BINUS UNIVERSITY

People
Innovation
Excellence

Course : COMP6577 – Machine Learning
Effective Period : February 2020

Introduction to Machine Learning 2

Session 03 & 04

BINUS UNIVERSITY

Learning Outcome

- LO1 : Student be able to explain the fundamental of machine learning concept

People
Innovation
Excellence

BINUS UNIVERSITY

Outline

- Classification
- Regression
- Machine learning challenges
- Testing and Validating
- Case Study

People
Innovation
Excellence

BINUS UNIVERSITY

People
Innovation
Excellence

Classification

- The goal in classification is to assign an unknown pattern to one out of a number of classes that are considered to be known.
- For example, in X-ray mammography, we are given an image where a region indicates the existence of a tumor. The goal of a computer-aided diagnosis system is to predict whether this tumor corresponds to the benign or the malignant class.
- Optical character recognition (OCR) systems are also built around a classification system, in which the image corresponding to each letter of the alphabet has to be recognized and assigned to one of the twenty-four (for the Latin alphabet) classes

BINUS UNIVERSITY

People
Innovation
Excellence

- The first step in designing any machine learning task is to decide how to represent each pattern in the computer.
- This is achieved during the preprocessing stage:
  - "encode" related information that resides in the raw data in an efficient and information-rich way.
  - By transforming the raw data in a new space with each pattern represented by a vecto $\boldsymbol{x} \in \mathbb{R}^l$ **(feature vector)**
  - Its *l* elements are known as features.
  - Each pattern becomes a single point in an *l*-dimensional space, known as the **feature space** or the **input space. (feature generation stage)**

BINUS UNIVERSITY

People
Innovation
Excellence

General classification process

- Having decided upon the input space, in which the data are represented, one has to train a classifier. This is achieved by first selecting a set of data whose class is known, which comprises the training set.
  - set of pairs, (y_n, x_n), n = 1, 2, ... , N
  - y_n is the (output) variable denoting the class in which x_n belongs
  - the class labels, y, take values over a discrete set, {1, 2, ... , M} for M-class classification task
- Then designs a function **(classifier)** , f , which predicts the output label given the input.

BINUS UNIVERSITY

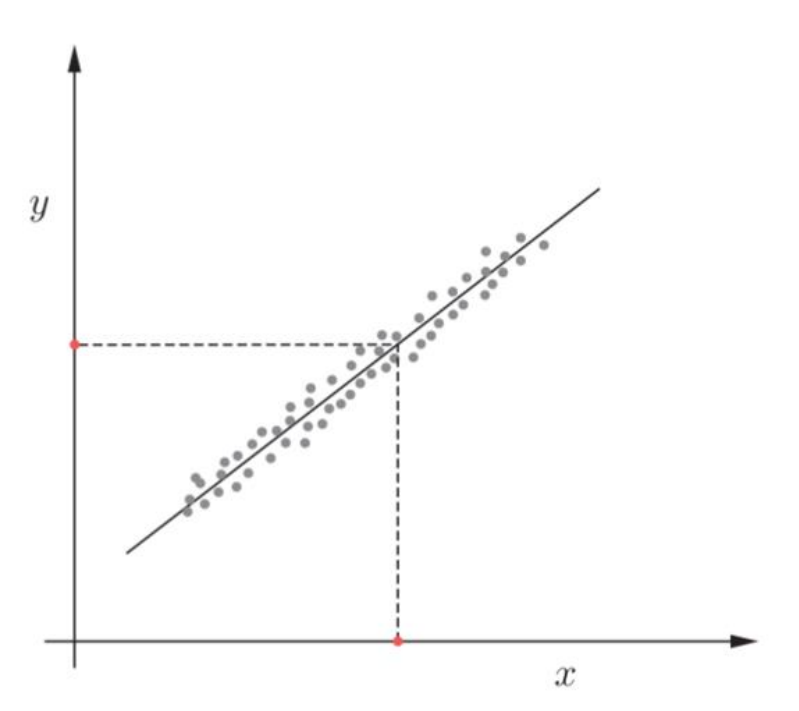
People
Innovation
Excellence

Regression

- The regression shares to a large extent the feature generation/selection stage, as described before; however, now the output variable, *y*, is not discrete but it takes values in an interval in the real axis or in a region in the complex numbers plane.
- The regression task is basically a curve fitting problem.
- We are given a set of training points, (y_n, x_n), $y_n \in R$, $x_n \in R^l$, $n = 1, 2, \ldots, N$
- The task is to estimate a function, f, whose graph fits the data.
- Once we have found such a function, when an unknown point arrives, we can predict its output value.

BINUS UNIVERSITY

People
Innovation
Excellence



- Once a function (linear in this case), *f* , has been designed, for its graph to fit the available training data set in a regression task, given a new (red) point, *x*, the prediction of the associated output (red) value is given by *y = f(x)*

BINUS UNIVERSITY

Other example of Regression Task

- In financial applications one can predict tomorrow's stock market price given current market conditions and all other related information. Each piece of information is a measured value of a corresponding feature.
- Signal and image de-noising can also be seen as a special type of a regression task. De-blurring is a typical image restoration task, where the de-blurred image is obtained as the output by feeding the blurred one as input to an appropriately designed function.

People
Innovation
Excellence

BINUS UNIVERSITY

People
Innovation
Excellence

Main Challenges of Machine Learning

1. Bad Data
   - Example:
     - Insufficient Quantity of Training Data
     - Nonrepresentative Training Data
     - Poor-Quality Data
     - Irrelevant Features (involves: Feature selection, Feature extraction, Creating new features by gathering new data)
2. Bad Algorithm
   - Example:
     - Overfitting the Training Data
     - Underfitting the Training Data

BINUS UNIVERSITY

Insufficient Quantity of Training Data

- Machine Learning takes a lot of data for most Machine Learning algorithms to work properly. Even for very simple problems you typically need thousands of examples, and for complex problems such as image or speech recognition you may need millions of examples (unless you can reuse parts of an existing model).

BINUS UNIVERSITY

Nonrepresentative Training Data

- In order to generalize well, it is crucial that the training data be representative of the new cases you want to generalize to. This is true whether you use instance-based learning or model-based learning.
- If the sample is too small, you will have sampling noise (i.e., nonrepresentative data as a result of chance), but even very large samples can be nonrepresentative if the sampling method is flawed. This is called sampling bias.

BINUS UNIVERSITY

People
Innovation
Excellence

Poor-Quality Data

- If your training data is full of errors, outliers, and noise (e.g., due to poor-quality measurements), it will make it harder for the system to detect the underlying patterns.
- It is often well worth the effort to spend time cleaning up your training data.
- For example:
  - If some instances are clearly outliers, it may help to simply discard them or try to fix the errors manually.
  - If some instances are missing a few features (e.g., 5% of your customers did not specify their age), you must decide whether you want to ignore this attribute altogether, ignore these instances, fill in the missing values (e.g., with the median age), or train one model with the feature and one model without it, and so on.

BINUS UNIVERSITY

People
Innovation
Excellence

Irrelevant Features

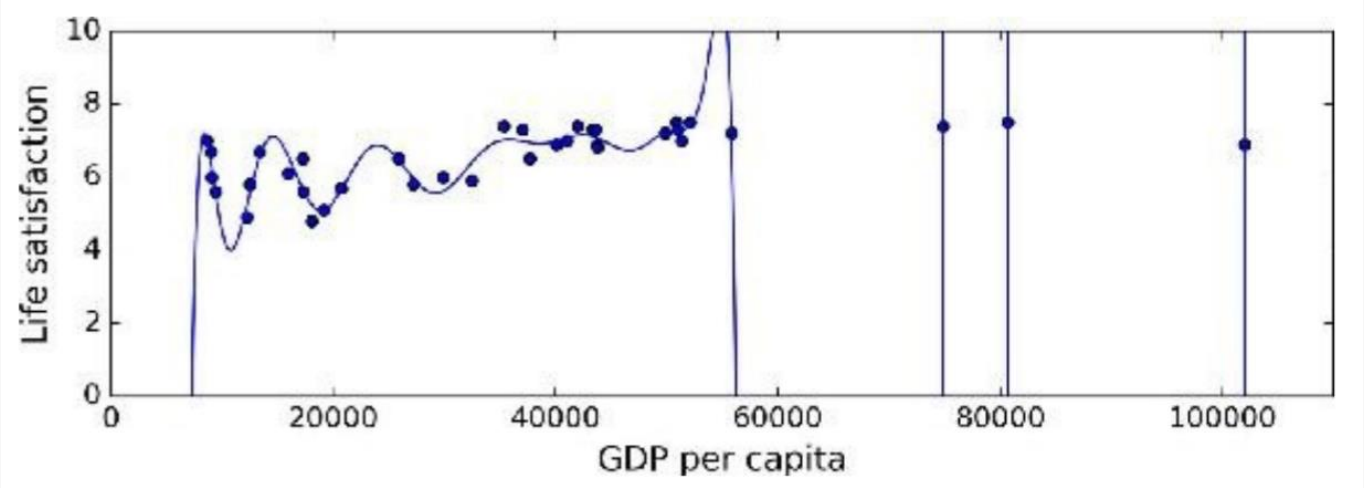
- A critical part of the success of a Machine Learning project is coming up with a good set of features to train on **(feature engineering)**.
- feature engineering, involves:
  - **Feature selection**: selecting the most useful features to train on among existing features.
  - **Feature extraction**: combining existing features to produce a more useful one (as we saw earlier, dimensionality reduction algorithms can help).
  - Creating new features by gathering new data.

BINUS UNIVERSITY

People
Innovation
Excellence

Overfitting the Training Data

- Overfitting means that the model performs well on the training data, but it does not generalize well.
- Example: a high-degree polynomial life satisfaction model that strongly overfits the training data. Even though it performs much better on the training data than the simple linear model, would you really trust its predictions?



BINUS UNIVERSITY

People
Innovation
Excellence

Underfitting the Training Data

- Underfitting is the opposite of overfitting: it occurs when your model is too simple to learn the underlying structure of the data.
- For example, a linear model of life satisfaction is prone to underfit; reality is just more complex than the model, so its predictions are bound to be inaccurate, even on the training examples.
- The main options to fix this problem are:
  - Selecting a more powerful model, with more parameters
  - Feeding better features to the learning algorithm (feature engineering)
  - Reducing the constraints on the model (e.g., reducing the regularization hyperparameter)

BINUS UNIVERSITY

People
Innovation
Excellence

Testing and Validating

- Once you have trained a model, you don't want to just "hope" it generalizes to new cases. You want to evaluate it, and fine-tune it if necessary.
- Split your data into two sets: the **training set** and the **test set.** Train your model using the training set, and test it using the test set.
- The error rate on new cases is called the generalization error (or out-of-sample error), and by evaluating the model on the test set, an estimation of this error is gotten. This value tells how well the model will perform on instances it has never seen before.
- If the training error is low (i.e., the model makes few mistakes on the training set) but the generalization error is high, it means that the model is overfitting the training data.

BINUS UNIVERSITY

People
Innovation
Excellence

Testing and Validating (2)

- Let's say that you measured the generalization error multiple times on the test set, and adapted the model and hyperparameters to produce the best model for that set. This means that the model is unlikely to perform as well on new data. A common solution to this problem is to have a second holdout set called the **validation set**.
- Train multiple models with various hyperparameters using the training set, select the model and hyperparameters that perform best on the validation set, and when you're happy with the model then run a single final test against the test set to get an estimate of the generalization error.

BINUS UNIVERSITY

People
Innovation
Excellence

Cross-validation

- To avoid “wasting” too much training data in validation sets, a common technique is to use **cross-validation**:
  - the training set is split into complementary subsets, and each model is trained against a different combination of these subsets and validated against the remaining parts. Once the model type and hyperparameters have been selected, a final model is trained using these hyperparameters on the full training set, and the generalized error is measured on the test set.

BINUS UNIVERSITY

Review Question

1. Can you name four of the main challenges in Machine Learning?
2. What is a test set and why would you want to use it?
3. What is the purpose of a validation set?
4. What is cross-validation and why would you prefer it to a validation set?

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

- From the link above we can define classification and regression case. For example classification case to identify room type or region. While for regression we can identify price of room or review score each month. You can also do clustering from the data above.

End of Session 03 & 04

BINUS UNIVERSITY

References

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