Machine Learning In Python

Subject : Classification Using Naïve Bayes, Overfitting vs. Underfitting

Lecturer : Reza Akbari Movahed

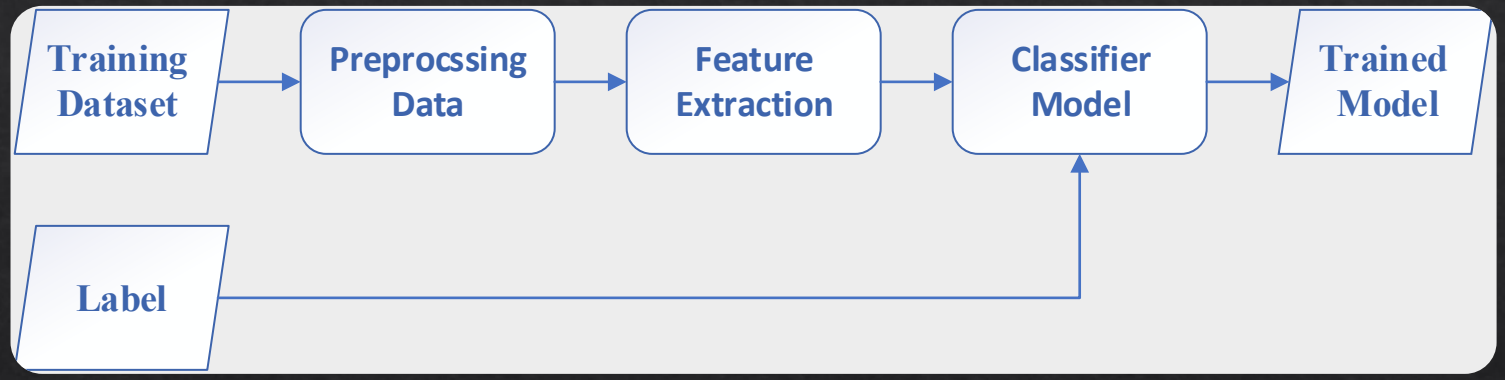
Hamedan University of Technology

Spring 2020

Classification Using Naïve Bayes

Classification In Supervised Learning Framework

Training Phase



Testing Phase

Classification Using Naïve Bayes

Classifier Models :

Naïve Bayes

- Naïve Bayes is a supervised classification algorithm which uses the Bayes' theorem for classification performance.
- Naïve Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naïve) independence assumptions between the features.

Bayes' theorem

- $P(C_k|X)$: Posterior Probability
- $P(X|C_k)$: Likelihood
- $P(X)$: Prior Probability

$$P(C_k|X) = \frac{P(C_k)P(X|C_k)}{P(X)}$$

$$\mathrm{X} = [\mathrm{x}_1, \mathrm{x}_2, \dots, \mathrm{x}_\mathrm{d}]$$

Classification Using Naïve Bayes

How to train and test a Naïve Bayes classifier?

- Suppose we have a binary classification problem (*k=1, 2*).
- Training:
  - Two probability distributions is fitted to the training samples of each class ($P(C_1|X)$, $P(C_2|X)$).
- Testing:
  - The testing sample is given to each probability distribution in order to generate probabilities belongs to each class ($P(C_1|X_T)$, $P(C_2|X_T)$).

$P(C_1|X_T) > P(C_2|X_T) \rightarrow X_T$ is classified to the C_1 by Naïve Bayes model

$P(C_2|X_T) > P(C_1|X_T) \rightarrow X_T$ is classified to the C_2 by Naïve Bayes model

Classification Using Naïve Bayes

How to train and test a Naïve Bayes classifier?

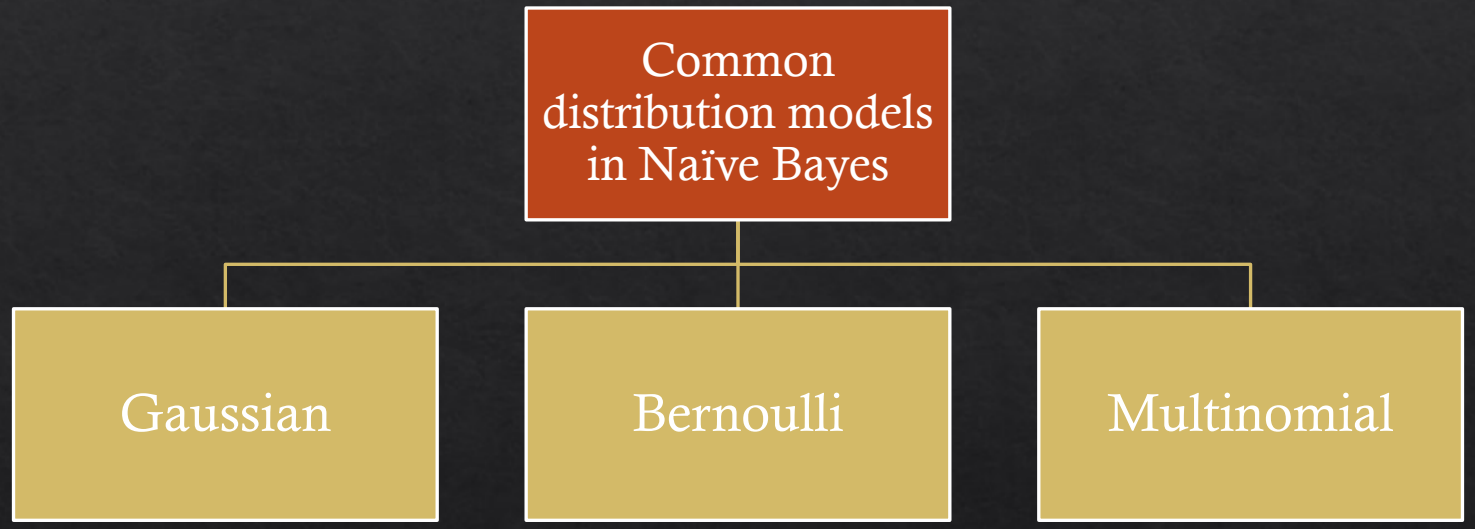
- Suppose we have a binary classification problem (*k=1, 2*).
- Training:
  - Two probability distributions is fitted to the training samples of each class ($P(C_1|X)$, $P(C_2|X)$).
- Testing:
  - The testing sample is given to each probability distribution in order to generate probabilities belongs to each class ($P(C_1|X_T)$, $P(C_2|X_T)$).

$P(C_1|X_T) > P(C_2|X_T) \rightarrow X_T$ is classified to the C_1 by Naïve Bayes model

$P(C_2|X_T) > P(C_1|X_T) \rightarrow X_T$ is classified to the C_2 by Naïve Bayes model

Classification Using Naïve Bayes

Probability Distribution Models in Naïve Bayes



Overfitting vs. Underfitting

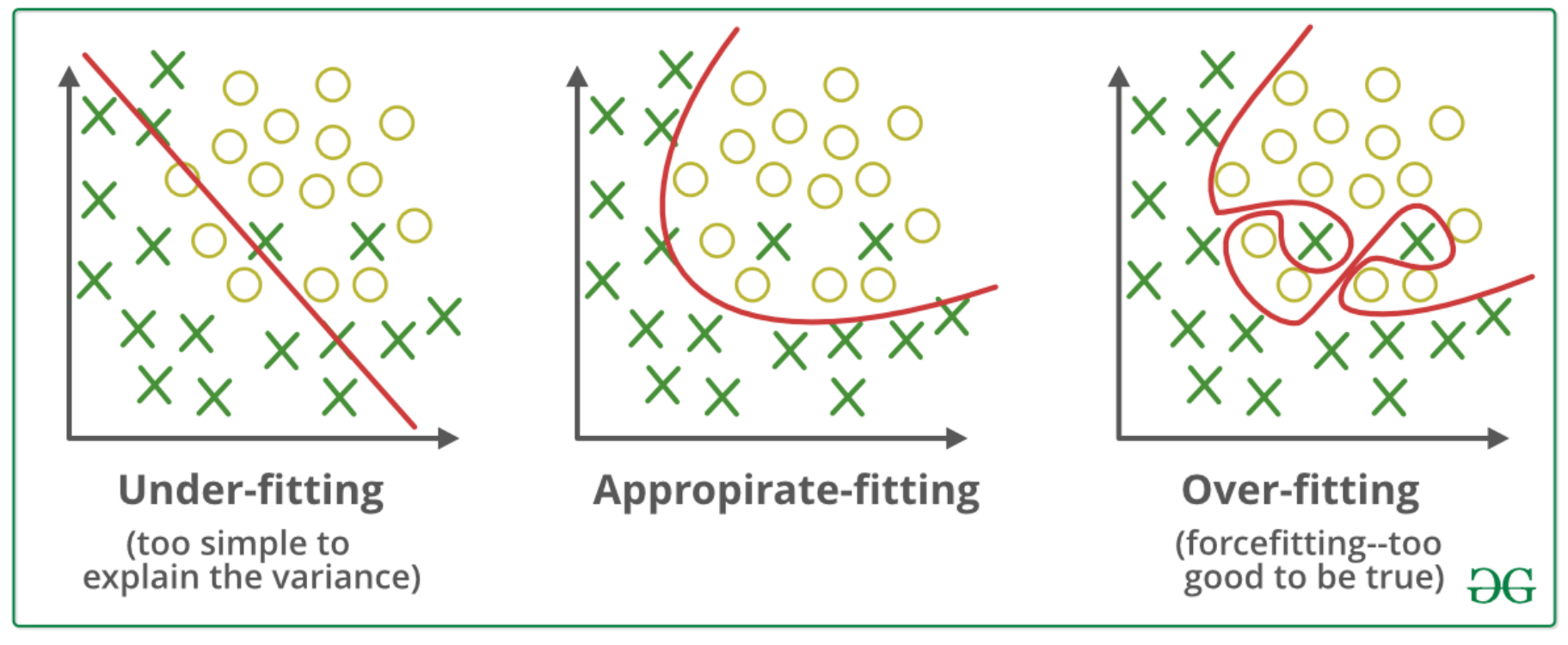
Overfitting:

- Overfitting refers to a model that models the training data too well.
- Overfitting happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data.
- This means that the noise or random fluctuations in the training data is picked up and learned as concepts by the model.
- The problem is that these concepts do not apply to new data and negatively impact the models ability to generalize.

Underfitting:

- Underfitting refers to a model that can neither model the training data nor generalize to new data.
- An underfit machine learning model is not a suitable model and will be obvious as it will have poor performance on the training data.

Overfitting vs. Underfitting



Overfitting vs. Underfitting

The main reasons of causing Overfitting:

- The high number of features of the dataset.
- Simplicity of data.
- The high number of training samples