Mathematical Approach of Machine Learning Algorithm

– A Short Introduction –

Julien Lin

December 2016

Part I

Introduction to

Supervised Machine Learning

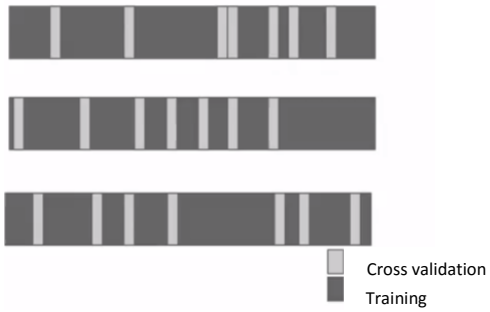
I. Supervised Learning Algorithm

Supervised learning is a machine learning task of inferring a function from **labelled** training data (a set of training examples). Each example is a pair consisting of an input variable x and a desired output value y, represented as $\{(x_1,y_1), (x_2,y_2), \ldots, (x_n,y_n)\}$.

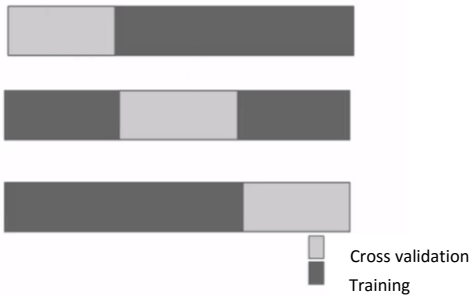
1) *The central dogma of Supervised Machine Learning*

The design of a hypothesis $h_\theta(x)$ using supervised learning algorithm consist of:

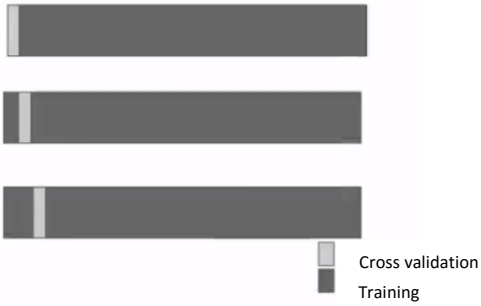
1) Identify a specific question: *"What do we try to predict", "What can we use to predict it?"*
2) Collect the best possible input data to try to answer the question.
3) Divide the all dataset into subset (60% into training set, 20% into cross-validation set, 20% into testing set) using Cross Validation technique (e.g. Random Subsampling, k-fold, Leave-one-out) .
   - **Random Subsampling**: Randomly divide the dataset into subsets across the folds such as:
     Build each hypothesis function on each training subsets and apply each of them on each cross-validation subsets



   - **k-fold:** Divide the dataset into k-equal sized datasets such as:
     Build each hypothesis function on each k(i^{th}) fold and apply each of them on each cross-validation k(i^{th}) fold



   - **Leave-one-out**: Divide the dataset by assigning for each fold only one sample as the cross validation set and the remaining samples as the training set.
     Build the hypothesis function on the remaining training set and apply it to the cross validation set. Then, move to the next sample on the next fold and assign it as the cross validation set and the remaining samples as the training set. Repeat the process until each sample had been assigned at least once as a cross-validation set.



4) In the **Training set** (60% of the overall dataset), analyses the labelled Training dataset and extract the most useful feature.
5) Build several hypothesis function (e.g. predictors, classifiers) that would fit to the training set well: minimize each of the training cost function $J_{train}(\boldsymbol{\theta}) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$ and obtain the optimal value of the parameters θ for each hypothesis model. Minimising the cost function refers to reducing the errors (i.e. false positive and false negative that the hypothesis function will get when fitting to the dataset). Thus, the goal of supervised machine learning is to avoid those errors when predicting or classifying.
6) Apply each hypothesis model with their optimal parameter θ on the **Cross Validation set** (20% of the overall dataset) to select the best hypothesis model with the best minimised cross validation cost function $J_{cv}(\boldsymbol{\theta}) = \frac{1}{2m_{cv}}\sum_{i=1}^{m_{cv}}(h_\theta({x_{cv}}^{(i)}) - {y_{cv}}^{(i)})^2$.
7) Optimize the model to avoid Overfitting/Underfitting, Bias/Variance
8) Apply the predictive function only once on an unseen **Testing set**

In other term,

1. $h_\theta = \theta_0 + \theta_1 x \longrightarrow minimize_\theta \; J(\theta) \longrightarrow \theta^{(1)} \longrightarrow J_{cv}\theta^{(1)}$
2. $h_\theta = \theta_0 + \theta_1 x + \theta_2 x^2 \longrightarrow minimize_\theta \; J(\theta) \longrightarrow \theta^{(2)} \longrightarrow J_{cv}\theta^{(2)}$
3. $h_\theta = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 \longrightarrow minimize_\theta \; J(\theta) \longrightarrow \theta^{(3)} \longrightarrow J_{cv}\theta^{(3)}$

   …

   $h_\theta = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \ldots + \theta_i x^i \longrightarrow minimize_\theta \; J(\theta) \longrightarrow \theta^{(i)} \longrightarrow J_{cv}\theta^{(i)} \longrightarrow J_{test}\theta^{(i)}$

   …

10. $h_\theta = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \cdots + \theta_{10} x^{10} \longrightarrow minimize_\theta \; J(\theta) \longrightarrow \theta^{(10)} \longrightarrow J_{cv}\theta^{(10)}$

Select the model with the lowest $J_{cv}\theta^{(i)}$ that generalise the best to new example (test set) and apply the model to estimate the $J_{test}\theta^{(i)}$