Machine Learning

Introduction

Introduction

- A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.
- In the case of a spam filter,
- the task T is to flag spam for new emails,
- the experience E is the training data (mails marked by the user as spam),
- the performance measure P is the ratio of correctly classified emails (accuracy).

Introduction

- Machine Learning applications:
- It is used in data based and data-rich application domains.
- A Machine Learning algorithm can often simplify code and perform better for problems for which existing solutions require a lot of complex coding / algorithms.
- It is suitable for complex problems for which there is no good solution at all using a traditional approach.
- A Machine Learning system can adapt to new data.
- It is used for getting insights about complex problems and large amounts of data.

Introduction

- The different types of Machine Learning systems are
- Supervised, Unsupervised, Semi-supervised and Reinforcement learning models.
- Online models / batch models.
- Instance-based models.

Challenges

- Challenges in Machine Learning surface up in the form of
  - Dataset on which the training algorithm should learn.
  - The right training model that suits the given data.
- Insufficient training data
- Even for very simple problems you one needs thousands of examples.
- And for complex problems such as image or speech recognition one may need millions of examples.

Challenges

- Nonrepresentative Training data / Imbalanced datasets
- In order to generalize well, it is crucial that your training data be representative of all the cases one wants to generalize to.
- Poor-quality data
- If the training data is full of missing data, outliers, and noise, it will make it harder for the training algorithm to detect the underlying patterns, so the system is less likely to perform well.

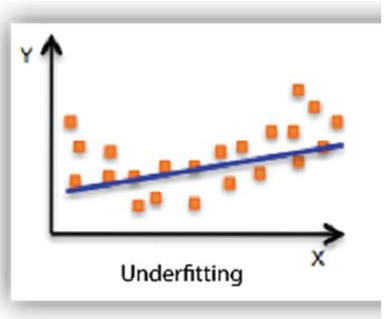
Challenges

- It is often well worth the effort to spend time cleaning up the training data.
- If some instances are clearly outliers, it may help to simply discard them or try to fix the errors manually.
- If some instances have missing features (e.g., 25% of the samples has null values for a feature), one must decide whether
  - to ignore this attribute/feature altogether,
  - to ignore these instances,
  - to fill in the missing values (e.g., with the median age), or
  - to train one model with the feature and one model without it, and so on.

Challenges

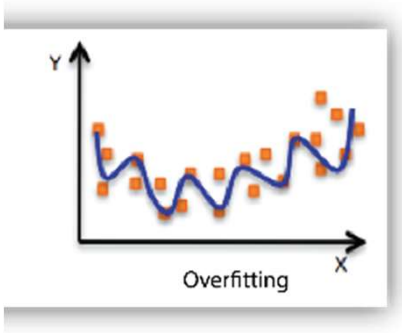
- Irrelevant features
- The training algorithm learns better if training data contains enough relevant features correlated with the target value and not too many irrelevant ones. One needs to perform,
- Feature selection: selecting the most useful features to train on among existing features.
- Feature extraction: combining existing features to produce a more useful ones.
- Creating new features by gathering new data.

Challenges



- Another challenge is to select and train the model such that it neither overfits nor underfits.
- Model - Underfitting
- If the model is too simplistic to capture the underlying structure of the data, it leads to underfitting.
- In other words, the model is unable to capture the true relationship between the features and the target variable.
- This is known as high bias can result in poor predictive performance because the model cannot represent the complexity of the data.

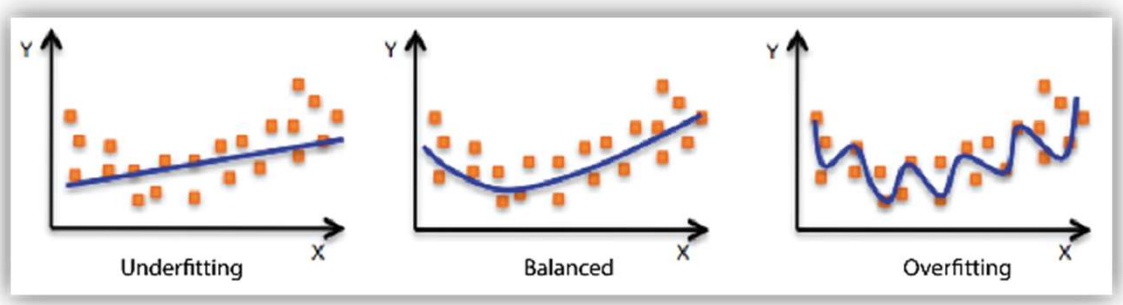
Challenges



- Model - Overfitting
- Overfitting occurs when a model learns the noise and specific patterns in the training data too well, to the extent that it fails to generalize to unseen data.
- This means that the model's predictions are highly sensitive to small fluctuations in the data.
- Such models tend to have excessively complex structures or too many parameters, allowing them to fit the training data closely. Such models have low bias but will be unable to generalize well to new, unseen data.

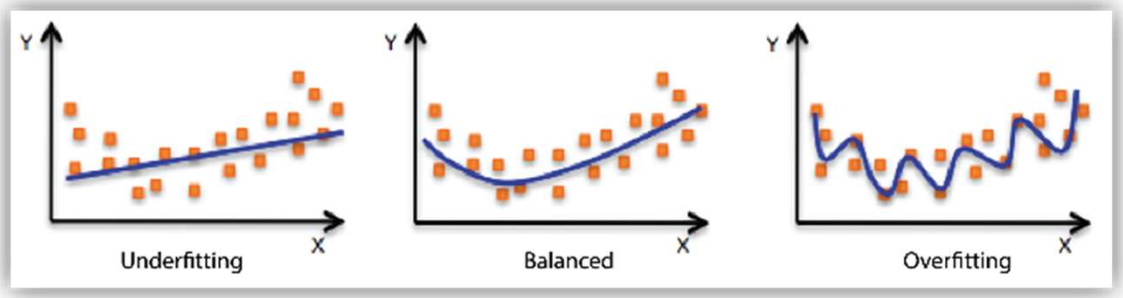
Challenges

Challenges



- Variance
- Underfitting:
- Conversely, in underfitting, the model fails to capture the underlying patterns in the data adequately. This leads to consistently large prediction errors across different subsets of the testing data.
- Although the errors are consistently large, they are not necessarily fluctuating wildly, which indicates low variance in the model's prediction errors.

Challenges



- Variance
- Overfitting:
- In the case of overfitting, the model captures not only the underlying patterns but also the noise present in the training data.
- As a result, the prediction errors made by the model may vary significantly when tested on different subsets of the data.
- This high variability or fluctuation in the predictions across different subsets of the training data indicates high variance in the model's prediction errors.

Applications

- Some of the applications are:
- Spam filtering
- Sales prediction
- Fraud detection
- Disease diagnosis
- Image classification / segmentation
- Sentiment analysis / Text classification
- Text summarization