Importance of Training, Testing and Generalization Error

Mehul - A023119820011

Raghav - A023119820007

- Training data for Machine Learning (ML) is a key input to algorithm that comprehend from such data and memorize the information for future prediction.

- Training data is a backbone of entire AI and ML project without that it is not possible to train a machine that learns from humans and predict for humans.

An Organized form of Unorganized Data

- Data collected from multiple sources are usually available in unorganized format, which is not useful for machines to ingest the useful information. But when such data is labeled or tagged with annotation it becomes a well-organized data that can be used to train the AI or ML model.
- And annotated or labeled data helps machines through computer vision to detect various objects from the group and store the information for future reference.

Recognition and Classification of Objects

- Another most important role of training data for machine learning is classifying the data sets into various categorized which is very much important for supervised machine learning.

- When your algorithm learns what are the features are important in distinguishing between two classes. It helps them to recognize and classify the similar objects in future, thus training data is very important for such classification.

Provides a Key Input to ML Algorithms

- To work with ML algorithm, we need certain inputs making our model understand the things in its own way. And training data is the only source, you can use as an input into your algorithms, that will help your AI model to gain the useful information from the data and take crucial decisions like human intelligence do.
- In a supervised machine learning, an additional input of labeled training data is required. And when your training data is not properly labeled, its not worth for supervised machine learning.

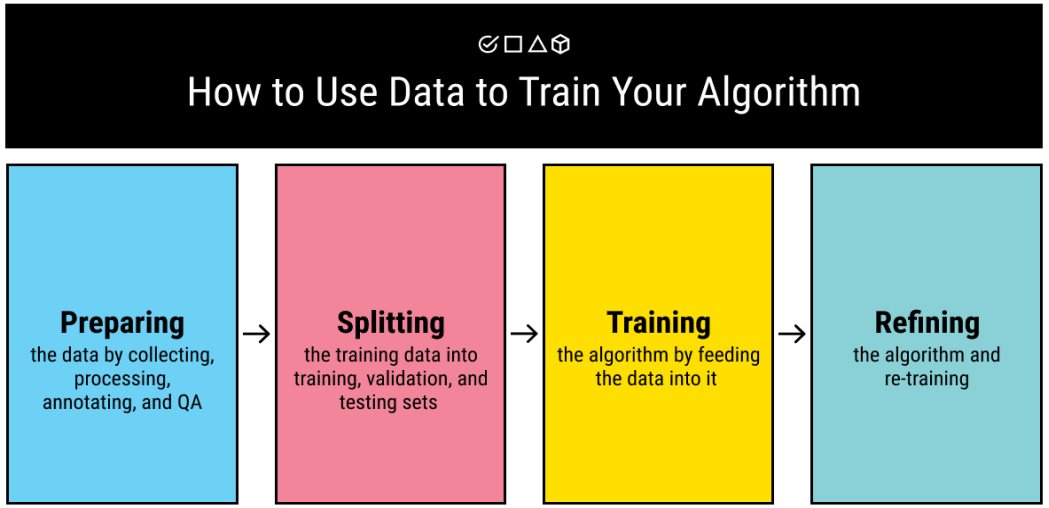
Validating the Machine Learning Model

- To validate or evaluate such AI model you need another set of training data which can be also called the validation data, use to check the accuracy level of model in different scenario.
- During ML model validation, labeled data is used to cross-check, whether machines has correctly detected the object or not. Training data is already labeled and if machine is unable to recognize the object, means either your labeled data is not right or algorithm is not capable to train your model in recognizing such things precisely. Once, you have checked the output given by machine you have to validate if its correct or not.

Testing Data

- After the machine learning model is built, you need to check its work. The AI platform uses testing data to evaluate the performance of your model and adjust or optimize it for better forecasts.
- **Unseen.** You cannot reuse the same information that was in the training set.
- **Large.** The data set should be large enough so that the machine can make predictions.
- **Representative.** The data should represent the actual dataset.

- The evaluation process in such systems is called the blind test. When building a model, AI splits the data in a ratio of about 70% to 30%, where the first figure is training data and the second is testing.



- After the model is built and tested, the machine calculates a special index that represents the quality of the model. Thus, users can decide to use this model for scoring or create another one.

Generalization Error

- In supervised learning applications in machine learning and statistical learning theory, generalization error (also known as the out-of-sample error) is a measure of how accurately an algorithm is able to predict outcome values for previously unseen data.
- Generalization error can be minimized by avoiding overfitting in the learning algorithm. The performance of a machine learning algorithm is visualized by plots that show values of *estimates* of the generalization error through the learning process, which are called learning curves

- An important way to understand generalization error is **bias-variance decomposition**.
- **bias** is the **error rate** in the world of big data. A model has a high bias when, for example, it fails to capture meaningful patterns in the data. Bias is measured by the differences between the *expected predicted values* and the *observed values*,
- In contrast with bias, **variance** is an algorithm's **flexibility** to learn patterns in the observed data. **Variance** is the amount that an algorithm will change if a **different dataset** is used. A model is of high variance when, for instance, it tries too hard that it not only captures the pattern of meaningful features but also that the meaningless error (**overfitting**).

D: the training dataset

x: our sample

y: the **real** values of the outcome variable

y_D: the **observed** values of the outcome variable **in dataset D**

$f(x; D)$: the **fitted** values of the outcome variable, i.e. the output of our model when input=x, and the model is learned from dataset D

E_D: the expectation on dataset D

Error(f;D): the generalization error of model f trained on dataset D

Generalization error

$$Error(f; D) = E[(f(x; D) - y_D)^2]$$

where

$$\overline{f(x)} = E_D[f(x; D)]$$

bias could be denoted as

$$bias^2(x) = (\overline{f(x)} - y_D)^2$$

Variance be

$$var(x) = E_D[(f(x; D) - \overline{f(x)})^2]$$

And **noise**

$$\epsilon^2 = E_D[(y_D - y)^2]$$

For noise, we have "zero assumption": the expectation of noise is zero.

$$E_D[y_D - y] = 0$$

$$E_D[(f(x;D) - y_D)^2] = E_D[(f(x;D) - \overline{f(x)} + \overline{f(x)} - y_D)^2]$$

$$= E_D[(f(x;D) - \overline{f(x)})^2] + E_D[(\overline{f(x)} - y_D)^2] + 2E_D[(f(x;D) - \overline{f(x)})(\overline{f(x)} - y_D)]$$

$$= E_D[(f(x;D) - \overline{f(x)})^2] + E_D[(\overline{f(x)} - y_D)^2]$$

$$= var(x) + E_D[(\overline{f(x)} - y + y - y_D)^2]$$

$$= var(x) + E_D[(\overline{f(x)} - y)^2] + E_D[(y - y_D)^2] + 2E_D[(\overline{f(x)} - y)(y - y_D)]$$

$$= var(x) + E_D[(\overline{f(x)} - y)^2] + E_D[(y - y_D)^2]$$

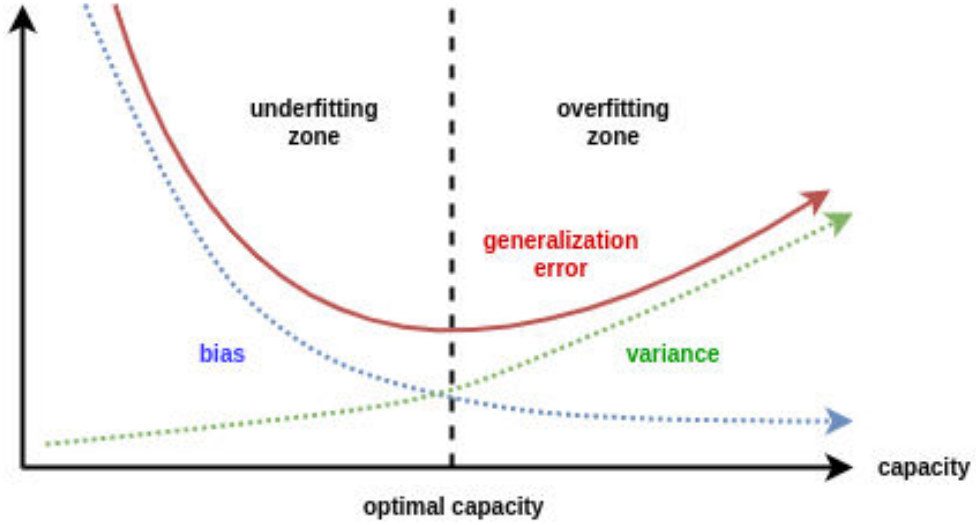
$$= var(x) + (\overline{f(x)} - y)^2 + E_D[(y - y_D)^2]$$

$$= var(x) + bias^2(x) + \epsilon^2$$

Interpretation

- Bias measures the deviation between the expected output of our model and the real values, so it indicates the **fit of our model**.
- Variance measures the amount that the outputs of our model will change if a different dataset is used. It is the impacts of using different datasets.
- Noise is the irreducible error, the **lowest bound of generalization error** for the current task that any model will not be able to get rid of, indicating the difficulty of this task.
- These 3 components above determine the model's ability to react to new unseen data rather than just the data that it was trained on.

- Generalization error could be measured by MSE. As the model capacity increases, the bias decreases as the model fits the training datasets better.



- However, the variance increases, as your model become sophisticated to fit more patterns of the current dataset, changing datasets (even if they come from the same distribution) would be impactful