




Statistical Machine Learning

2주차
담당: 14기 박상준



Contents

1. Data Preprocessing

2. Learning Process

3. Model Evaluation

4. Prediction

1. Data Preprocessing

보다 높은 정확성을 갖는 분석을 위해 원자료에 대해 전환 및 가공을 거치는 단계
(AutoML의 등장으로 그 중요도 및 비중이 높아지고 있음)

정규화와 표준화

특성변수의 단위 등에서 나타나는
차이를 조정해주는 역할

One-Hot Encoding

범주형 변수를 수치형 변수로 변환

Bag of Words

텍스트를 수치형 변수로 변환해주는
방법

차원축소

Feature이 많을 때 발생하는 overfitting을
방지하기 위하여 진행

이상치/결측치 처리

머신러닝 모형은 직접 결측치를 처리할
수 없음

불균형 자료처리

불균형 자료 문제 해소를 위한 과대표집
방법

1. Data Preprocessing

PCA

Principal Component Analysis

Linear Dimensionality Reduction Technique

Reduce dimensionality of highly correlated data by transforming original set of vectors to a new set

t-sne

T-distributed stochastic neighborhood Embedding

Non-Linear Dimensionality Reduction Technique

Minimize the Kullback-Leibler Divergence (KL Divergence) between the two distributions

1. Data Preprocessing

SMOTE

Oversampling

K-nearest neighbors

$$\mathbf{x}_{syn} = \mathbf{x}_i + \lambda (\mathbf{x}_k - \mathbf{x}_i), \mathbf{x}_k \in S_i$$

ADASYN

Oversampling

SMOTE + α

표본 수에 대한 weight로 추출

2. Learning Process

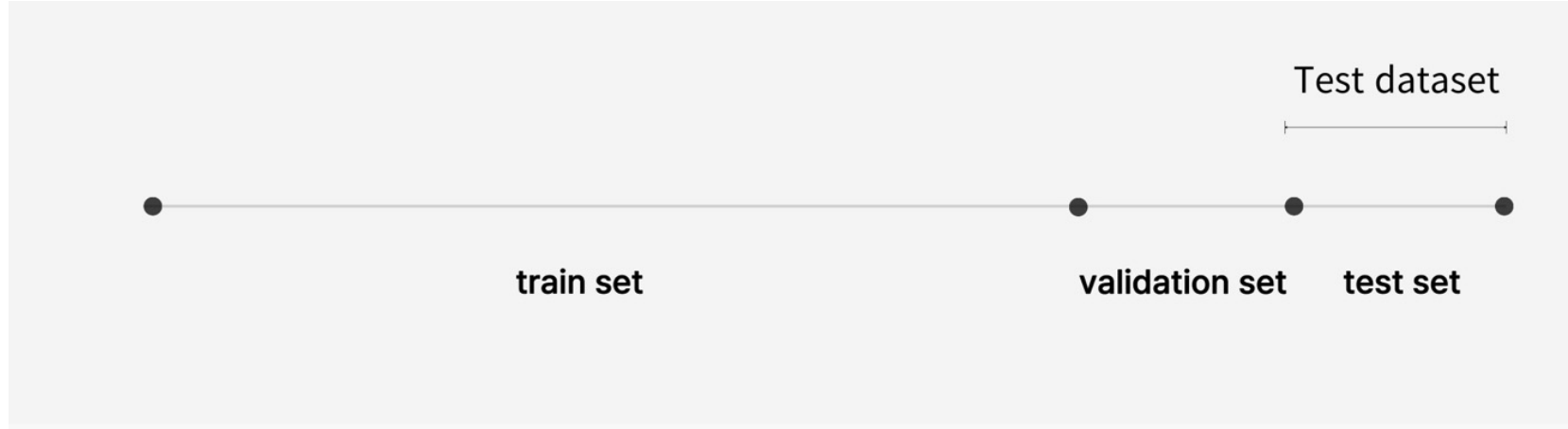
Use train set to fit a model + adjust hyperparameters of model



3. Model Evaluation

Apply fitted model to a test set to check model performance

To overcome overfitting: Increase data size, regularization, ensemble method etc.



4. Prediction

Apply a generalized model into the field

