Machine Learning Workshop

December 2, 2017

Overview

Overview of Machine Learning

- What is Machine Learning?
- Supervised and Unsupervised Learning
- Regressing and Classification
- Feature Engineering and Dimensionality Reduction
- Overfit/Underfit Challenges
- Batch vs. Online Operation

Before We Start

Before We Start

- Be sure to install the following in order to run the examples, you can still follow along in the browser regardless.

https://www.anaconda.com/download

```
pip install pandas matplotlib seaborn scikit-learn
```

Before We Start

- The following is a link to the Juptyer Notebook for this workshop, it can be viewed directly in the browser, if you installed the software earlier you can also run it locally on your computer.

https://git.io/vbmxJ

What is Machine Learning?

What is Machine Learning?

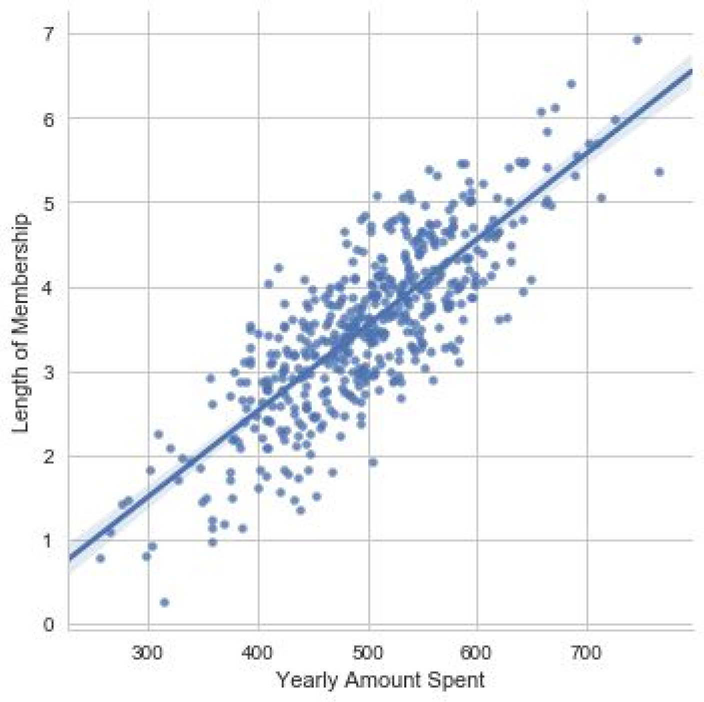
- Machine Learning is the field of study that makes it possible for problems to be solved by computers without requiring direct programming to be implemented.
- It's foundations are grounded in statistics, particularly Bayesian, and has since progressed with the early advent of Support Vector Machines and Logistic Regression.
- The field is broad and the main focus now is on Deep Learning, involving creating Artificial Neural Networks (ANN) with many layers, these are leading to many promising results!

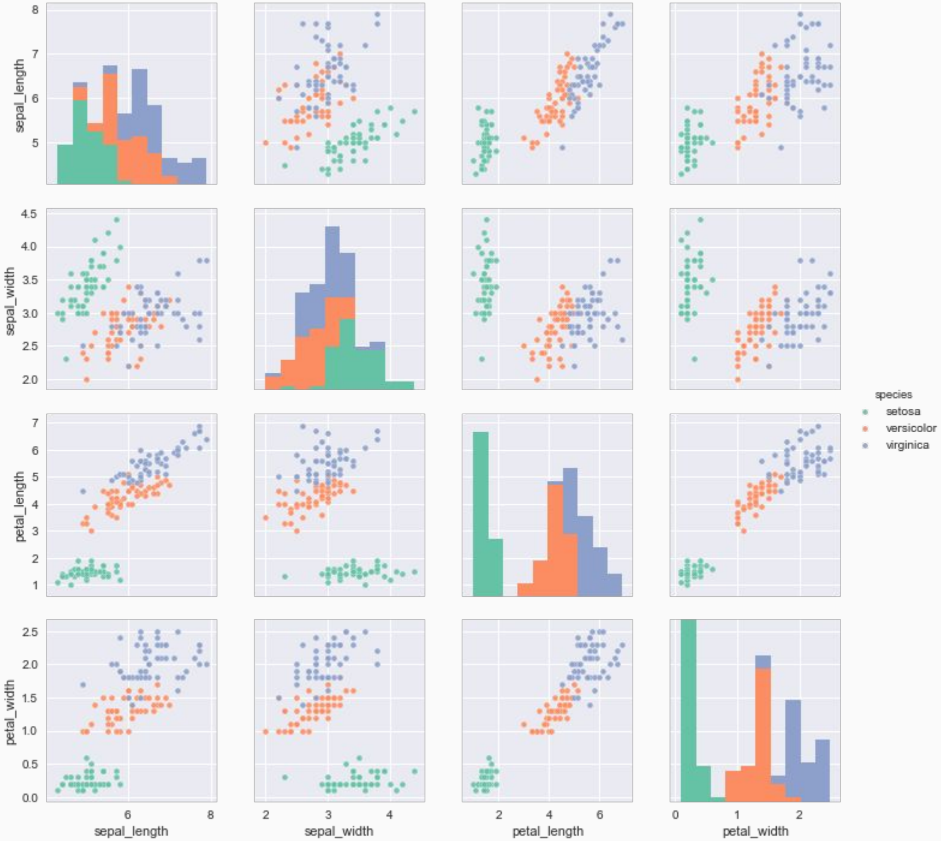
Supervised and Unsupervised Learning

- Are the most common learning methods in Machine Learning, although there are others, we will focus on these two.
- Supervised learning is as it sounds, it's where explicit known labels are provided, the supervision an outside informer (e.g. human) providing a list of known labels to train and evaluate the model.
- Unsupervised learning involves having a model attempt to identify patterns or any underlying information that can be extracted from the data without any oversight or external input. A common example would be clustering to find related groups within a dataset.

Regressing and Classification

- Regression involves trying to estimate the relationship among variables, traditionally with statistical methods this involved ordinary least squares (OLS), Machine Learning models attempt to achieve the same, estimating the outcome for a real number.
- Classification involves determining the class of an outcome from a discrete set of possible outcomes, traditionally this was performed using logistics regression, Machine Learning models are applied to attempt to determine the class of an estimated outcome.

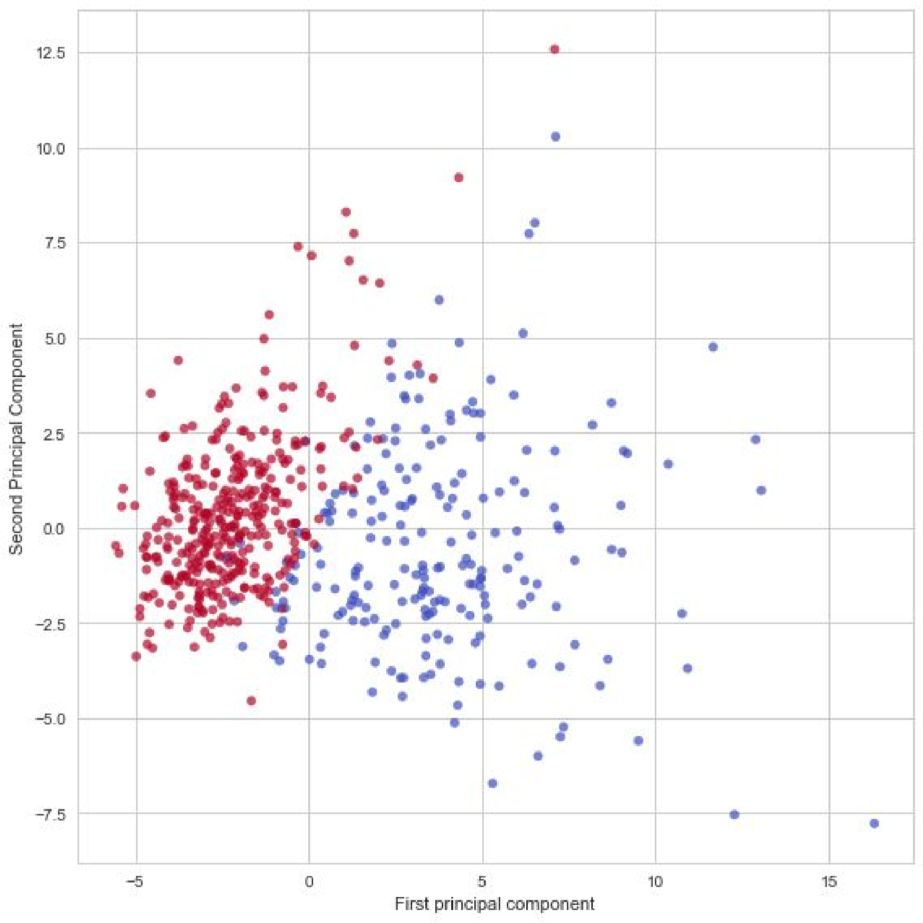




Examples of regression and classification problems solved using Machine Learning methods

Feature Engineering and Dimensionality Reduction

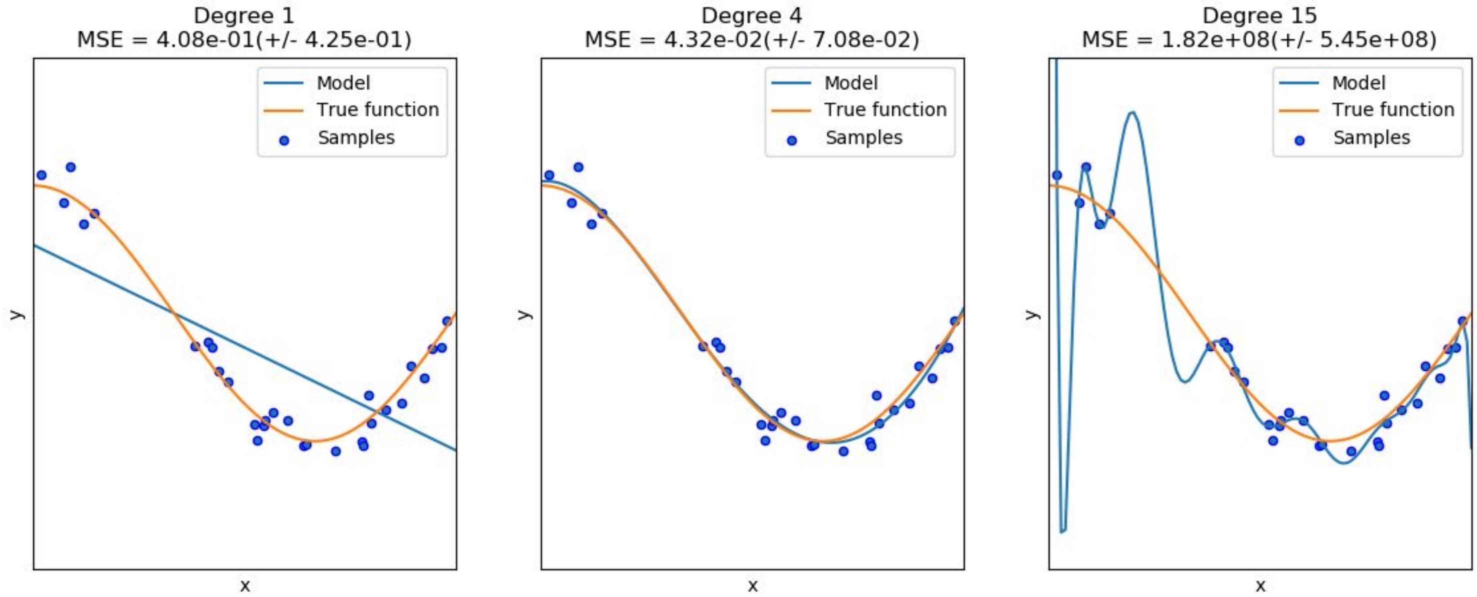
- One of the biggest challenges in Machine Learning is getting the data in the optimal format and also extracting/encoding the information in such a way that it's optimal for the model to infer the desired outcome.
- Another challenge is dealing with noisy/erroneous data, which is often the bulk of the initial Machine Learning Work.
- Dimensionality reduction involves reducing the dimensionality of the data by analyzing the influence of vectors on describing the variance of the data (e.g. Principal Component Analysis).



Example of linearly separable data classes after performing Principal Component Analysis.

Overfit/Underfit Challenges

- Overfitting occurs when the model is over trained using the training data, resulting in an output that excels at predicting the provided training data, but fails to generalize to test data.
- Underfitting occurs normally when a model is too generalized and does not match the training data closely enough to be beneficial for either training data or test data.
- There is no free lunch, there are always trade-offs and it may take many experiments to find the right balance between the two, with Deep Learning sometimes the challenge is amplified by exploding / vanishing gradients.



Examples of overfitting and underfitting

Batch vs. Online Operation

- Once you have a model how do you use it in a real-world problem? This is often a challenge, the traditional examples shown involving separating the data into training and testing sets.
- Training on a predefined set of data makes the model often incapable of incremental learning, training and improving based on new incoming data, the model was made just for the train/test data given.
- Online involves ensuring that the model can continue to learn as new data becomes available and then apply that data to the desired outcome.

Machine Learning Walkthrough

Machine Learning Walkthrough

- The following is a link to the Juptyer Notebook for this workshop, it can be viewed directly in the browser, if you installed the software earlier you can also run it locally on your computer.

https://git.io/vbmxJ