

단단한 머신러닝

1. 서론

I. 서론

- 특성, 차원
- 레이블 공간
- 특성 공학
- 데이터의 분리
- 지도 학습
- 비지도 학습

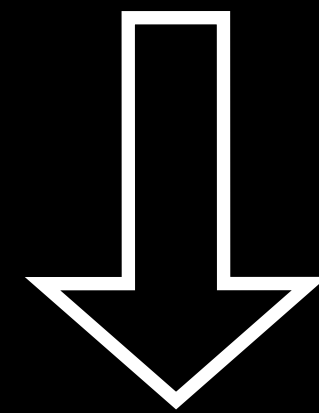
특성, 차원 및 레이블 공간

특성 (Feature)

색깔
(청록색, 진녹색)

꼭지 모양
(곧음, 말림)

소리
(둔탁함, 맑음, 혼탁함)



레이블

잘 익은 수박?
(예 / 아니오)

특성, 차원 및 레이블 공간

특성 (Feature)

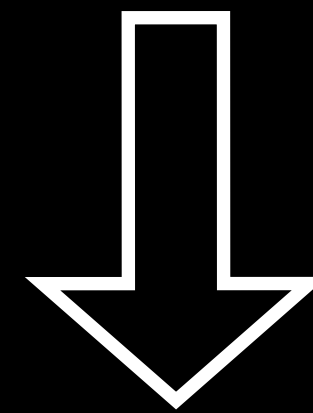
성별

나이

좌석 클래스

가족 수

운임



레이블

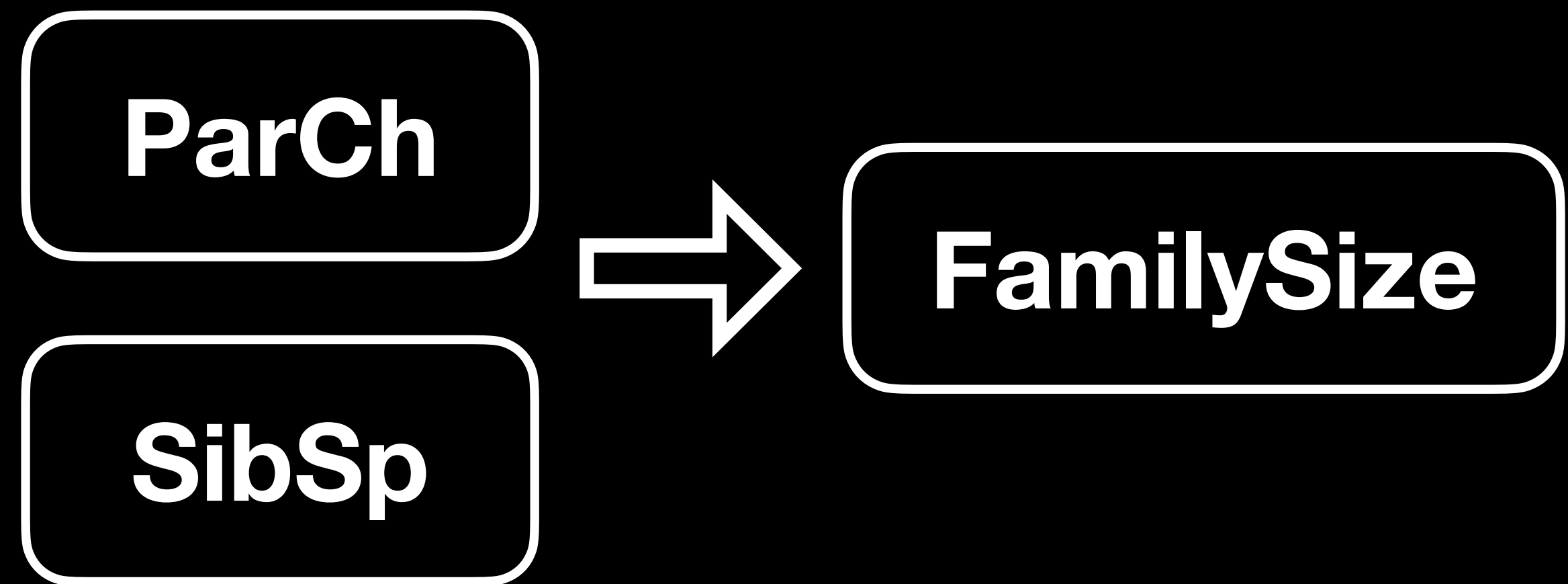
생존 여부
(예 / 아니오)

특성 공학 (Feature Engineering)

- 특성의 유용한 표현을 추출하기 위해 사람이 초기 입력 데이터를 수동으로 변환하는 것
- Categorize
- Scaling
- One - Hot encoding

특성 공학 (Feature Engineering)

- 특성의 유용한 표현을 추출하기 위해 사람이 초기 입력 데이터를 수동으로 변환하는 것
- Categorize
- Scaling
- One - Hot encoding



데이터의 분리

훈련 데이터 / 검증 데이터 / 테스트 데이터

실제 데이터
(정답 X)

(정답이 있는) 전체 데이터
Labeled Data

데이터의 분리

훈련 데이터 / 검증 데이터 / 테스트 데이터

실제 데이터

훈련 데이터
Train Data

테스트 데이터
Test Data

데이터의 분리

훈련 데이터 / 검증 데이터 / 테스트 데이터

실제 데이터

모델링 후 테스트

훈련 데이터
Train Data

테스트 데이터
Test Data

데이터의 분리

훈련 데이터 / 검증 데이터 / 테스트 데이터

실제 데이터
“수능”

모델링 후 테스트

훈련 데이터
Train Data

“연습문제”

테스트 데이터
Test Data

“모의고사”

데이터의 분리

훈련 데이터 / 검증 데이터 / 테스트 데이터

실제 데이터
“수능”

모델링 후 테스트

— 모델 검증 →

훈련 데이터
Train Data

검증 데이터
Validation Data

테스트 데이터
Test Data

“연습문제”

“꼭지시험”

“모의고사”

GitLens

1. kNN

```
[47] ▶ ▶≡ M↓  
test_model(  
    KNeighborsClassifier(n_neighbors=13)  
)
```

Average Score: 80.92

2. Decision Tree

```
[48] ▶ ▶≡ M↓  
test_model(  
    DecisionTreeClassifier()  
)
```

Average Score: 79.69

3. Random Forest

```
[49] ▶ ▶≡ M↓  
test_model(  
    RandomForestClassifier(n_estimators=13)  
)
```

Average Score: 80.36

4. Naïve Bayes

```
[50] ▶ ▶≡ M↓  
test_model(  
    GaussianNB()  
)
```

Average Score: 71.83

5. SVM (selected)

```
[51] ▶ ▶≡ M↓  
test_model(  
    SVC()  
)
```

Average Score: 82.83

Testing

```
[52] ▶ ▶≡ M↓  
clf = SVC()  
clf.fit(train_data, train_target)
```

지도 학습

Supervised Learning

- k-최근접 이웃 기법
- 서포트 벡터 머신
- 결정 트리 / 랜덤 포레스트
- 선형 회귀
- 로지스틱 회귀
- 신경망

비지도 학습

Unsupervised Learning

- 클러스터링
- 차원 축소
- 주성분 분석