we run a **regression model**, we are **"training"** the model with the data. As the data changes, re-running the model with the updated data yields revised model parameters – i.e., the model "learns" from the new data.

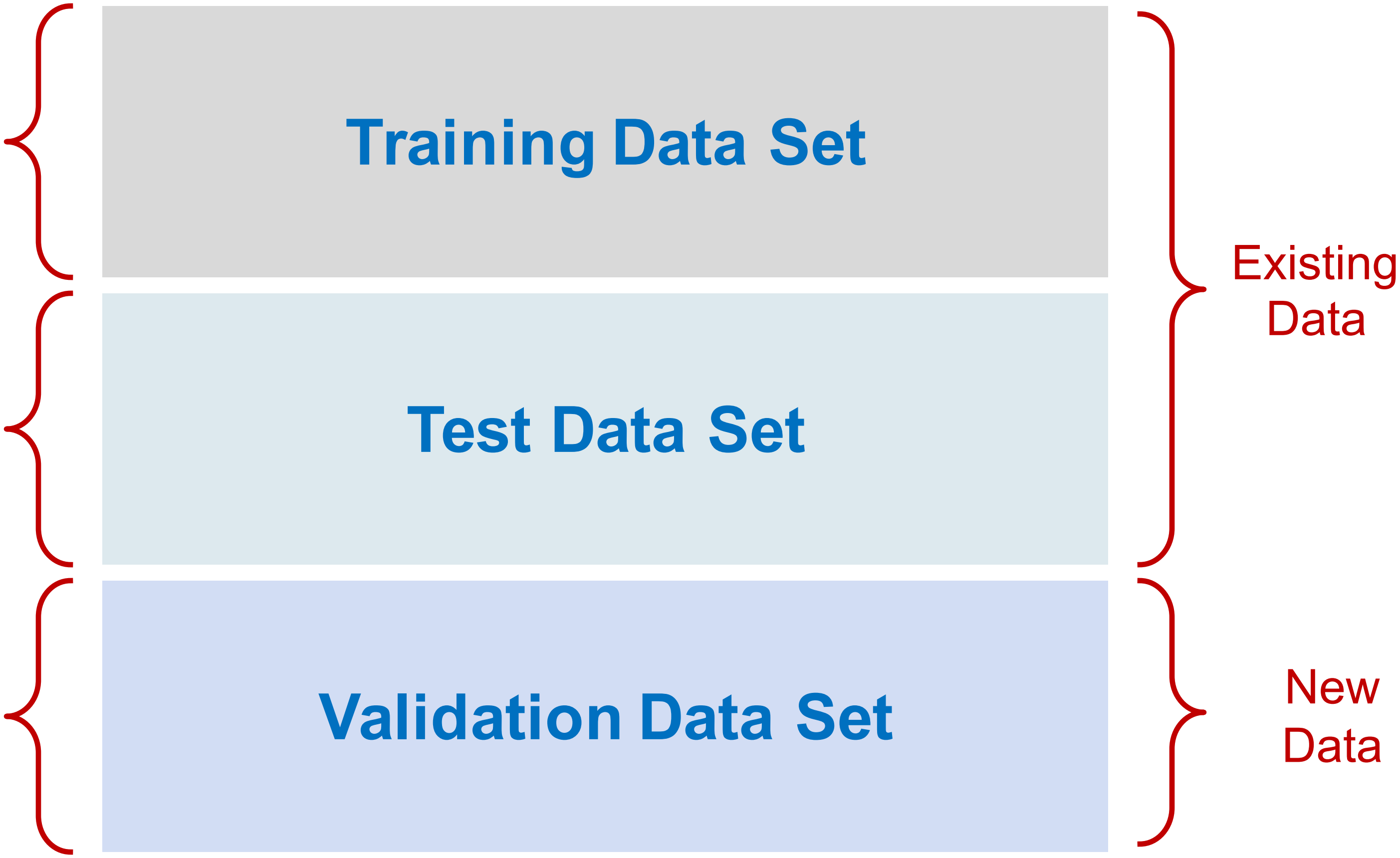
Supervised vs. Unsupervised Learning

- Machine learning can be **supervised** or **unsupervised**.
- We learn many things in life (e.g., **how to walk**) without even thinking or **without** specific **goals** in mind.
- Likewise, when we apply machine learning methods to learn from the data without a specific goal, this is called **"unsupervised"** learning.
- **Data mining** is ***"the computational process of discovering trends in data (ACM)"*** which were previously unknown – i.e., more closely associated with **"unsupervised learning"**.
- When you explore **descriptive statistics** and **correlations** you are using unsupervised learning methods.
- We learn other things in life with a **specific purpose** or **goal** (e.g., predictive analytics).
- Likewise, when we apply machine learning methods with a specific goal in mind, we call this **"supervised"** learning
- Most **predictive analytics** methods fall under the category of **"supervised"** learning – there is a specific **prediction goal**

Key Machine Learning Concepts

- **Training** – when we **run models** on the data to obtain parameters to help us understand the data (e.g., means, variance, regression coefficients), this is called "training" the model.
- **Testing** – some time we use all the data to develop predictive models, but this is not very useful for evaluating the predictive accuracy of the data. So, it is customary to set aside a portion of the data to test the model.
- **Training/Test/Validation Data Sets** – when the data is **partitioned** into a part to train the model and the other part to test the model, we referred to these data portions as the **training** and **test** data **sets**. When we obtain new data to evaluate the model we call this new data the **"validation"** data set.
- This is particularly important because **over-identified** models are notorious for fitting the existing data well, but **performing poorly** with different data
- Partitioning the data into various training and test sets and computing aggregate predictive accuracy scores is **central to machine learning**.

Machine Learning Illustration



Develop Model

Training Data Set

Evaluate Model with Existing Data

Test Data Set

Validate Model with New Data

Validation Data Set

Existing Data

New Data

Training vs. Test Error

- **Training Error** – is the **MSE** calculated with the same training data used to develop the model.
- When models are **over-identified** the training error tends to be small because, particularly when the **R^2** is high.
- But **over-identified** models and models with high **multicollinearity** are notorious for having high predictive accuracy with the same data used to build the model (i.e., the training set) but **perform poorly** when **tested** with **different data**.
- You can test the robustness of your model by dropping a few random data points and see if your model parameter estimates change substantially – **"shaking the tree"**
- **Test Error** – is the **MSE** calculated with new observations or existing observations not used in the training method.
- The **training** and **test** error are often **very different**
- And the **training error** can substantially **underestimate** the **test error**

Kogod School
of
Business