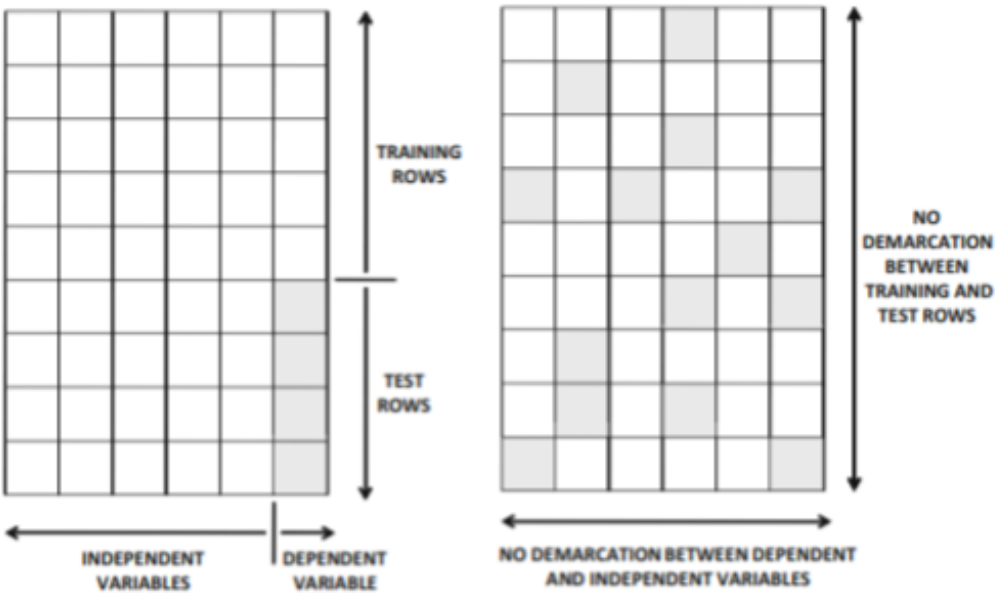
제3장 모델 기반 협업 필터링

3.1 개요

2장에서 소개한 이웃 기반 방법(neighborhood-based method)은 ML에서 흔히 사용하는 k-nearest neighbor classifier의 일반화로 볼 수 있습니다. **모델 기반의 방법**에서는 지도학습 또는 비지도학습 머신러닝 방법과 마찬가지로 데이터의 모델 요약이 앞부분에 작성됩니다. 따라서, 학습이 예측 단계와 확연히 구분됩니다.

하지만, Classifier와 달리 Matrix Completion Problem은 다음과 같은 결정적인 차이점이 존재합니다.



(a) Classification (b) Collaborative filtering

- feature 변수와 class 변수의 명확한 구분이 없습니다.
- 행렬의 행 사이에서 명확한 구분이 없습니다.

- 기존 행렬과 전치 행렬에 같은 방식의 접근을 적용할 수 있습니다.

분류 분야에서 개발된 대부분의 **앙상블 방법**에 대한 이론이 추천 시스템에 적용되고 있습니다. 그러나 데이터 분류 모델을 바로 행렬 완성 문제로 일반화하는 것은 쉽지 않고, 특히 대부분 항목이 누락된 경우 더 어렵습니다.

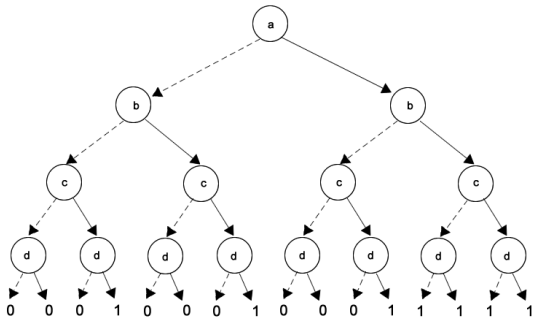
모델 기반 추천 시스템은 이웃 기반 방법보다 **용량 효율성, 학습 속도와 예측 속도, 과적합 방지** 면에서 장점이 있습니다.

아래부터 추천 시스템에서으 의사 결정, 회귀 트리 활용법과 규칙 기반 협업 필터링과 나이브-베이즈 모델과 **잠재 요인 모델** 등에 대해 다루어보겠습니다.

3.2 의사 결정 및 회귀 트리

의사 결정 트리(Decision Tree)는 독립변수에서 분리 기준으로 알려진 계층적 결정 집합을 사용해 데이터 공간을 계층적으로 분리합니다.

의사 결정 트리의 노드가 두 자식을 갖는 경우, 결과적으로 나온 의사 결정 트리를 이진 의사 결정 트리라고 합니다.



분리의 품질은 가중 평균 **지니 계수**를 통해 평가할 수 있으며, 지니 계수는 0에서 1 사이의 값을 갖고 0에 가까울수록 큰 분별 능력을 의미합니다.

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

탑-다운 방식의 계층 분리는 가장 작은 지니 계수를 갖는 속성이 선택되며 진행되며 각 노드가 오직 하나의 클래스에만 속할 때까지 수행합니다. 그렇게 진행되어 마지막 노드가 leaf node입

니다. 데이터의 독립변수들은 Decision Tree의 root부터 leaf까지 경로를 만드는 데 사용합니다.

많은 경우 트리는 과적합을 줄이기 위해 가지치기(pruning)를 합니다.

3.2.1 의사 결정 트리를 협업 필터링으로 확장

협업 필터링에 분류 모델인 의사 결정 트리를 적용할 때의 문제점은,

첫째, 대부분 항목이 누락된 평점 행렬은 매우 희소한 행렬이라는 점입니다.

둘째, 예측 항목과 관측된 항목이 열 방식으로 피처 변수와 클래스 변수로 구분이 안 된다는 점과

두번째 문제는 각 아이템의 평점을예측하는 별도의 의사 결정 트리를 구축해 해결할 수 있습니다. 각각의 의사 결정 트리에서 하나의 특성만이 종속 변수이고 나머지 특성은 독립 변수입니다.

반면, 첫번째 문제는 다루기 쉽지 않습니다. 차원 축소 방법을 이용하는 방법이 합리적인 접근 방법입니다. User마다 완전히 명시된 평점 벡터를 얻을 수 있게 되기 때문에 차원 축소와 분류 모델을 결합한 접근은 매우 유용합니다.

3.3 규칙 기반 협업 필터링

연관 규칙(Association Rules)은 이진 데이터에서 자연스럽게 정의됩니다. 연관 규칙 mining의 핵심은 transaction database에서 상관관계가 높은 items의 집합을 정하는 일입니다. 이 작업은 **지지도**와 **신뢰도** 개념에 의해 이뤄지며 이 지표는 아이템 집합들 사이의 관계를 수치화합니다.

만약 itemset의 **지지도(support)**가 최소한 미리 정해둔 임계값과 같다면 itemset는 fregent하다고 합니다. 다시 말해, 임계값을 minimum support라 하며 frequent itemsets의 기준이 됩니다. 기업 입장에서 높은 지지도를 갖는 패턴을 찾는 것은 매우 유용하며 이 연관 규칙은 $X \Rightarrow Y$의 형태로 표시됩니다. 이러한 규칙의 강도는 신뢰도를 통해 측정됩니다.

신뢰도(confidence)는 $X \cup Y$의 지지도를 X의 지지도로 나눈 값으로 얻어집니다. 높은 신뢰도 값은 항상 규칙의 강도가 더 강하다는 표현입니다.

연관 규칙을 찾는 과정은 위 과정의 알고리즘을 사용합니다. 우선, 최소 지원 임계값을 만족하는 itemset가 결정되고, 각 set에서 가능한 모든 양방향 파티션을 사용해 잠재저 규칙 $X \Rightarrow Z - X$를 생성합니다. 최소 신뢰도를 만족시키는 규칙이 유지됩니다.

연관 규칙은 단항 평점 행렬의 맥락에서 추천을 수행하는 데 특히 유용합니다. 규칙 기반 협업 필터링의 첫 단계는 사전 정의된 최소 지원 및 최소 신뢰 수준(매개변수)에서 모든 연관 규칙을 발견하는 것입니다.

3.4 나이브 베이즈 협업 필터링

가장 먼저 각 평점이 범주 값으로 취급될 수 있는 **구별되는 평점**이 적다고 가정합니다. 즉, 평점 사이의 순서는 무시된 채 이산값으로 처리합니다.

나이브 베이즈 모델은 일반적으로 분류에 사용되는 생성 모델입니다.

아이템 j에 대해 사용자 u의 관찰되지않은 평점 r_{uj}를 예측해야 하는 경우, 확률론의 베이즈 법칙을 사용해 표현식을 단순화할 수 있습니다.

Bayes rule : $P(A|B) = \frac{P(A) \cdot P(B|A)}{P(B)}$

$P(r_{uj} = v_s | I_u\text{에서관측된평점}) = \frac{P(r_{uj}=v_s) \cdot P(I_u\text{에서관측된평점} I_u | r_{uj}=v_s)}{P(I_u\text{에서관측된평점})}$

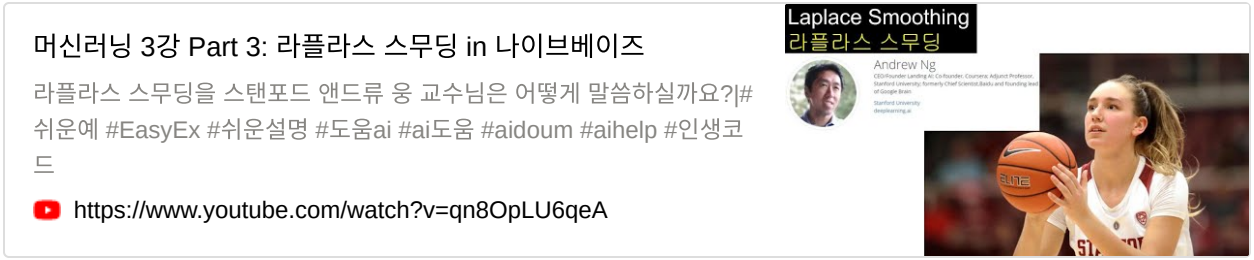
$P(r_{uj} = v_s | I_u\text{에서관측된평점})$은 **나이브 가정**을 사용해 추정합니다.

나이브 가정은 평점 간의 **조건부 독립성**을 기반으로 합니다.

평점 r_{uj}의 사후 확률 추정치는 최대 확률을 차지하는 평점으로 결정하거나 예측된 값 모두의 가중 평균으로 결정할 수 있습니다. 전자의 경우, 가능한 평점 수가 적을 때 합리적인 사용 방법이며 후자는 평점 분포의 세분성이 큰 경우에 바람직합니다.

3.4.1 과적합 조정

행렬이 희소하고 관측된 평점의 수가 적으면 데이터 중심 추정치가 확실하지 않거나 결정되지 않습니다. 다시 말해, 소량의 데이터에서 모델 파라미터를 추정하기 때문에 잘못되고 과적합된 결과를 얻을 위험이 있습니다.



이를 해결하기 위해 라플라시안 스무딩(Laplacian smoothing) 방법이 일시적으로 사용됩니다. 라플라시안 스무딩 파라미터 α를 통해 스무딩 레벨을 조절하고 값이 클수록 스무딩 효과는 더 커져 과적합이 방지되지만 그 결과는 기본 데이터를 덜 반영하는 방향으로 나아갑니다.

3.4.2 이진 평점에 대한 베이즈 방법의 적용 예

교재 상 예시이기에 생략

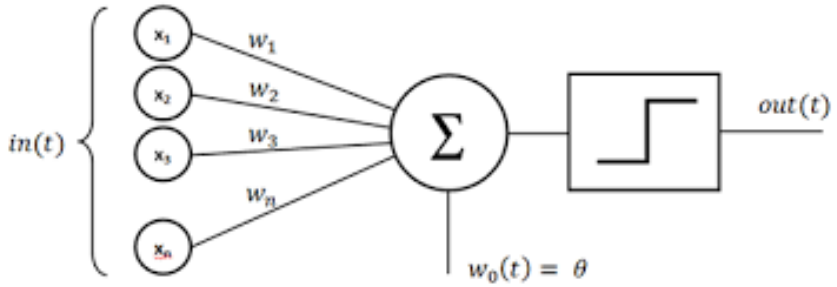
3.5 임의의 분류 모델을 블랙 박스로 사용

다른 많은 분류 방법은 협업 필터링으로 확장될 수 있는데 이러한 방법의 가장 큰 문제점은 기본 데이터의 불완전성입니다. 차원 축소 방법을 사용해 데이터의 저차원 표현을 만들 수 있지만 이 또한 분류 과정에서 해석 능력이 손실될 우려가 있습니다.

따라서, 연속적으로 세부적으로 훈련 열의 누락된 값을 반복적으로 채우며 문제를 해결합니다.

3.5.1 예: 신경망을 블랙 박스로 사용

신경망은 시냅스 연결을 통해 서로 연결된 뉴런을 사용해 인간의 뇌를 시뮬레이션합니다. 인공 신경망에서 기본 연산 단위는 뉴런이라고도 하며 시냅스 연결의 강도는 가중치에 해당합니다.



$$out(t) = sign\{\overline{W} \cdot \overline{X_i} + b\}$$

$\overline{X_i}$ 는 i번째 훈련 인스턴스의 d입력을 정의하는 d-차원 행 벡터이고 $\overline{W}$는 계수 벡터입니다. 파라미터 b는 bias를 나타내고 $out(t)$는 예측 값으로 예측된 출력의 오차는 선형 회귀와 유사한 방식으로 가중치를 W로 업데이트하는 데 사용됩니다.

일반적으로 신경망은 여러 계층을 가질 수 있으며중간 노드는 비선형함수를 계산할 수 있습니다. 해당 학습 알고리즘을 **역전파 알고리즘(back-propagation algorithm)**이라고 합니다. 신경망의 주요 장점은 다층 아키텍처가 다른 분류 방법으로는 쉽게 계산할 수 없는 복잡한 비선형 함수를 계산할 수 있다는 것입니다. 평점 행렬과 같은 noise data의 경우 **정규화**를 사용해 noise의 영향을 줄입니다.

3.6 잠재 요인 모델

행렬 완성을 위한 직접적인 방법으로 잠재 요인 모델을 사용하는데, 기본 아이디어는 데이터 행렬의 행과 열의 중요한 부분이 서로 밀접하게 관련돼 있다는 사실을 활용하는 것입니다. 데이터에는 중복성이 내장돼 있고 그 결과 데이터 행렬은 종종 저차원 행렬에 의해 상당히 잘 추정됩니다.

이로서 희소 행렬의 작은 부분으로도 완전히 지정된 하위 근사치를 결정할 수 있고 데이터 행렬의 항목을 재구성하기 위해 **기대치값-최대화(EM: Expectation-Maximization) 기법**을 차원 축소와 결합합니다.

차원 축소 방법은 행과 열의 상관관계를 활용해 완전히 지정되고 축소된 표현을 만듭니다.

*세부 사항 추후 기록 예정

3.7 분해와 이웃 모델 통합

*세부 사항 추후 기록 예정

3.8 요약

협업 필터링을 위한 여러 모델을 설명했습니다. 협업 필터링 문제는 분류 문제의 일반화로 볼 수 있고 따라서 분류에 적용되는 많은 모델은 일부 일반화를 사용해 협업 필터링에도 적용됩니다.

예외적인 것 중 주목할 만한 경우는 협업 필터링 문제에 고도로 맞춤화된 **잠재 요인 모델**입니다.

잠재 요인 모델은 평점을 예측하기 위해 다양한 유형의 분해를 사용합니다. 이러한 다양한 분해 유형은 목적함수의 특성과 기본 행렬에 대한 제약 조건이 다릅니다. 또한 정확성, 과적합 및 해석 가능성 측면에서 서로 다른 절충안이 있을 수 있습니다.

잠재 요인 모델은 협업 필터링의 최첨단입니다. 목적함수 및 최적화 제약 조건의 선택을 기반으로 다양한 잠재 요인 모델이 제안되었으며 잠재 요일 모델을 이웃 방법과 결합해 통합 모델을 생

성할 수도 있으며, 이는 잠재 요인 모델과 이웃 방법 모두의 이점을 누릴 수 있습니다.