

# Algorithmic Learning Theory

## Spring 2017

### Lecture 1

**Instructor:** Farid Alizadeh

**Scribe:** Yuan Qu

01/18/2017

There are 4 section in the Lecture 1:

1. Overview
2. Fundamental Concepts
3. Distinction of Concepts
4. Bayesian Decision Rule

## 1 Overview

Brief of the Syllabus and the tools needed.

### 1.1 Textbooks

See *Syllabus* in *Sakai*.

### 1.2 Student Work

See *Syllabus* in *Sakai*.

No test, **BUT**:

1. Course note, in LaTeX
2. Student projects (Can be in small group)

### 1.3 Topics

See *Syllabus* in *Sakai*.

### 1.4 Tools

*R* & *RStudio*

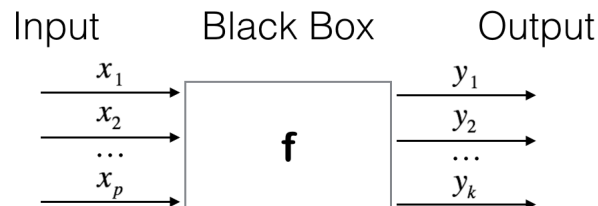
## 2 Fundamental Concepts

Introduction of basic concepts and ideas of machine learning.

### 2.1 Black Box

**Definition by Wikipedia:** *In science, computing, and engineering, a black box is a device, system or object which can be viewed in terms of its inputs and outputs (or transfer characteristics), without any knowledge of its internal workings.*

Consider a black box with a bunch of input and output,



Which consists of 3 parts, Input, Output and the Function  $f$ :

**Input**  $x_1, x_2, \dots, x_p$ , also called independ vars, predictors, features and controls.

The input can be numerical or non-numerical(categorical), if input is in multi-class, it can be ordered(ordinal) or unordered.

**Output**  $y_1, y_2, \dots, y_k$ , also called response, depend vars and target vars. For example,

f()	Output
regression	numerical
classification	categorical

## 2.2 Data

**Definition by Wikipedia: Data(computing)** *is any sequence of one or more symbols given meaning by specific act(s) of interpretation.*

Data is also called example and observation and usually organized by table, like this:

row	$x_1$	$x_2$	...	$x_p$	$y_1$	$y_2$	...	$y_k$
1	...	...	...	...	...	...	...	...
2	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...
N	...	...	...	...	...	...	...	...

The data table can be regarded as  $\hat{f}$ , an estimation of  $f$ , which is considered an algorithm or a rule.

For most application scenarios, data is given or generated by an algorithm called generation method.

In a dataset, there may be some missing data(N/A) in the table.

## 2.3 Big Data

Data that can't be located in one computer.

**"Big"** large scale in 2 dimensions, the number of records/items(rows) and the number of features(columns).

**More features is good?** There may be some irrelevant or redundant features in dataset, so that feature selection is a major problem.

## 3 Distinctions

The distinctions between some concepts in machine learning.

### 3.1 Parametric vs Non-parametric

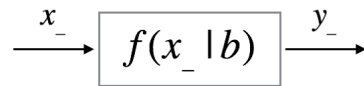
e.g.  $x = \text{height}, y = \text{weight}, y = f(x)$

#### 3.1.1 Parametric

e.g. Regression

Assumption:  $y = b_0 + b_1x$ ,  $b_0, b_1$  are the parameters.

So, the problem can be showed like this:



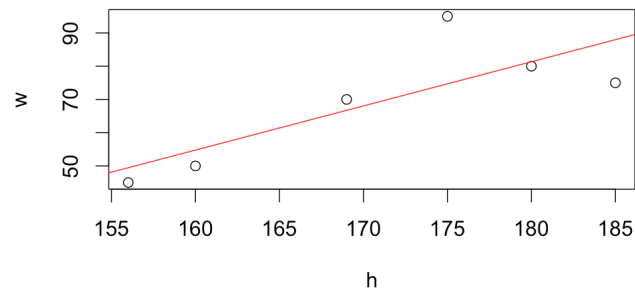
The work of machine learning is to get data and find the parameters.

Supposed we get the dataset like this:

W	H
$w_1$	$h_1$
$w_2$	$h_2$
...	...
$w_n$	$h_n$

Assume:  $w = b_0 + b_1h$

So, the regression plot is:  $\hat{w} = \hat{b}_0 + \hat{b}_1h$

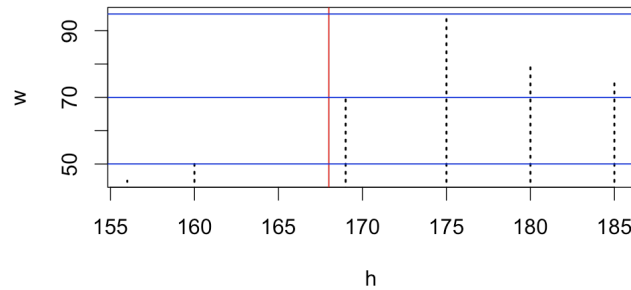


#### 3.1.2 Non-parametric

e.g. kNN.(k-nearest neighbors algorithm)

if  $k = 3$ , find  $k$  near point of  $h_{\text{new}}$ , so,

$$w_{\text{new}} = \text{average}(k \text{ closest points to } h_{\text{new}}) = \frac{w_1 + w_2 + w_3}{3}$$



In this case, the  $h_{new}$  is the red line, so, the  $w_1, w_2, w_3$  are the  $w$  in blue lines.

### 3.1.3 Conclusion

**Parametric** Assumption:  $y = b_0 + b_1x$ ,  $b_0, b_1$  are the parameters.

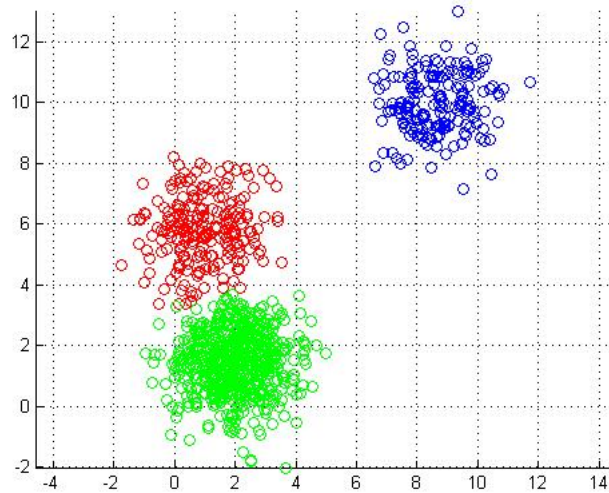
$$\xrightarrow{x_{new}} \boxed{\hat{f} = \hat{b}_0 + \hat{b}_1x} \xrightarrow{y_{new}}$$

**Non-parametric** More flexible. No parameters to impose, no assumption.

## 3.2 Supervised vs Unsupervised(non-supervised)

**Supervised** Name  $x$  and  $y$ , the input and output are specific by labels.

**Unsupervised** Every column of data is  $x$ , no labels.  
e.g k-means



When the number of dimensions increase, the difficulty of recognition increase, so that the selection of  $x$  is a major problem.

### 3.3 Online learning vs Offline(not online) learning

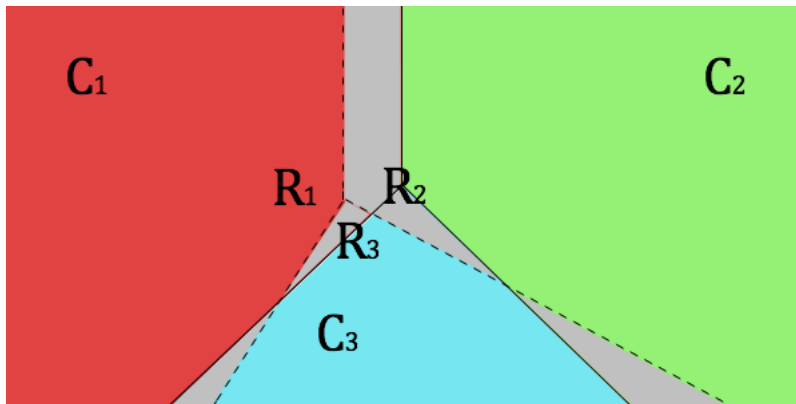
**Online learning** data becomes available in a sequential order and is used to update our best predictor for future data at each step.

**Offline(not online) learning** generate the best predictor by learning on the entire training data set at once.

### 3.4 Loss vs Overfitting

#### 3.4.1 Loss for classification

**Definition by Wikipedia** a loss function or cost function is a function that maps an event or values of one or more variables onto a real number intuitively representing some "cost" associated with the event.



The true classification is  $f$ , and the classes are  $C_1, C_2$ , and  $C_3$ , which are divided by line and painted in red, green and blue. Supposed that we don't know the true classification.

The estimated classification is  $\hat{f}$ , and the classes are  $R_1, R_2$ , and  $R_3$ , divided by dotted line.

The mistake area, painted by gray, is the difference between estimation and true classification.

Loss function abstract the size of mistake area.

$P_x(t) \rightarrow$  Probability density function

Misclassification rate:

$$\Pr_x[f(x) \neq \hat{f}(x)]$$

But misclassification rate above is not computable, because we have to know two function or classification.

Empirical error/misclassification rate:

$$\frac{\text{Number of misclassified data}}{\text{Total data number}}$$

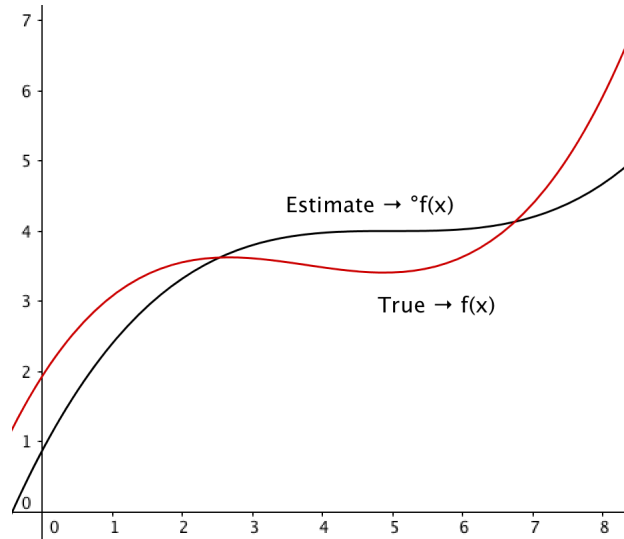
**Loss matrix** To show the cost of error in a matrix.

		Prediction			
		R	G	B	
True	R	0			cost of error
	G		0		cost of correct is 0
	B			0	

e.g.

		Prediction	
		Cancer	No
True	Cancer	0	10
	No	1	0

### 3.4.2 Loss for regression



Mean square error is used to eliminate the + and -.

Loss function:  $E_x(\ell(f, \hat{f}, x))$

Error rate:

$$\int \ell(f, \hat{f}, x) p(x) dx = \int (f(x) - \hat{f}(x))^2 p(x) dx$$

Major of loss:  $\text{Max}|f(x) - \hat{f}(x)|$

Empirical error:

$$\frac{\sum_{i=1}^N (f(x) - \hat{f}(x))^2}{N}$$

### 3.4.3 Overfitting

**Definition by Wikipedia** In 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.

In this case, overfitting is that the  $\hat{f}(x)$  performs in perfect result. Generally, overfitting is a bad phenomenon, because the data self may have error or miss relevant vars.

**Validation set** The technique to guard overfitting from complex model.

1. Set a set of data aside(validation set)
2. learning is done on the remaining data(training set)



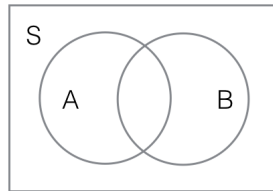
3. Use validation set to predict response

## 4 Bayesian Decision Rule

**Bayes Law:** For Event A and Event B

$$0 \leq P[A] \leq 1, 0 \leq P[B] \leq 1$$

(A and B:  $A \cap B$ , A or B:  $A \cup B$ , not A:  $\bar{A}$  or  $A^*$ )



If,

$$A \cap B \neq \emptyset, \quad (1)$$

So, according to (1),

$$P[A|B] = \frac{P[A \cap B]}{P[B]} \Rightarrow P[A \cap B] = P[A|B] \times P[B], \quad (2)$$

And, the same with (2),

$$P[B|A] = \frac{P[B \cap A]}{P[A]} \Rightarrow P[B \cap A] = P[B|A] \times P[A] \quad (3)$$

Obviously, we have,

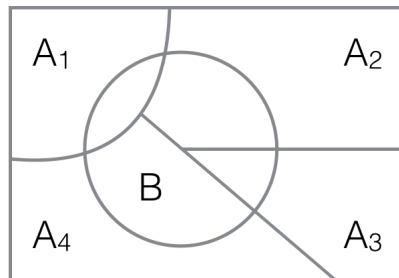
$$P[A|B]P[B] = P[B|A]P[A] \quad (4)$$

So,

$$P[A|B] = \frac{P[B|A] \times P[A]}{P[B]}$$

$$P[B|A] = \frac{P[A|B] \times P[B]}{P[A]}$$

**e.g.** If, a partition of A,



$$A_1 \cup A_2 \cup A_3 \cup A_4 = S, A_i \cap A_j = \emptyset, i, j = 1, 2, 3, 4$$

So,

$$P[B] = \sum_{i=1}^4 P[B|A_i]P[A_i]$$