Problem Set 6

Nov. 11, 2021

1. Normal distribution of data is an assumption commonly used by many machine learning models. It is important to know whether data are under normal distribution before the training process. Study how to check the "normality" of data and answer the following questions:
   a) QQ-plot is an efficient tool to visually check the similarity between two distributions through. Briefly introduce how to judge normal distribution of data through QQ-plot and sketch an example plot to illustrate.
   b) To quantitatively check the normality of data, we need to perform hypothesis testing. Search for some literature and introduce one method of normality testing.
   c) Search for methods to transform data to normal distribution when the original data are not. Choose one method and briefly introduce how it works.
2. Consider the hypothetic scenario of Covid-19 case tracing. If someone got some minor symptoms of fever or cough and checked negative for Sars-Cov-2 virus, it is still not sufficient to judge whether he is free of Covid-19. A common practice is to query whether he had any possible known cause of the symptoms such as getting cold due to sudden weather change or some event in his recent life, and whether he had any possible overlap with anyone who might be a close contact with Covid-19 patients. Inference is then made about the possible that he got Covid-19. Try to formulate this decision-making procedure using the framework of Bayesian Inference or Bayesian Networks.

Due date: Nov. 17 (Wednesday) 23:00 Beijing time