In this section machine learning and its application will be discussed.

Machine Learning

Computational method using experience to improve performance or to make accurate prediction. ML algorithms depend on:

- Sample complexity
- Time complexity
- Space complexity

Experience

Refers to the past information available to the learner. Related to:

- Data analysis
- Statistics
- Probability
- Optimization

Application

- Text or document classification
- Natural language processing
- Speech recognition
- Computational biology
- Computer vision
- Fraud detection
- Games
- Medical diagnosis
- Search engines

What kind of problems can be solved with machine learning methods?

Learning Problems

- **Classification**: Assign a category to each item: documents- politics, business, sports, weather.
- **Regression**: Predict a real value to each item: prediction of stock values.
- **Ranking**: Order items according to some criterion: Web search.
- **Clustering**: Partition items into homogeneous regions: context of social network analysis.
- **Dimensionality reduction or Manifold learning**: Transform initial representation of items into a lower dimensional representation: preprocessing digital images.

Keywords

Learning problem- Spam detection:

- **Examples**: Items or instances of data used for learning or evaluation: Collection of emails.
- **Features**: The set of attributes: length of message, sender, header, subject, keywords, …
- **Labels**: Values or categories assigned to examples: Spam and Non-Spam.

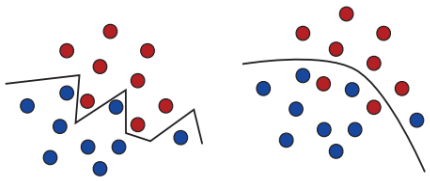


Figure 1.1 The zig-zag line on the left panel is consistent over the blue and red training sample, but it is a complex separation surface that is not likely to generalize well to unseen data. In contrast, the decision surface on the right panel is simpler and might generalize better in spite of its misclassification of a few points of the training sample.

Learning Stages of Spam problem

1. Give collection of labeled examples.
2. Partition the data into a training sample, a validation sample, and a test sample.
3. Associate relevant features to the examples.
4. Assign free parameters and hypothesis set.
5. Predict the labels of the examples in the test sample.
6. Calculate the Loss.

The most frequent words in machine learning will be defined in this section.

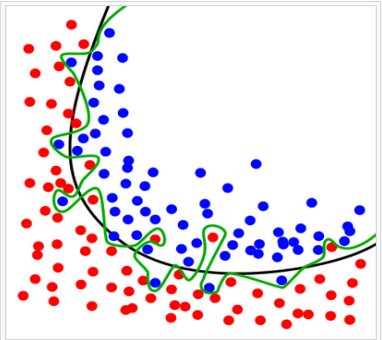
Definitions

- **Training Sample**: Examples used to train a learning algorithm: a set of email examples with labels.
- **Validation Sample**: Examples used to tune the parameters of a learning algorithm when working with labeled data to avoid overfitting.
- **Test Sample**: Examples used to evaluate the performance of a learning algorithm.
- **Loss Function**: A function that measures the difference, or loss, between a predicted label and a true label.
- **Hypothesis set**: A set of functions mapping features (feature vectors) to the set of label Y.

Overfitting

a statistical model describes random error or noise instead of the underlying relationship. Overfitting occurs when a model is excessively complex, such as having too many parameters relative to the number of observations.

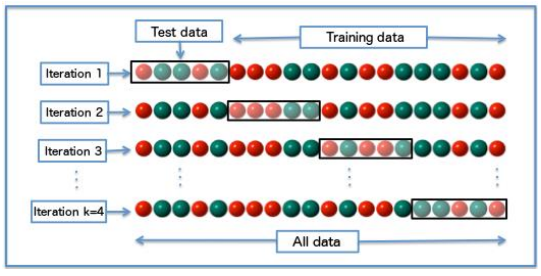
A model that has been overfit has poor predictive performance, as it overreacts to minor fluctuations in the training data.



The green line represents an overfitted model and the black line represents a regularised model. While the green line best follows the training data, it is too dependent on it and it is likely to have a higher error rate on new unseen data, compared to the black line.

Cross Validation

In k-fold cross-validation, the original sample is randomly partitioned into k equal sized subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k − 1 subsamples are used as training data.



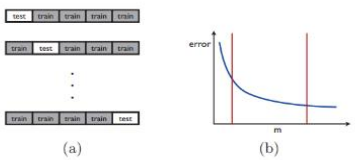


Figure 1.2 n-fold cross validation. (a) Illustration of the partitioning of the training data into 5 folds. (b) Typical plot of a classifier's prediction error as a function of the size of the training sample: the error decreases as a function of the number of training points.

Supervised and unsupervised learnings will be discussed in this section.

Learning Scenarios

- **Supervised learning**: The learner receives a set of labeled examples as training data and makes predictions for all unseen points: classification, regression, and ranking problems, and Spam detection.
- **Unsupervised learning**: The learner exclusively receives unlabeled training data, and makes predictions for all unseen points: Clustering and dimensionality reduction.
- **Semi-supervised learning**: The learner receives a training sample consisting of both labeled and unlabeled data, and makes predictions for all unseen points: classification, regression, ranking tasks.