기계학습 2주차 과제

12191656 이채연

HomeWorks #1

Regression (회귀)문제에 대해 설명 하고 실생활에 사용되는 대표적인 응용(application)사례 쓰시오.

Regression에서는 response variable, 즉 label이라는 yi가 discrete(categorical)한 값이 아니라 continuous한 값인 실제 연속적인 실수 값이다. 그래서 Regression은 실수인 single real-valued input xi와 실수인 single real-valued response yi를 가지고 무엇인가를 학습하는 문제이다.
이러한 Regression 문제는 어떤 데이터에 맞는 어떤 모델을 fitting하는 문제라고 많이 얘기를 한다. 그리고 그러한 모델은 straight line(직선)이 될 수도 있고, quadratic function(2차 함수)나 또다른 다차원의 함수가 될 수 있다.
실생활에서 Regression을 다룰 때 high-dimensional input에 대해서 regression을 해야하기 때문에 쉽지 않다는 문제가 있다. 또한 outlier, 즉 실제 값이 아닌 어떤 noise에 의해서 값이 흐트러지거나 변형이 될 수 있다. 그리고 어떤 값들이 연속적으로 부드럽게 있지 않고 넓게 산재해 있는 non-smooth responses 문제 도 많이 발생이 되고 있다. 그래서 이러한 문제들 속에서 올바른 모델을 fitting해야 한다.
실생활에서 Regression을 사용하는 대표적인 응용(application)사례는 market condition들이나 주변정보들을 기반으로 주식시장 값 예측하기, YouTube에서 시청자가 보는 비디오를 기반으로 시청자의 나이 예측하기, 로봇의 3d space에서의 위치(SLAM) 찾아내기, 의료장비의 측정치를 기반으로 전립선의 특이 항원의 양 측정하기, 건물 안 위치에서 기온을 예측하기 등이 있다.

HomeWorks #2

입력 데이터 D 가 주어졌을 때 Unsupervised learning의 대표적인 기술인 K-means clustering의 동작 방법에 대해 설명 하시오.

K-means clustering은 어떤 cluster 개수에 대해서 distribution을 estimation하는 것이다.
K-means clustering의 동작 방법은 두 단계로 나뉜다. 첫 번째 단계의 목적은 먼저 cluster의 개수는 사전에 알 수 없으므로, cluster의 개수를 찾아내는 것이다. 그래서 입력

데이터 D가 주어졌을 때 cluster의 개수인 K를 estimation하면서 확률 분포를 그려보는 방식으로 동작하는데, mode selection에 의해서 그 mode에서 K값을 approximation한다. 이때 사용하는 수식은 K* = argmax P(K|D) 라고 해서 이 density, distribution의 확률 값을 최대화할 수 있는 K값을 찾는다. 이 K 값이 입력 데이터 D에 맞는 cluster의 개수이다.
이어서 두번째 단계에서는 cluster의 개수가 정해졌으면 각 cluster에 대해서 각각의 data point를 하나씩 다 assign한다. 즉, 어떤 데이터가 어느 cluster에 속하는지를 찾는 것이다. 이때 사용하는 수식은 zi* = argmax P(zi = k | xi, D)인데, 여기서 zi은 1부터 k 사이의 값으로, cluster를 나타내는 variable이다. 그리고 이 zi에 대해서 어떤 data point i를 assign하는 것이다. 어떤 data point i에 대해서 각각의 cluster를 찾는 방법은 argmax P(zi = k | xi, D)를 해서 xi와 D가 given으로 주어질 때 이 density를 최대화 할 수 있는 k를 찾아서 그것을 특정 data point i에 대한 cluster인 zi로 나타낸다.