3.1 Introduction

Model-Based Methods

- data에 대한 model이 먼저 생성됨
- Training 과 Prediction 단계가 확실히 구분되어 있음
- 전통적 machine learning model 들이 collaborative filtering에 적용될 수 있음
- Classification, regression 문제들이 matrix completion 또는 collaborative filtering 의 특이 케이스 이기 때문

3.1 Introduction

Data Classification vs Collaborative Filtering

- Data Classification 에서는 feature(독립) 변수와 class(종속) 변수간 확실한 구분 / Matrix Completion 에서는 확실한 구분이 존재 하지 않음 각 column은 종속 이면서 독립 변수
- Data Classification 에서는 Training Data 와 Test Data 간의 확실한 구분 / Matrix Completion에서는 Matrix의 row를 따라서 확실한 구분이 존재 하지 않음 specified된 항목을 Training Data, Unspecified Data를 Test Data로 보는게 최선
- Data Classification 에서는 Column은 Feature들을 표현하고 Row들은 케이스를 표현 / Collaborative Filtering 에서는 이 접근 법이 Rating Matrix와 Transpose 모두에 적용 가능

3.1 Introduction

Model-based vs Neighborhood-based

Space Efficiency

- 대체적으로 학습된 모델은 기존 Rating Matrix 보다는 크기가 작음 (공간 요구사항이 작음)
- Neighborhood-Based는 User 또는 Item의 수의 제곱에 비례

Training Speed and Prediction Speed

- Model-based는 대체적으로 Training 속도가 빠르고 예측이 더 효율적임
- Neighborhood-based는 전처리 과정이 User나 Item수의 제곱에 비례

Avoiding Overfitting

- Model-based는 overfitting을 피하는데 도움이 됨
- Regularization 이 Model을 더 Robust 하게 해줌

3.2 Decision and Regression Trees

Decision and Regression Trees

- Decision Tree는 종속 변수가 범주형인 경우
- Regression Tree는 종속 변수가 수치형인 경우
- Decision Tree는 독립 변수에 split criteria 라고 부르는 계층적 의사 결정의 집합으로 data space에 대한 계층적 분리를 의미함
- split 의 품질은 split에 의해서 생긴 자식 노드들의 Gini index의 가중 평균을 통해서 산정됨

3.2 Decision and Regression Trees

Gini Index

- 0부터 1사이의 값을 가지며 작은 값을 가질수록 구별할 수 있다는 표현
- p1부터 pr이 노드 S에서 r개의 다른 class의 data record 일때 노드의 Gini Index는 아래와 같음

$$G(S) = 1 - \sum_{i=1}^{r} p_i^2$$

- split의 전체적인 Gini index는 자식 노드들의 Gini index의 가중 평균과 일치
- s_1과 s_2 가 Binary Decision Tree에서 노드 S의 두 자식 노드인 경우, n_1 과 n_2 가 data record일때

$$\text{Gini}\,(S \Rightarrow [S_1, S_2]) = \frac{n_1 \cdot G(S_1) + n_2 \cdot G(S_2)}{n_1 + n_2}$$

3.2 Decision and Regression Trees

Decision Trees

- Gini Index는 split 진행시 적절한 attribute 선택을 위해 사용됨 (가장 작은 Gini Index의 attribute가 선택)
- 각 노드가 특정 클래스의 data records들만 보유할때 까지 위 선택과정을 반복
- 특정 클래스 보유가 아닌 특정 클래스에 대해 정해진 수치 보다 적은 data record 보유시 중단 가능
- 위 케이스에서는 node의 주보유 클래스가 label이 되고 Leaf Node라고 함
- Decision Tree는 계층적 분할이기 때문에 테스트 케이스는 특정 경로를 따라 Leaf로 이동함
- 수치 데이터에 적용을 위해서는 attribute value가 구간으로 변환되어 split을 진행

3.2 Decision and Regression Trees

Decision Trees with Numerical Variables

- 수치 데이터에 적용을 위해서는 attribute value가 구간으로 변환되어 split을 진행
- 구간으로 split이 진행되기 때문에 Multi-way split이 이루어질 수 있음
- Numeric 종속 변수에서는 Gini Index 대신 분산을 기준으로함
- 낮은 분산일수록 나은 분별 능력을 의미
- Prediction 진행시에는 Leaf Node의 평균 값이나 Linear Regression Model을 적용함

3.2 Decision and Regression Trees

Pruning in Decision Trees

- Overfitting을 피하기 위해서 적용됨
- Tree 생성 단계에서 Training Data의 일부를 사용하지 않음
- Pruning의 효과를 확인하기 위해서 사용하지 않은 데이터를 가지고 테스트
- node의 제거가 Training시 사용하지 않은 데이터를 통한 데이터에 대해 정확도를 상승 시키는 경우 node 는 제거됨

3.2.1 Extending Decision Trees to Collaborative Filtering

Main Challenge

- Column 기준으로 구분이 확실히 되어 있지 않다는점
- Rating Matrix의 주요 entry 들이 비어 있다는점
- Collaborative Filtering에서는 종속과 독립 변수가 확실히 구분되어 있지 않음
- 위 이슈는 각 item에 대한 rating을 예측하기 위해 각각 decision tree를 생성

3.2.1 Extending Decision Trees to Collaborative Filtering

Missing Independent Features

- user의 item에 대한 rating을 threshold 기준으로 분리시 비어 있는 데이터를 양쪽 브랜치로 모두 할당시 tree가 엄격한 분할이 아님
- 테스트 케이스가 tree에서 여러 path를 따라 매핑
- 2.5장에서 제시된것처럼 낮은 차원의 표현으로 대채하는 접근
- m x (n-1) 차원의 matrix에서 d<<n-1인 mxd 차원의 fully specified된 matrix로 변환
- 위 과정을 통해서 각 user의 d차원 rating vector가 생성
- 이 축소된 표현으로 해당 item에 대한 decision tree를 생성
- 총 n개의 item에 대해 n개의 decision tree가 생성되며 j번째 tree로 j번째 item에대한 rating 예측이 가능

3.3 Rule-Based Collaborative Filtering

Association Rule Mining

- Transaction Database T={T_1 … T_m }이 n개의 item I에 대해서 정의
- I는 item의 전체 집합이며 각 Transaction Ti는 I의 부분집합
- Association Rule Mining의 key는 Transaction Database에서 상관관계 높은 item의 집합들을 찾는것
- Support와 Confidence라는 개념을 통해서 item 집합들간의 관계가 측정됨

Support

- 정의: T안의 부분집합 itemset X의 비율
- itemset의 support가 미리 정해진 s에 적어도 일치하는 경우 itemset은 frequent하다고 함
- Threshold s는 minimum support, 위 itemset들은 frequent item sets, frequent patterns

3.3 Rule-Based Collaborative Filtering

Confidence

- 정의: X $\Rightarrow$ Y의 Confidence는 Transaction T가 X를 포함할때, Y를 포함할 조건부 확률
- X U Y 의 support를 X의 support로 나눈것
- Rule에대한 Confidence는 항상 (0,1) 범위에 존재
- Confidence가 높을 수록 Rule의 강하다는 표시

3.2.1 Extending Decision Trees to Collaborative Filtering

Association Rule

- 정의: Rule X ⇒ Y 는 최소 support s 와 최소 confidence c에서 아래 조건을 만족시 association rule이라함

1. X U Y 의 support가 적어도 s
2. X ⇒ Y의 confidence가 적어도 c

- Association Rule을 찾는건 두 단계 프로세스
- 첫 단계 에서는 최소 support s 를 만족하는 모든 itemset들을 정의
- 위 각 items Z에서 (X,Z-X)의 모든 분할을 구성하여 X ⇒ Z-X 룰을 생성
- 위 룰 들 중 최소 confidence를 만족한 Rule 들만 유지

3.3.1 Leveraging Association Rules for Collaborative Filtering

Association Rule

- Association Rule 들은 단항 Rating Matrix 에서 추천시 유용
- Rule-based Collaborative Filtering의 첫 스텝은 미리 정의 되어 있는 최소 support와 최소 confidence에 해당하는 모든 association rule들을 찾는것
- 최소 support 와 최소 confidence는 예측정확도를 최대화 하기위해 tuning 되는 paramete로 볼 수 있음
- 결과가 정확히 한 item 만 있는 rule들이 남음
- 이 rule들의 집합이 모델이며 특정 사용자에게 추천시 사용가능
- Antecedent가 user에 대한 동일한 pseudo-item 부분집합을 갖는 rule들이 정의되고 confidence 감소 순서대로 정렬
- Top -k 개의 pseudo-item을 선택함으로서 정렬된 Rule들을 통해 item에대한 rating 예측에 사용 가능
- Pseudo-item간 발생 가능한 모순은 평균을 측정하여 해소 할 수 있음

3.3.2 Item-Wise Models versus User-Wise Models

Item-Wise Models versus User-Wise Models

- 앞서 Item-Wise에 대해서 진행됬기 때문에 User-Wise를 위해서 Transpose에 적용
- Transpose에 적용하기 위해서 pseudo-users로 생성, Item에 대한 pseudo-users가 Transaction
- Association Rules들이 최소 support와 최소 confidence를 기준으로 mined
- Association Rule 접근은 collaborative Filtering 뿐 아니라 특정 item에 소비자를 매칭 시키는 content-based recommender system에서도 유용
- 위 Rule들을 profile association rules라함

Classification Problem vs Recommender Systems

- Classification은 생성된 rule의 Consequent가 항상 class variable을 포함
- Recommender system에서 생성된 rule의 Consequent는 어떤 item도 포함할 수 있음
- 둘의 주 차이점은 Collaborative Filtering 에서는 feature variable과 class variable에 대한 확실한 구분이 없다는것