BINUS UNIVERSITY

People
Innovation
Excellence

Course : COMP6577 – Machine Learning
Effective Period : February 2020

Learning in Parametric Modeling 2

Session 11 & 12

Learning Outcome

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

BINUS UNIVERSITY

Outline

- Regularization
- Maximum likelihood method
- Bayesian inference
- Case Study

People
Innovation
Excellence

BINUS UNIVERSITY

People
Innovation
Excellence

Regularization

- **Regularization** is a mathematical tool to impose a priori information on the structure of the solution, which comes as the outcome of an optimization task.
- Regularization was first suggested by the great Russian mathematician Andrey Nikolayevich Tychonoff (sometimes spelled Tikhonov) for the solution of integral equations.
- Sometimes, it is also referred as Tychonoff-Phillips regularization, to honor David Phillips as well, who developed the method independently

BINUS UNIVERSITY

People
Innovation
Excellence

Regularization (2)

- Reformulate the LS minimization task

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

$$\text{subject to:} \quad \|\boldsymbol{\theta}\|^2 \leq \rho,$$

- where $||\cdot||$ stands for the Euclidean norm of a vector. In this way, we do not allow the LS criterion to be completely "free" to reach a solution, but we limit the space in which to search for it.
- Obviously, using different values of ρ, we can achieve different levels of shrinkage.
- The optimal value of ρ cannot be analytically obtained, and one has to experiment in order to select an estimator that results in a good performance.

BINUS UNIVERSITY

People
Innovation
Excellence

Regularization (3)

- For the LS loss function and the constraint used before, the optimization task can equivalently be written as

$$\text{minimize:} \quad L(\boldsymbol{\theta}, \lambda) = \sum_{n=1}^{N} \left(y_n - \boldsymbol{\theta}^T \boldsymbol{x}_n\right)^2 + \lambda \|\boldsymbol{\theta}\|^2 : \quad \text{Ridge Regression.}$$

- It turns out that, for specific choices of $\lambda \geq 0$ and ρ, the two tasks are equivalent. Note that this new cost function, $L(\theta, \lambda)$, involves one term that measures the model misfit and a second one that quantifies the size of the norm of the parameter vector.

BINUS UNIVERSITY

Regularization (4)

- Taking the gradient of *L* in the previous equation with respect to θ and equating to zero, we obtain the regularized LS solution for the linear regression task:

$$\left(\sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^T + \lambda I\right) \hat{\boldsymbol{\theta}} = \sum_{n=1}^{N} y_n \boldsymbol{x}_n$$

- where I is the identity matrix of appropriate dimensions. The presence of λ biases the new solution away from that which would have been obtained from the unregularized LS formulation. The task is also known as **ridge regression**.

BINUS UNIVERSITY

Regularization (5)

- Ridge regression attempts to reduce the norm of the estimated vector and at the same time tries to keep the sum of squared errors small; in order to achieve this combined goal, the vector components, θi, are modified in such a way so that the contribution in the misfit measuring term, from the less informative directions in the input space, is minimized.

BINUS UNIVERSITY

People
Innovation
Excellence

Regularization (6)

- In practice, the bias parameter, θ_0, is left out from the norm in the regularization term; penalization of the bias would make the procedure dependent on the origin chosen for *y*.
- Indeed, it is easily checked out that adding a constant term to each one of the output values, y_n, in the cost function, would not result in just a shift of the predictions by the same constant, if the bias term is included in the norm.
- Hence, usually, ridge regression is formulated as:

$$\text{minimize} \quad L(\boldsymbol{\theta}, \lambda) = \sum_{n=1}^{N} \left(y_n - \theta_0 - \sum_{i=1}^{l} \theta_i x_{ni} \right)^2 + \lambda \sum_{i=1}^{l} |\theta_i|^2$$

BINUS UNIVERSITY

People
Innovation
Excellence

Maximum Likelihood Method

- ML and LS are two of the major pillars on which parameter estimation is based and new methods are inspired from. The ML method was suggested by Sir Ronald Aylmer Fisher.
- Once more, we will first formulate the method in a general setting, independent of the regression/classification tasks.
- We are given a set of say, N, observations, $X = \{x_1, x_2, \ldots, x_N\}$, drawn from a probability distribution. We assume that the joint pdf of these N observations is of a known parametric functional type, denoted as $p(X; \theta)$, where the parameter vector $\theta \in \mathbb{R}^K$ is unknown and the task is to estimate its value. This is known as the **likelihood function** of θ with respect to the given set of observations, X.

BINUS UNIVERSITY

Maximum Likelihood Method (2)

- According to the ML method, the estimate is provided by

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} := \arg\max_{\boldsymbol{\theta} \in \mathcal{A} \subset \mathbb{R}^K} p(\mathcal{X}; \boldsymbol{\theta}) : \quad \text{Maximum Likelihood Estimate.}$$

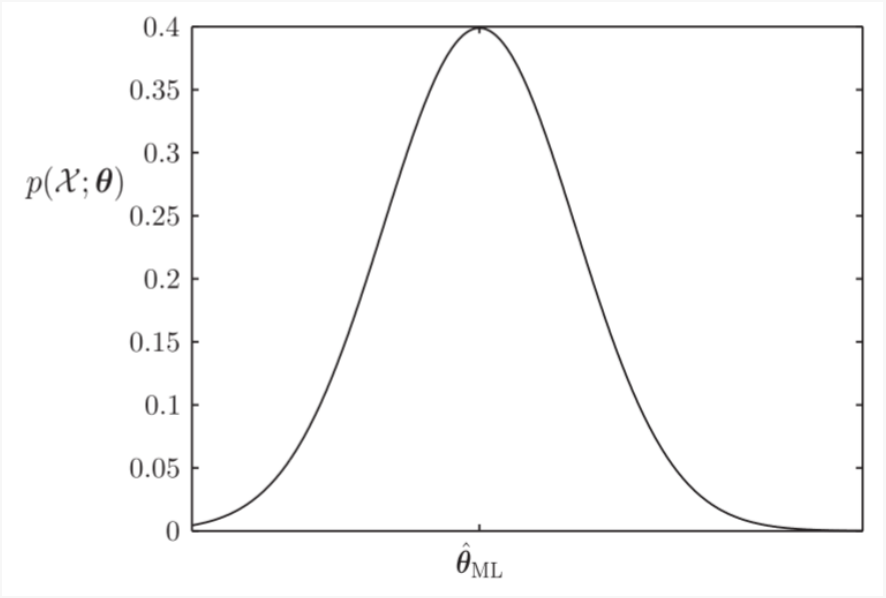
- For simplicity, we will assume that the parameter spa $\mathcal{A} = \mathbb{R}^K$, and that the parameterized family $\{p(X\,;\,\theta) : \theta \ \mathbb{R}^K \ \}$ enjoys a unique minimizer with respect to the parameter θ.

People
Innovation
Excellence

BINUS UNIVERSITY

People
Innovation
Excellence

Maximum Likelihood Method (3)



- According to the maximum likelihood method, we assume that, given the set of observations, the estimate of the unknown parameter is the value that maximizes the corresponding likelihood function.

BINUS UNIVERSITY

People
Innovation
Excellence

- The ML estimator has some very attractive properties, namely:
  - The ML estimator is asymptotically unbiased; that is, assuming that the model of the pdf, which we have adopte here exists a true parameter θ_o, the $\lim_{N\to\infty} \mathbb{E}[\hat{\theta}_{\mathrm{ML}}] = \boldsymbol{\theta}_o$
  - The ML estir ate is asymptotically consistent so that given an $\lim_{N\to\infty} \mathrm{Prob}\left\{\left|\hat{\theta}_{\mathrm{ML}} - \boldsymbol{\theta}_o\right| > \epsilon\right\} = 0$

that is, for large values of N, we expect the ML estimate to be very close to the true value with high probability.

BINUS UNIVERSITY

People
Innovation
Excellence

Bayesian Inference

- In our discussion, so far, we have assumed that the parameter associated with the functional form of the adopted model is a deterministic constant, whose value is unknown to us.
- In this session, we will follow a different rationale. The unknown parameter will be treated as a random variable.
- Hence, whenever our goal is to estimate its value, this is conceived as an effort to estimate the value of a specific realization that corresponds to the observed data.
- As the name Bayesian suggests, the heart of the method beats around the celebrated Bayes theorem. Given two jointly distributed random vectors, say, x, θ, Bayes theorem states that $p(x, \theta) = p(x|\theta)p(\theta) = p(\theta|x)p(x)$

BINUS UNIVERSITY

People
Innovation
Excellence

Bayesian Inference (2)

- Assume that x, θ are two statistically dependent random vectors. Let $\mathcal{X} = \{\mathbf{x}_n \in \mathbb{R}^l, n = 1, 2, \ldots, N\}$, be the set of the observations resulting from *N* successive experiments. Then, Bayes theorem gives

$$p(\boldsymbol{\theta}|\mathcal{X}) = \frac{p(\mathcal{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{X})} = \frac{p(\mathcal{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathcal{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}}$$

- Obviously, if the observations are i.i.d., then we can write

$$p(\mathcal{X}|\boldsymbol{\theta}) = \prod_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{\theta})$$

- In the formulas, p(θ) is the a priori pdf concerning the statistical distribution of θ, and p(θ|*X*) is the conditional or a posteriori pdf, formed after the set of *N* observations has been obtained. The prior probability density, p(θ), can be considered as a constraint that encapsulates our prior knowledge about θ.

BINUS UNIVERSITY

People
Innovation
Excellence

Bayesian Inference (3)

- If the adopted assumptions about the underlying models are sensible, we expect the posterior pdf to be a more accurate one to describe the statistical nature of θ.
- We will refer to the process of approximating the pdf of a random quantity, based on a set of training data, as ***inference***, to differentiate it from the process of estimation, that returns a single value for each parameter/variable.
- So, according to the inference approach, one attempts to draw conclusions about the nature of the randomness that underlies the variables of interest. This information can in turn be used to make predictions and to take decisions.

BINUS UNIVERSITY

People
Innovation
Excellence

Case Study

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

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

1. What do you think about regularization function to solve overfitting problem. (Clue: Sergios Theodoridis. (2015). Chapter 3, Section 3.8)

2. From the same dataset that we use in the case study, estimate the parameter using one of the techniques discussed in this session (regularization, Maximum likelihood or Bayesian inference method)

End of Session 11 & 12

BINUS UNIVERSITY

People
Innovation
Excellence

References

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