Elements of Machine Learning

Assigment 1 - Problem 1

Sangeet Sagar(7009050), Philipp Schuhmacher(7010127)
{sasa00001,phsc00003}@stud.uni-saarland.de

November 11, 2021

1 Principles of statistical learning

Explain in short and concise words the concepts of these pairs of terms:

1. unsupervised and supervised learning
2. prediction and inference
3. classification and regression
4. training and test data
5. parametric and non-parametric models

You are allowed to use a maximum of 300 words for this exercise, every ten more words will lead to losing one point.

Solution:

1. In supervised learning, both input and output data points (x_i, y_i) are supplied to build a model or learn a function that can be used to map new data points. Linear regression and logistic regression are commonly used supervised learning algorithms.
   However, in unsupervised learning, we are only given input data points $x_1, x_2, \ldots, x_n$, i.e. there is no information on the relationship between input and output variables, and the goal is to discover patterns or groupings in the data points. Clustering is a popular algorithm of this kind that is used to group similar data points.

2. In prediction we use estimated function $\hat{f}$ learned on the training data, to predict output to unseen data points. In other words, output variable Y is predicted using $\hat{Y} = \hat{f}(X)$.
   Inference is about understanding the relationship between X and Y rather than estimating f, and the goal is to understand how the output is affected as the input variables change.

3. Classification is used to categorize a data point into a class or label. E.g. classifying an email as spam or safe. Regression uses training data to learn a function that best describes the relationship between input and output variables and then predict values for unknown data points.

4. Data used to build (or train) a model or fit a function that holds the relationship between input and output data points is training data. Now the model built on train data must be evaluated to check it's performance on unseen data. This is done using test data. The test data provides an unbiased evaluation of the trained model.

5. Parametric and non-parametric models are two different ways used to estimate $\hat{f}$ that best describes observations (X, Y). The parametric model follows a 2-step approach. We assume a function f and then use training data to estimate the coefficients of the function. For e.g. assume a linear function,

   $$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

   the goal is to estimate these $p + 1$ coefficients $\beta_0, \beta_1, \ldots, \beta_p$. While a non-parametric model does not make any assumption about the function f, it directly estimates f that best fits the training data. This way, it can fit a wide range of possible shapes of f.