機器學習導論

Introduction to Machine Learning

Homework 4 Dimensionality Reduction & Clustering

(Thursday, May 19th, 2022)

一、Dimensionality Reduction

”auto-mpg.csv”這筆資料是一份汽車規格的數據集，其中包含 9 個 feature。

mpg: continuous
cylinders: multi-valued discrete
displacement: continuous
horsepower: continuous
weight: continuous
acceleration: continuous
model year: multi-valued discrete
origin: multi-valued discrete
car name: string (unique for each instance)

除了 origin 及 car name 以外，其餘 feature 可以用來預測汽車的續航力(mpg)，請跟隨下列三種方式，探討何種 dimensionality reduction 適合處理此資料集。

1. High Correlation filter
2. Backward Selection
3. PCA(reduce to having up to 95% variance)

完成降維後，請將剩下的 feature，以 linear regression 之方式，預測 mpg，檢視各自的 MSE 及相關係數，並討論各個方法的優缺點。

(Data preprocessing：5%, Each method implement：15*3%, result & discuss：15%)

二、Clustering

Cluster_data.csv 是一個 x-y 二維資料，以散點圖呈現可看出有類似群聚分群的狀況，請利用 K-means cluster 演算法做聚類分群，並套用不同群的顏色呈現在散點圖上，最後解釋你分群數量的依據，並討論你的發現。

(K-means cluster implement：15%, discuss & result presentation：20%)

Note：You must hand craft code of PCA & K-means algorithm(*numpy & pandas is allowed). However, High Correlation filter and Backward Selection is allowed to use existing library