Course : COMP6577 – Machine Learning
Effective Period : February 2020

Mean-Square Error Linear Estimation

Session 13 & 14

People
Innovation
Excellence

Learning Outcome

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

People
Innovation
Excellence

Outline

- The cost function surface
- A geometric viewpoint (orthogonality condition)
- Case Study

The Normal Equations

- The general estimation task has been introduced in the previous Topic.
- Given two dependent random vectors, *y* and *x*, the goal of the estimation task is to obtain a function, *g*, so as, given a value *x* of *x*, to be able to predict (estimate), in some optimal sense, the corresponding value *y* of *y*, or $\hat{y}$ = g(x).
- The optimal MSE estimate of *y* given the value *x* = *x* is

$$\hat{y} = \mathbb{E}[\mathbf{y}|\boldsymbol{x}].$$

- In general, this is a nonlinear function.

BINUS UNIVERSITY

People Innovation Excellence

Mean-Square Error Linear Estimation

- We now turn our attention to the case where g is constrained to be a linear function.
- Let (y, x) ∈ $\mathbb{R} \times \mathbb{R}^l$ be two jointly distributed random entities of **zero mean values**. In case the mean values are not zero, they are subtracted.
- Our goal is to obtain an estimate of θ ∈ $\mathbb{R}^l$ in the linear estimator model,

$$\hat{y} = \boldsymbol{\theta}^{\mathrm{T}}\mathbf{x},$$

- So that the cost function is minimum,

$$J(\boldsymbol{\theta}) = \mathbb{E}[(y - \hat{y})^2], \qquad \boldsymbol{\theta}_* := \arg\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta}).$$

- In other words, the optimal estimator is chosen so as to minimize the variance of the error random variable

$$\mathrm{e} = \mathrm{y} - \hat{\mathrm{y}}.$$

- Minimizing the cost function J(θ) is equivalent with setting its gradient with respect to θ equal to zero,

$$\begin{aligned}\nabla J(\boldsymbol{\theta}) &= \nabla\, \mathbb{E}\left[\left(\mathrm{y} - \boldsymbol{\theta}^T \mathbf{x}\right)\left(\mathrm{y} - \mathbf{x}^T \boldsymbol{\theta}\right)\right] \\ &= \nabla \left\{\mathbb{E}[\mathrm{y}^2] - 2\boldsymbol{\theta}^T\, \mathbb{E}[\mathbf{x}\mathrm{y}] + \boldsymbol{\theta}^T\, \mathbb{E}[\mathbf{x}\mathbf{x}^T]\boldsymbol{\theta}\right\} \\ &= -2\boldsymbol{p} + 2\Sigma_x \boldsymbol{\theta} = \mathbf{0}\end{aligned}$$

$$\boxed{\Sigma_x \boldsymbol{\theta}_* = \boldsymbol{p} : \quad \text{Normal Equations,}}$$

People
Innovation
Excellence

People
Innovation
Excellence

- where the input-output cross-correlation vector p is given by

$$\boldsymbol{p} = \left[\mathbb{E}[\mathrm{x}_1\mathrm{y}], \ldots, \mathbb{E}[\mathrm{x}_l\mathrm{y}]\right]^{\mathrm{T}} = \mathbb{E}[\mathbf{x}\mathrm{y}],$$

- and the respective covariance matrix is given by

$$\Sigma_x = \mathbb{E}\left[\mathbf{x}\mathbf{x}^{\mathrm{T}}\right].$$

- Thus, the weights of the optimal linear estimator are obtained via a linear system of equations, provided that the covariance matrix is **positive definite** and hence it can be inverted. Moreover, in this case, the solution is *unique*.
- On the contrary, if Σ_x is singular and hence cannot be inverted, there are infinitely many solutions.

The Cost Function Surface

- Elaborating on the cost function, J(θ), as it is defined before, we get that $\boldsymbol{J(\theta)} = \sigma_y^2 - 2\theta^T p + \theta^T \Sigma_x \theta$.
- Adding and subtracting the term $\theta_*^T \Sigma_x \theta_*$ and taking into account the definition of θ_* from the normal equation, it is readily seen that

$$\boxed{J(\boldsymbol{\theta}) = J(\boldsymbol{\theta}_*) + (\boldsymbol{\theta} - \boldsymbol{\theta}_*)^{\mathrm{T}} \Sigma_x (\boldsymbol{\theta} - \boldsymbol{\theta}_*),}$$

- Where

$$J(\boldsymbol{\theta}_*) = \sigma_y^2 - \boldsymbol{p}^{\mathrm{T}} \Sigma_x^{-1} \boldsymbol{p} = \sigma_y^2 - \boldsymbol{\theta}_*^{\mathrm{T}} \Sigma_x \boldsymbol{\theta}_* = \sigma_y^2 - \boldsymbol{p}^{\mathrm{T}} \boldsymbol{\theta}_*,$$

Is the minimum achieved at the optimal solution.

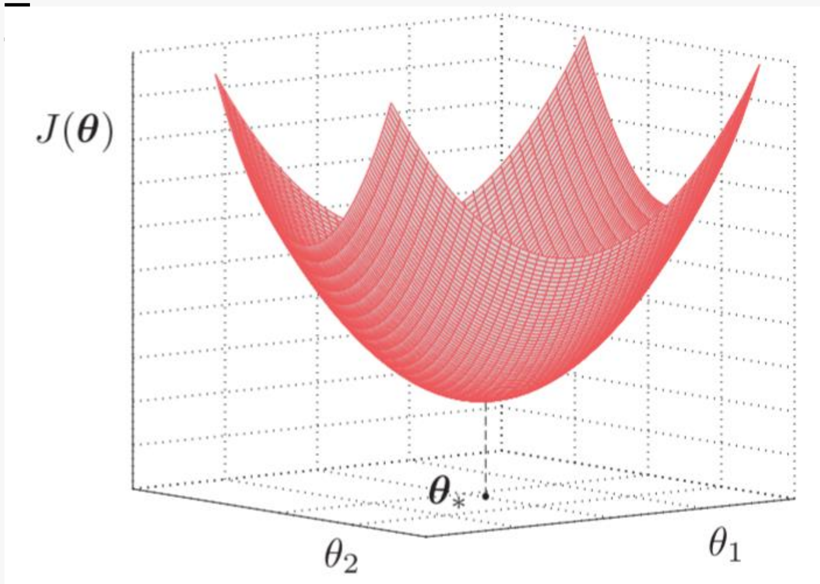
People
Innovation
Excellence

The Cost Function Surface (continue)

- The cost at the optimal value $\boldsymbol{\theta}_*$ is always less than the variance $E[y^2]$ of the output variable. This is guaranteed by the positive definite nature of Σ_x or Σ_x^{-1}, which makes the second term on the right-hand side always positive, unless **p = 0**; However, the cross-correlation vector will only be zero if x and y are uncorrelated.
- In this case, one cannot say anything (make any prediction) about *y* by observing samples of *x*, at least as far as the MSE criterion is concerned, which turns out to involve information residing up to the second order statistics.
- In this case, the variance of the error, which coincides with $J(\theta_*)$, will be equal to the variance σ_y^2 ; the latter is a measure of the "intrinsic" uncertainty of *y* around its (zero) mean value.
- On the contrary, if the input-output variables are correlated, then observing *x* removes part of the uncertainty associated with *y*.

The Cost Function Surface (continue)

- For any value θ, other than the optimal θ_* ,the error variance increases as suggests, due to the positive definite nature of

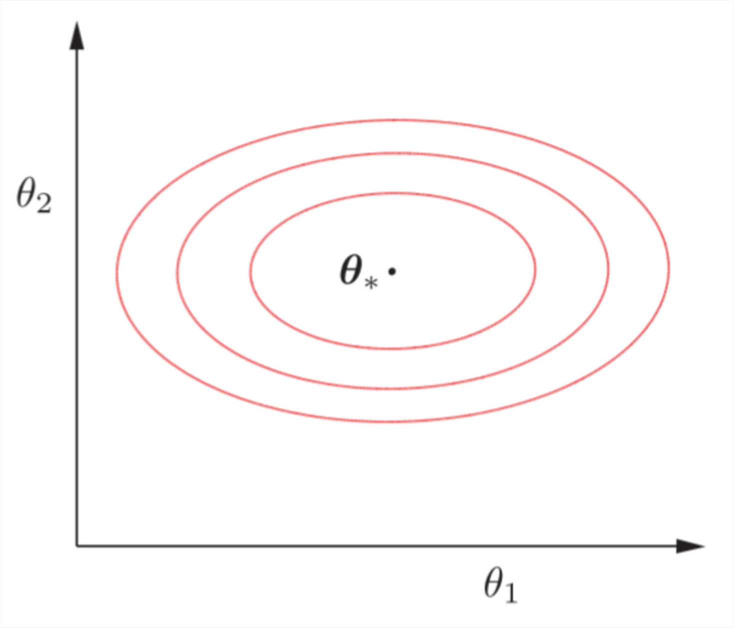


- The figure shows the cost function (mean-square error) surface defined by J(θ).

People
Innovation
Excellence

The Cost Function Surface (continue)

- The corresponding isovalue contours are shown in figure below. In general, they are ellipses, whose axes are determined by the eigenstructure of Σ_x. For $\Sigma_x = \sigma^2 I$, where all eigenvalues are equal to σ^2, the contours are circles



People
Innovation
Excellence

A Geometric Viewpoint: Orthogonality Condition

- What we have discussed so far comes from the geometric interpretation of the random variables.
- The set of random variables is a vector space over the field of real (and complex) numbers.
- If x and y are any two random variables then x + y, as well as αx, are also random variables for every $\alpha \in \mathbb{R}$
- this vector space equipped with an inner product operation, which also implies a norm and makes it a **Euclidean space**.
- The mean value operation has all the properties required for an operation to be called an inner product.

- Indeed, for any subset of random variables

- $\mathbb{E}[xy] = \mathbb{E}[yx]$,
- $\mathbb{E}[(\alpha_1 x_1 + \alpha_2 x_2)y] = \alpha_1 \mathbb{E}[x_1 y] + \alpha_2 \mathbb{E}[x_2 y]$,
- $\mathbb{E}[x^2] \geq 0$, with equality if and only if $x = 0$.

- Thus, the norm induced by this inner product $||x||$ coincides with the respective standard deviation (assuming $E[x] = 0$).

$$\|x\| := \sqrt{\mathbb{E}[x^2]}.$$

- Given two uncorrelated random variables, x, y, or E[xy] = 0, we can call them **orthogonal**, because their inner product is zero.
- We are now free to apply to our task of interest any one of the theorems that have been derived for Euclidean spaces.
- Let us now rewrite the equation

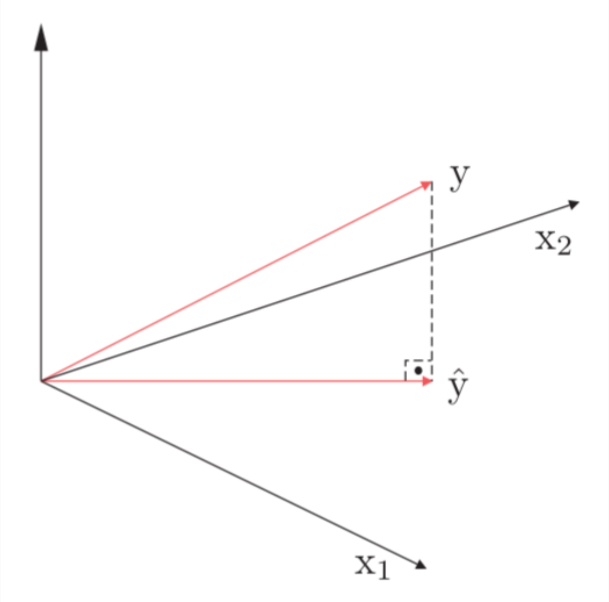
$$\hat{y} = \boldsymbol{\theta}^{T}\mathbf{x}, \longrightarrow \hat{y} = \theta_1 x_1 + \cdots + \theta_l x_l.$$

- Thus, the random variable, ŷ, which is now interpreted as a point in a vector space, results as a linear combination of *l* elements in this space.
- Thus, the estimate, ŷ, will necessarily lie in the subspace spanned by these points. In contrast, the true variable, *y*, will not lie, in general, in this subspace.

- Because our goal is to obtain a ŷ that is a good approximation of *y*, we have to seek the specific linear combination that makes the norm of the error, $e = y - \hat{y}$, minimum.
- This specific linear combination corresponds to the **orthogonal** projection of *y* onto the subspace spanned by the points $x_1, x_2, \ldots, x_l$. This is equivalent with requiring

$$\boxed{\mathbb{E}[\mathrm{e}\mathrm{x}_k] = 0, \quad k = 1, \ldots, l: \quad \text{Orthogonality Condition.}}$$

- The error variable being orthogonal to every point x_k, k = 1, 2, … , *l*, will be orthogonal to the respective subspace.



- Such a choice guarantees that the resulting error will have the minimum norm; by the definition of the norm, this corresponds to the minimum MSE, or $E[e^2]$.

- The set of Orthogonality Condition equations can now be written as:

$$\mathbb{E}\left[\left(\mathrm{y} - \sum_{i=1}^{l} \theta_i \mathrm{x}_i\right) \mathrm{x}_k\right] = 0, \quad k = 1, 2, \ldots, l,$$

Or

$$\sum_{i=1}^{l} \mathbb{E}[\mathrm{x}_i \mathrm{x}_k] \theta_i = \mathbb{E}[\mathrm{x}_k \mathrm{y}], \quad k = 1, 2, \ldots, l,$$

Which leads to Normal equations. Another name is Wiener-Hopf equations.

Case Study

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

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

- From the parameter estimated in the last session, discuss and give the overview of the cost function.

End of Session 13&14

BINUS UNIVERSITY

References

- Sergios Theodoridis. (2015). *Machine Learning: a Bayesian and Optimization Perspective*. Jonathan Simpson. ISBN: 978-0-12-801522-3. Chapter 4.
- https://www.kaggle.com/jojoker/singapore-airbnb

People
Innovation
Excellence