BINUS UNIVERSITY

People
Innovation
Excellence

Course	: COMP6577 – Machine Learning
Effective Period	: February 2020

Learning in Parametric Modeling 1

Session 08 & 09

Learning Outcome

- LO2: Student be able to interpret the distribution of dataset using regression method

Outline

- Parametric Modeling
- Parameter estimation
- Linear regression
- Classification
- Case Study

Parametric Modeling

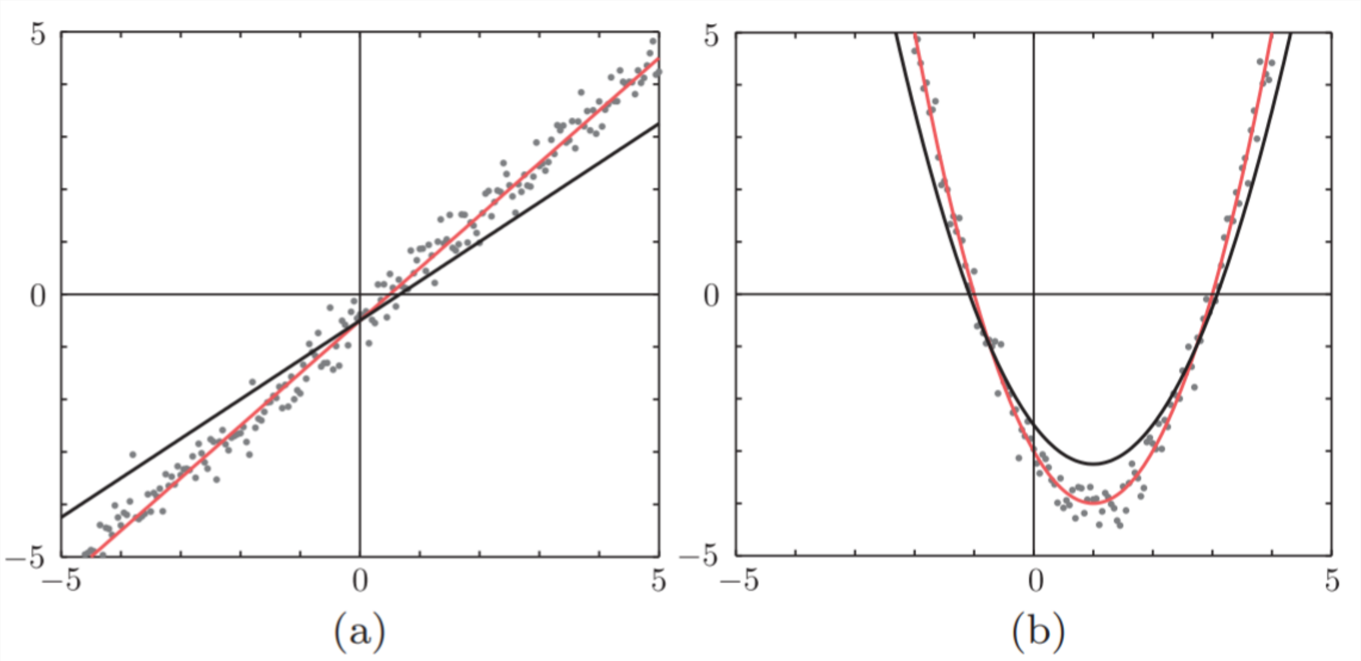
- Parametric models are mobilized to describe the available data.
- In parametric modeling, the aforementioned functional dependence is defined via a set of unknown parameters, whose number is fixed. In contrast (non-parametric methods), unknown parameters may still be involved, yet their number depends on the size of the data set.
- In parametric modeling, there are two possible paths to deal with the uncertainty imposed by the unknown values of the parameters. According to the first one, specific values are obtained and assigned to the unknown parameters. In the other approach, which has a stronger statistical flavor, parametric models are adopted in order to describe the underlying probability distributions, which describe the input and output variables, without it being necessary to obtain specific values for the unknown parameters

Parameter Estimation

- The task of estimating the value of an unknown parameter vector, θ, has been at the center of interest in a number of application areas.
- For example:
  - Given a set of data points, one must find a curve or a surface that “fits” the data. The usual path to follow is to adopt a functional form, such as a linear function or a quadratic one, and try to estimate the associated unknown coefficients so that the graph of the function “passes through” the data and follows their deployment in space as close as possible.

People
Innovation
Excellence

Parameter Estimation (2)



- Fitting (a) a linear function and (b) a quadratic one. The red lines are the optimized ones.

Parameter Estimation (3)

- The data lie in the $\mathbb{R}^2$ space and are given to us as a set of points *(y_n, x_n), n* = 1, 2, ... , *N*. The adopted functional form for the curve corresponding to the figure (a) before is:

$$y = f_\theta (x) = \theta_0 + \theta_1 x$$

- and for the case of figure (b) is:

$$y = f_\theta (x) = \theta_0 + \theta_1 x + \theta_2 x^2$$

- The unknown parameter vectors are $\theta = [\theta_0 , \theta_1]^T$ and $\theta = [\theta_0 , \theta_1, \theta_2]^T$, respectively. In both cases, the parameter values, which define the curves drawn by the red lines, provide a much better fit compared to the values associated with the black ones.

Parameter Estimation (4)

- In both cases, the task comprises two steps:
  - Adopt a specific parametric functional form, which we reckon to be more appropriate for the data at hand.
  - estimate the values of the unknown parameters in order to obtain a “good” fit.
- The task can be defined as follows:
- Given a set of data points, (y_n, x_n), y_n $\mathbb{R}$, , x_n ∈ $\mathbb{R}^l$, n = 1, 2, ... , *N*, and a parametric set of functions, find a function $\mathcal{F}$, , which will be denoted as $f(\cdot) := f_{\theta*}(\cdot)$, such that given a value $\mathbb{R}^l$ x ∈ , *f(x)* best approximates the corresponding $\mathbb{R}$, value y ∈

$$\mathcal{F} := \left\{ f_{\boldsymbol{\theta}}(\cdot) : \boldsymbol{\theta} \in \mathcal{A} \subseteq \mathbb{R}^K \right\}$$

Parameter Estimation (5)

- The value θ_* is the value that results from the estimation procedure. The values of θ_* that define the red line curves in Figures (a) and (b) in previous slides are:

$$\theta_* = [-0.5, 1]^T , \theta_* = [-3, -2, 1]^T$$

- Having adopted a parametric family of functior $\mathcal{F}$, one has to get an estimate for the unknown set of parameters. To this end, a measure of fitness has to be adopted. The more classical approach is to adopt a **loss function**, which quantifies the deviation/error between the measured value of *y* and the predicted one using the corresponding measurements *x*, as in $f_\theta(x)$.

Parameter Estimation (6)

- In a more formal way, a nonnegative (loss) function is adopted:

$$\mathcal{L}(\cdot,\cdot) : \mathbb{R} \times \mathbb{R} \longmapsto [0, \infty)$$

- and compute θ_* so as to minimize the total loss, or as we say the cost, over all the data points, or

$$f(\cdot) := f_{\boldsymbol{\theta}_*}(\cdot) : \ \boldsymbol{\theta}_* = \arg\min_{\boldsymbol{\theta} \in \mathcal{A}} J(\boldsymbol{\theta})$$

- Where $J(\boldsymbol{\theta}) := \sum_{n=1}^{N} \mathcal{L}\big(y_n, f_{\boldsymbol{\theta}}(\boldsymbol{x}_n)\big)$ assuming that a minimum exists.

- Note that, in general, there may be more than one optimal values $\theta*$, depending on the shape of $J(\theta)$.

People
Innovation
Excellence

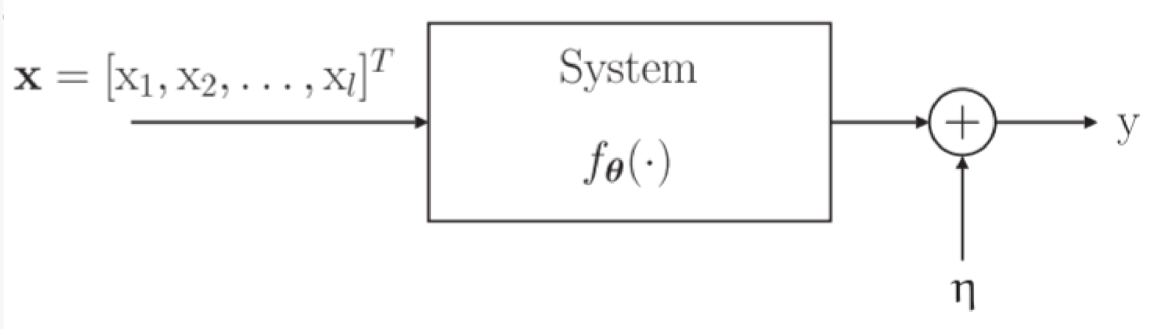
LS Loss Function

- The LS loss function is credited to the great mathematician Carl Frederich Gauss, who proposed the fundamentals of the LS method in 1795 at the age of eighteen.
- However, it was Adrien-Marie Legendre who first published the method in 1805, working independently.
- Gauss published it in 1809. The strength of the method was demonstrated when it was used to predict the location of the asteroid Ceres.
- Since then, the LS loss function has “haunted” all scientific fields, and even if it is not used directly, it is, most often, used as the standard against which the performance of more modern alternatives are compared.

$$\mathcal{L}\left(y, f_\theta(\boldsymbol{x})\right) = \left(y - f_\theta(\boldsymbol{x})\right)^2.$$

Linear Regression

- In statistics, the term **regression** was coined to define the task of modeling the relationship of a dependent random variable, *y*, which is considered to be the response of a system, when this is activated by a set of random variables, x_1, x_2, ... , x_l, which will be represented as the components of an equivalent random vector *x*. The relationship is modeled via an additive disturbance or noise term, η (unobserved random variable).
- This is the block diagram of the process, which relates the involv


$\mathbf{x} = [x_1, x_2, \ldots, x_l]^T$
System
$f_{\boldsymbol{\theta}}(\cdot)$
$+$
y
η

Linear Regression (2)

- The goal of the regression task is to estimate the parameter vector, *θ*, given a set of measurements, *(y_n, x_n)*, *n* = 1, 2, ... , *N*, that we have at our disposal. This is also known as the ***training data set***, or the ***observations***. The dependent variable is usually known as the output variable and the vector *x* as the input vector or the ***regressor***. If we model the system as a linear combiner, the dependence relationship is written as $y = \theta_0 + \theta_1 x_1 + \cdots + \theta_l x_l + \eta \quad = \theta_0 + \boldsymbol{\theta}^T \mathbf{x} + \eta$
- The parameter θ_0 is known as the ***bias*** or the ***intercept***. Usually, this term is absorbed by the parameter vector *θ* with a simultaneous increase of the dimension of x by adding the con Indeed, we can write:

$$\theta_0 + \boldsymbol{\theta}^T \mathbf{x} + \eta = [\boldsymbol{\theta}^T, \theta_0] \begin{bmatrix} \mathbf{x} \\ 1 \end{bmatrix} + \eta$$

People
Innovation
Excellence

Linear Regression (3)

- the regression model will be written as **y = θ^T x + η** and, unless otherwise stated, this notation means that the bias term has been absorbed by *θ* and *x* has been extended by adding 1 as an extra component.
- Because the noise variable is unobserved, we need a model to be able to predict the output value of *y*, given the value *x*.
- In linear r g prediction model:

$$\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 x_1 + \cdots + \hat{\theta}_l x_l := \hat{\boldsymbol{\theta}}^T \boldsymbol{x}$$

$\hat{\boldsymbol{\theta}}$

$\hat{y}_n$

- Using the LS loss function, the estimate is set equal to θ$_*$, which minimizes the square differe d y$_n$, over the set of the available observa minimizing, with respect to θ, the cost function

$$J(\boldsymbol{\theta}) = \sum_{n=1}^{N} (y_n - \boldsymbol{\theta}^T \boldsymbol{x}_n)^2$$

People
Innovation
Excellence

Linear Regression (4)

- Taking the derivative (gradient) with respect to θ and equating to the zero vector, 0, we obtain:

$$\left(\sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^T\right)\hat{\boldsymbol{\theta}} = \sum_{n=1}^{N} \boldsymbol{x}_n y_n \quad \Rightarrow \quad X := \begin{bmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_N^T \end{bmatrix} = \begin{bmatrix} x_{11} & \dots & x_{1l} & 1 \\ x_{21} & \dots & x_{2l} & 1 \\ \vdots & \ddots & \vdots & \\ x_{N1} & \dots & x_{Nl} & 1 \end{bmatrix}.$$

- Then, it is straightforward the equation can be written as $X^T X\hat{\boldsymbol{\theta}} = X^T\boldsymbol{y}$, where $\boldsymbol{y} := [y_1, y_2, \dots, y_N]^T$,

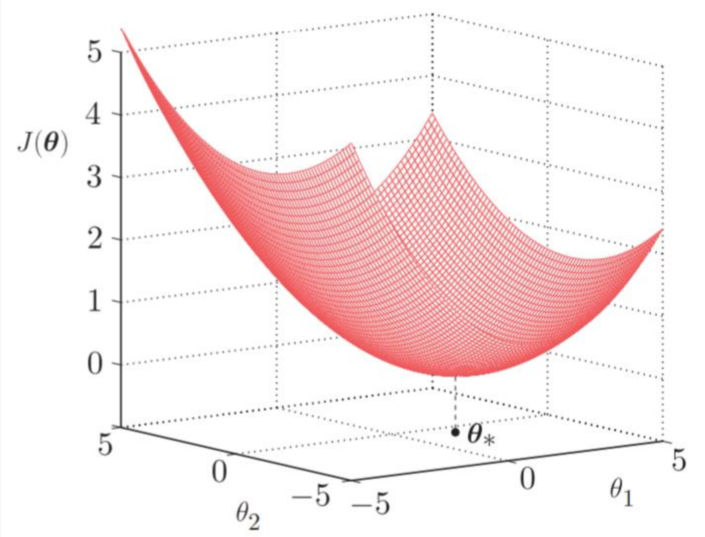
- and the LS estimate is given by

$$\boxed{\hat{\boldsymbol{\theta}} = (X^T X)^{-1} X^T \boldsymbol{y} : \quad \text{The LS Estimate,}}$$

that $(X^T X)^{-1}$ exists

Linear Regression (5)

- This is a major advantage of the LS loss function, when applied to a linear model. Moreover, this solution is unique, provided that the *(l + 1)* × *(l + 1)* matrix X^TX is invertible.
- The uniqueness is due to the parabolic shape of the graph of the LS cost function. The figure below is the illustration of two-dimentional space. The least-squares loss function has a unique minimum at the point θ



Classification

- Classification is the task of predicting the class to which an object, known as pattern, belongs. The pattern is assumed to belong to one and only one among a number of a priori known classes.
- Each pattern is uniquely represented by a set of measurements, known as **features**. One of the early stages in designing a classification system is to select an appropriate set of feature variables. These should “encode” as much class-discriminatory information, so that, by measuring their value for a given pattern, to be able to predict, with high enough probability, the class of the pattern.
- Selecting the appropriate set of features for each problem is not an easy task and it comprises one of the most important areas within the field of Pattern Recognition

Classification (2)

- To formulate the task in mathematical terms, each class is represented by the class label variable, *y*. For the simple two-class classification task, this can take either of two values, depending on the class, e.g., 1, −1, or 1, 0, etc. Then, given the value of *x*, corresponding to a specific pattern, its class label is predicted according to the rule,

$$\hat{y} = \phi(f(\boldsymbol{x}))$$

- where $\phi(\cdot)$ is a nonlinear function that indicates on which side of the decision surface, f(x) = 0, *x* lies.

Classification (3)

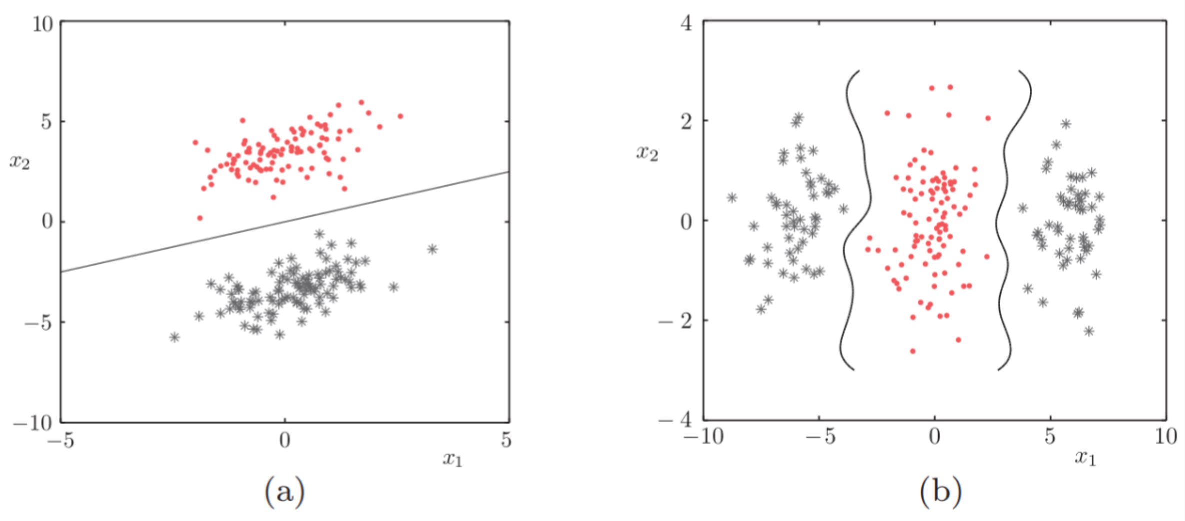
- So, is the classification any different from the regression task?
  - they are similar, yet different
- Note that in a classification task, the dependent variables are of a discrete nature, in contrast to the regression, where they lie in an interval. This suggests that, in general, different techniques have to be adopted to optimize the parameters.
- For example, the most obvious choice for a criterion in a classification task is the probability of error. However, in a number of cases, one can attack both tasks using the same type of loss functions. Even if such an approach is adopted, in spite of the similarities in their mathematical formalism, the goals of the two tasks remain different.

Classification (4)

- In the regression task, the function *f(·)* has to “explain” the data generation mechanism. The corresponding surface in the *(y, x)* sp$\mathbb{R}^{l+1}$ should develop so as to follow the spread of the data in the space, as close as possible.
- In contrast, in classification, the goal is to place the corresponding surface *f(x) = 0*, ir$\mathbb{R}^{l}$, so as to separate the data that belong to different classes as much as possible. The goal of a classifier is to partition the space where the features vectors lie into regions and associate each region with a class.

Classification (5)

- Example of two cases of classification tasks. The first one is an example of two linearly separable classes, where a straight line can separate the two classes, and the second one of two nonlinearly separable classes, where the use of a linear classifier would have failed to separate the two classes.



(a) (b)

Classification (Example)

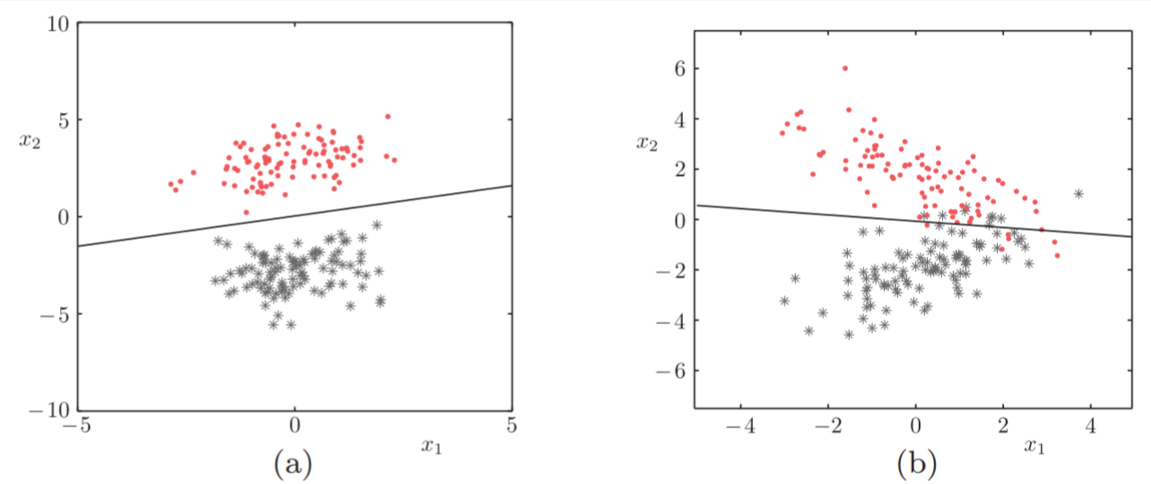
- We are given a set of training patterns, $x_n \mathbb{R}^l$, n = 1, 2, ... , *N*, that belong to either of two classes, say ω_1 and ω_2. The goal is to design a hyperplane $f(x) = \theta_0 + \theta_1 x_1 + \cdots + \theta_l x_l = \boldsymbol{\theta}^T \boldsymbol{x}$ where we have absorbed the bias θ_0 in θ and extend the dimension of x= *0.* Our aim is to place this hyperplane in between the two classes.
- Obviously, any point lying on this hyperplane scores a zero, f(x) = 0, and the points lying on either side of the hyperplane score either a positive (f(x) > 0) or a negative value (f(x) < 0), depending on which side of the hyperplane they lie. Therefore, we should train the classifier so that the points from one class score a positive value and the points of the other a negative one.

Classification (Example)

- This can be done, for example, by labeling all the points from class, say ω_1, with $y_n = 1$, $\forall_n : x_n \in \omega_1$, and all the points from class ω_2 with $y_n = -1$, $\forall_n : x_n \in \omega_2$. Then the LS loss is mobilized to compute θ so as to minimize the cost

$$J(\boldsymbol{\theta}) = \sum_{n=1}^{N} \left(y_n - \boldsymbol{\theta}^T \boldsymbol{x}_n \right)^2$$

- This figure shows the resulting LS classifiers for two cases of data.



(a)

(b)

People
Innovation
Excellence

Classification (Example)

- In case figure (b), the classifier cannot classify correctly all the data points. Our desire to place all the data, which originate from one class, on one side and the rest on the other cannot be satisfied. All that our LS classifier can do is to place the hyperplane so that the sum of squared errors, between the desired (true) values of the labels, y_n, and the predicted outputs, $\theta^T x_n$, are a minimum.
- It is mainly for cases such as overlapping classes, which are usually encountered in practice, where one has to look for an alternative to the LS criteria and methods, in order to serve better the needs and the goals of the classification task.

Case Study

Given data of Singapore Airbnb which can be downloaded in this link

https://www.kaggle.com/jojoker/singapore-airbnb

- By using the data from the link above, you can simulate linear regression and classification. You can try and choose the attributes yourself.

End of Session 08 & 09

References

- Sergios Theodoridis. (2015). *Machine Learning: a Bayesian and Optimization Perspective*. Jonathan Simpson. ISBN: 978-0-12-801522-3. Chapter 3.
- Aurélien Géron. (2017). 01. *Hands-on Machine Learning with Scikit-Learn and Tensorflow*. O'Reilly Media, Inc..LSI: 978-1-491-96229-9. Chapter 3.
- https://www.kaggle.com/jojoker/singapore-airbnb