기계학습 및 데이터마이닝 assignment1 report

이름 : 유성민

학과 : 소프트웨어

학년 : 3

학번 : 201520908

문제1.

스크린샷, 조류이(가) 표시된 사진

자동 생성된 설명1) make\_onehot method 구현

스크린샷이(가) 표시된 사진

자동 생성된 설명2) 결과

문제2.

스크린샷이(가) 표시된 사진

자동 생성된 설명

텍스트, 지도이(가) 표시된 사진

자동 생성된 설명<California google map 사진>

스크린샷이(가) 표시된 사진

자동 생성된 설명알 수 있는 사실:

<level\_1은 행정구역내 주택가격의 중앙값이 0.2이상 1.7미만인 경우>

분포를 보면 행정구역내 소득의 중앙값인(Medinc)의 값이 0 ~ 5인 구간과 경도가 -117 ~ -122인 구간에 집중적으로 분포 되어있다. 샌디에고 ~ 샌프란시스코 사이이고 캘리포니아의 중부에 집중적으로 거주하는 것을 알 수 있다. 값 싼 주택의 경우 중부에 주로 위치하는 것을 알 수 있으며 이 주택에 거주하는 시민들의 소득수준이 다른 level에 비해 낮은 것을 쉽게 알 수 있다.

소득 수준이 8 ~ 10인 구간에 위치하면서 값 싼 주택에 거주하는 시민들도 보이지만 소수이다.

전 처리 후 데이터의 수는 9536개로 세 개의 level과 비교해 제일 많은 수의 시민들이 이 구간에 위치하고 있었다.

즉 결론적으로, 값 싼 주택에 사는 주민은 중부에 주로위치, 소득수준 낮음이라고 한마디로 정리할 수 있겠다.

다음 구간을 살펴보면,

스크린샷이(가) 표시된 사진

자동 생성된 설명<level\_2는 주택가격의 중앙값이 1.7이상 2.5미만인 경우>

경도가 -117 ~ -119사이(특히 -118도)와 -121 ~ -123사이(특히 -122도)에 집중적으로 분포 되어있다. 위치는 샌프란시스코를 관통하는 경도와 LA를 관통하는 경도에 집중적으로 위치해 있다.

소득 수준을 보면 2 ~ 6사이 정도에 집중적으로 분포 되어있고, 소득을 기준으로 6이상부터 다수의 outlier가 보이는 것을 관찰할 수 있다. 소득 수준이 높음에도 불구하고 중간 가격의 주택에 거주하는 시민이 값 싼 주택에 거주하거나 비싼 주택에 거주하는 경우보다 많음을 지칭한다.

전체적인 분포는 값 비싼 주택에 거주하는 level\_3과 크게 차이 나지는 않고 분포가 비슷하다.

결과적으로 값이 중간 정도인 주택에 거주하는 시민의 특성으로는 북서부와 남동부에서 집중적으로 거주하고, 소득수준은 2 ~ 6까지 다양하다.

전 처리 후 데이터는 5317개로 셋 중에 두 번째로 컸다.

스크린샷이(가) 표시된 사진

자동 생성된 설명< level\_3는 주택가격의 중앙값이 2.5이상 5미만인 경우>

이는 비싼 가격의 주택이 경도가 -121 ~ -123(특히 -122)과 -119 ~ -117(특히 -118) level\_2와 굉장히 유사한 분포를 보이고 있다. 샌프란시스코, LA를 관통하는 경도에 주로 위치해 있다는 것을 알수 있다. 이 주택에 사는 시민들의 소득의 중앙값은 3~ 8사이에 집중적으로 분포 되어있다. 이 또한 level\_2의 소득구간과 굉장히 유사함이 보인다. 비싼 주택에 거주하는 만큼 소득수준도 높은 곳에 주로 분포하는 것을 관찰할 수 있다.

전 처리 후 4718개의 데이터가 나왔고, 세 개의 level중 가장 적은 수를 기록했다.

또 하나 알 수 있는 점은 level\_1과 level\_2, 그리고 level\_3의 소득의 중앙 값를 기준으로 outlier가 주로 이 경도구간(-118과 -122)에 위치하는 것을 알 수 있는데, 소득이 높으면서 값싼 주택에 사는 경우나 중간 가격정도의 주택에 거주하는 시민들은 북서부나 남동부에 거주함을 알 수있다.

전체적으로 정리하면, 소득이 낮고 값 싼 주택에 거주하는 시민의 경우 중부에 집중 되어있고, 소득이 높아질수록, 주택의 가격이 비싸질 수록, 서부나 동부로 조금씩 이동해가는 분포였다.

문제3.

스크린샷이(가) 표시된 사진

자동 생성된 설명<약 100번 정도 실행해 본 결과 위의 결과가 제일 비슷했던 것 같아 캡쳐 했습니다.>

jittering이 필요한 경우:

정수형 데이터나 데이터 겹치는 부분이 많을 때 사용하면 좋은 기법으로써,

0에 가까운 임의의 노이즈를 랜덤으로 생성하여 결과가 더욱 자연스러워지고, 겹치는 부분이

빗겨 나가기 때문에 원 분포에 대한 정확한 정보를 얻을 수 있는 장점이 있다.

즉 데이터가 많이 겹쳐 있거나 정수형 데이터처럼 연속적이지 않은 데이터에 약간의 노이즈를 가해서 더욱 자연에 가까운 데이터를 만들 때 사용합니다.