**교육일지**

**교육 제목 : 머신러닝**

**교육 장소 : YGL C6 강의실**

**교육 일시 : 2021/10/14**

**초매개변수조절(Hyperparametertuning)**

**-초매개변수는 학습 과정을 제어하는 데 사용되는 매개 변수를 의미**

**-초매개변수는 모델 학습과정이 아닌 모델 개발자에 의해서 지정됨**

**K-nearestneighborsclassification**

**지도학습으로서 분류(Classification) 나 회귀(Regression)에 사용되는 비모수적 방법**

**파라메터 학습을 위한 훈련과정이 없으나 훈련집합은 필요**

**각 데이터 간에 거리를 계산하기 위한 거리척도가 필요**

**초매개변수 k를 설정해야 함**

**거리에 대한 가중치**

**반응변수전처리(Targetengineering)**

**·주로 parametric model에서 예측 및 모델 적용을 위해서 사용**

**·e.g. Gaussian distribution, Ordinary linear regression**

**1. Log transformation**

**2. Box-cox transformation**

**Logtransformation**

**· 오른쪽으로 치우친 분포 (Right skewed)가 정규 분포로 변환**

**Feature 표준화(Standardization)**

**·각 각 feature의 측정 단위에 대한 보정**

**- 예) 아파트 값을 추정하기 위한 feature들 중 평수(30평)와 주변지역의 땅값**

**(3,000,000/평)**

**·Centering and scaling을 통해서 평균이 0, 표준편차가 1이 되도록 변환 해 줌**

**결측치종류**

**· 무작위 결측치 (Random missing value)**

**-완전무작위 결측치 (MCAR: Missing Completely At Random)**

**- 예) 단순한 결측치**

**무작위 결측치 (MAR: Missing At Random)**

**- 예) 여성(X1)의 경우 체중(X2)에 대한 답이 없음**

**비무작위 결측치 (NMAR: Not Missing At Random)**

**- 체중(X2) 무거운 사람은 체중(X2)에 대한 답이 없음**

**결측치대체(Imputation)**

**결측치를 “최상의 추측”값으로 대체**

**Estimated statistic (e.g., Mean, Median, Mode, Regression)**

**K-nearest neighbor**

**Tree-based**

**중요하지않은Feature제거(filtering)**

**의미없는 변수들 (non-informative predictors) 을 포함했을 때 RMSE 의 변화**

**중요하지않은Featurefiltering**

**의미없는 변수들 (non-informative predictors) 을 포함했을 때 학습시간변화**

**제로분산feature(Zerovariancefeatures)**

**제로 분산 features를 판단하는 일반적 기준**

**- 전체 샘플중에 서로 다른 관측값의 비율이 낮은 경우 (약 ≤ 10%)**

**- 가장 빈도가 높은 관측값과 두 번째로 높은 관측값 과의 비가 높은 경우 (약 ≥ 20배)**

**범주형데이터(Categoricalfeature)engineering**

**재범주화 (Lumping)**

**One-hot & dummy encoding**

**Label encoding**

**Replacing with the mean or proportion**

**Lumping**

**· 매우 작은 빈도를 갖는 범주들을 모아서 하나의 범주로 재범주화**

**One-hot&dummyencoding**

**각 범주를 1 또는 0 (True or False) 로 표시**

**Labelencoding**

**· 각 범주 자료를 연속형 변수로 바꾸어 표현 (순서형 자료의 경우)**

**- e.g.) Very high (=5) , high (=4), moderate (=3), low (=2), very low (=1)**