January 23, 2018

# Problem Set 1

Sanchit Singhal

INF 385T – Introduction to Machine Learning with Danna Gurari

Spring 2018

School of Information

## 1. Supervised versus Unsupervised Learning

Based on the type and amount of supervision a machine learning system gets during training, it can be classified into four categories – supervised and unsupervised being two of them.

Supervised Learning is when the training set used to teach the algorithm contains labels (desired solutions) for instances.

Unsupervised Learning is when the training set does not include labels but instead the system, through various methods, attempts to learn by itself.

As mentioned in the definitions, the primary difference between the learning approaches remains with the training data. Supervised learning includes both input and output variables to approximate the mapping function so that output can be estimated for new input information. In contrast, unsupervised learning has no corresponding output variable for the input data. The goal is to model the underlying pattern or structure in the data. Therefore, unlike supervised learning, there is no "correct" answer and the algorithm is tasked with discovering interesting structure of patterns.

Because of this difference, the approaches employ different learning algorithms. The most important ones for each are listed below.

Supervised Learning algorithms:
- K-Nearest Neighbors
- Linear Regression
- Logistic Regression
- Supper Vector Machines (SVMs)
- Decision Trees and Random Forests
- Neural Networks

Unsupervised Learning algorithms:
- Clustering
- Dimensionality reduction
- Association rule learning

## 2. Supervised Learning: Regression versus Classification

Supervised Learning can be used to solve different kinds of learning problems including regression and classification.

Regression problems are when the task is to predict a specific numeric value using a set of features called predictors and their labels.

Classification problems are when the task is to categorize new input into classes based on examples in the training data.

As stated, the difference between the two learning problems lie with the type of output that is expected of them. Regression problems are when the output is a real value and classification is when the output variable is a category. An example of real world solutions that use regression

might be time series forecasts however recommendation systems probably use classification. The algorithms used for each problem are different too. An example might be to use linear regression for regression problems whereas classification problems can use support vector machines.

**3. Supervised Learning: Generalization**

**a)** The motivation behind splitting data into a training and testing datasets is to evaluate the machine's learning. If one were to test the system with the same data that it learned from, this would be non-beneficial. Because the machine has learned from these examples itself, it will do well but still may not generalize to new data points very well. By using part of the available data to test with instead of train, it can be used to assess the performance of the model and corrections can be made when needed.

**b)** If a model performs well on the training data but generalize poorly to new instances, the model is overfitting. This means the model is accounting too much for the noise in the training data and finding patterns that may not apply outside the sample.

**c)** If a model performs poorly on both the training data and new instances, the model is underfitting. This means the model is too simple to learn about the patterns in the data and therefore does not do well in either the training data or new input.

**d)** The motivation behind regularizing a model to simplify a complex problem. When reality is more complex than a model can be, reducing the constrains on the model can help resolve the issue of underfitting.

**4. Instance-Based versus Model-Based Learning**

Making predictions in machine learning tasks involves taking training examples to learn and then generalizing to examples that the system has not encountered. There are two main approaches for this – instance-based learning and model-based learning.

Instance-based learning is when the system memorizes the training data and based on the measure of similarity that is set, uses it to generalize for new input.

Model-based learning is when the system creates a model out of the training data and then uses this model to predict for new input.

The primary difference between the two learning approaches is the process used to generalize. While instance-based literally uses examples and similarities to judge new instances, model-based learning creates a model with the example to evaluate new data.

## 5. Online versus Online Learning

Based on the ability to incrementally learn, machine learning systems can be classified into online or offline learning approaches.

Offline Learning, or batch learning, is when a system can not learn incrementally and therefore must be taught with all the data as it does not have the capability to learn after launching.

Online Learning, or incremental learning, is when the system can learn incrementally through sequential data input – both in mini-batches or one-by-one.

As stated in the definitions, the obvious difference is the ability to learn incrementally or not but this leads to various implications that must be considered when designing the system. Because offline learning can't learn after deployment, it must be retrained with the entire dataset if new training data is to be added - this is expensive in terms of computing. In comparison, online learning systems work better with a constant flow of data and are able to adapt faster and on larger datasets. Although, online learning systems are more susceptible to bad data and must be monitored more carefully.