비즈니스애널리틱스 기말고사 문제은행: 총 50 개 문항

2023 년 6 월

■ Week02. 비즈니스 애널리틱스 개요 (I)

1. 의사결정지원시스템(DSS, Decision Support System)의 '모델(Model)'을 개발하는 2 가지 접근법에 대해 설명하고, 각 접근법을 어떤 상황에서 사용해야 하는지 설명하시오.

의사결정지원시스템의 모델을 개발하는 방법에는 귀납적 접근법, 연역적 접근법이 있다. 귀납적 접근법은 개별적인 사실에서부터 일반적인 사실들을 이끌어 내는 방법이고, 연역적 접근법은 일반적 사실에서 특수한 사실들을 이끌어 내는 방법이다. 데이터가 많은 경우, 데이터들을 토대로 일반적 원리를 이끌어내는 귀납적 접근법을 주로 사용하나, 데이터가 부족(부실)한 경우, 귀납적 접근법을 적용하는 것은 적절하지 못하므로 이럴 경우, 연역적 접근법을 사용하는 것이 대안이 될 수 있다.

2. 데이터(data) 정보(information), 지식(knowledge)의 차이를 설명하시오.

데이터는 그 자체로는 의미를 갖고 있지 않은 날(raw) 것 그대로의 원시 자료를 의미한다.

정보는 배경정보가 함께 고려되어 의미가 갖게 된(라벨링이 된) 데이터를 의미한다.

지식은 문제를 해결하거나 의사결정을 내리는데 활용될 수 있는 정보들 사이에 내재되어 있는 패턴을 의미한다.

[DIKW 피라미드를 통해 차이에 대한 예시]

Data: A 는 100 원, B 는 200 원에 연필을 판매한다.

Information: A 연필이 더 저렴하다.

Knowledge: 더 저렴한 A 로부터 연필을 사야겠다.

3. 지식공학(Knowledge Engineering) 기법의 5 가지 유형을 개괄으로 설명하시오.

예측(분류)(Classification) : 축적된 과거의 정보를 활용하여, 미래 시점의 사건 발생이나 결과를 예측하는 통계 혹은 기계학습 모델로, 연속적인 값(Continuous Value)을 예측하는데 활용되는 회귀(Regression) 모델과 범주형 값(Class Label)을 예측하는 데 활용되는 분류(Classification) 모델이 있다.

전처리(Preprocessing): 예측(분류) 모델 등의 성과를 향상시키기 위해, 입력 데이터에 대해 적절

한 사전 처리를 수행하는 기법 (information gain 의 효과를 기대)을 의미한다. 전처리를 위한 지식공학 기법으로는 군집화 기법, 변수 이산화(Feature Discretiazation) 기법, 주성분 분석(PCA, Principal Component Analysis), 퍼지 이론(Fuzzy Theory) 등이 있다.

군집화(Clustering): 사전에 정해진 기준 없이, 서로 동질한 데이터들을 같은 그룹으로 묶어주는 기법이다. 군집화를 위한 지식공학 기법으로는 계층적 군집분석, 비계층적 K-means 군집분석, 비계층적 자기조직화지도(SOM, Self-Organizing Maps) 군집분석이 있다.

가치평가(Valuation): 정성적인 측정대상의 가치나 성과를 정량화하여 평가하는 기법이다. 가치평가를 위한 지식공학 기법으로는 분석적 계층 프로세스(AHP: Analytic Hierarchy Process), 분석적 네트워크 프로세스(ANP: Anaylytic Network Process), 자료포락분석(DEA: Data Envelopment Analysis)가 있다.

최적화(Optimization): 주어진 제약조건 하에서, 특정 목적함수를 최대/최소화하는 변수들의 최적값을 도출하는 기법이다. 최적화를 위한 지식공학 기법으로는 선형계획법(LP: Linear Programming), 유전자 알고리즘(GA: Genetic Algorithms), 시뮬레이티드 어닐링(SA, Simulated Annealing), 개미 집단 최적화 알고리즘(ACO, Ant Colony Optimization) 등이 있다.

4. 예측(prediction)의 2 가지 종류(유형)에 대해 설명하고, 각 유형의 성능은 어떤 지표로 평가하게 되는지 예를 들어 설명하시오.

예측(prediction)의 2 가지 종류는 회귀(Regression) 모델과 분류(Classification) 모델이 있다.

회귀(Regression) 모델은 연속적인 값(Continuous Value)을 예측하는데 활용되며, 성능 평가 시 오차를 최소화하는 것을 목적으로 한다. 대표적인 성능 평가 지표로는 오차값을 기준으로 평가하는 RMSE, MAE, MAPE 등이 있으며, 예측값과 실제값 차이에 대해 평균을 측정하는 방식에 따라 그 기준이 다르다.

분류(Classificaion) 모델은 범주형 값(Class Label)을 예측하는 데 활용되며, 성능 평가 시 오분류를 최소화하는 것을 목적으로 한다. 대표적인 성능 평가지표로는 정확도를 기준으로 평가하는 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F1 스코어(F1 Score) 등이 있으며, 예측값과 실제값의 True 와 False 의 비율에 대해 측정하는 방식에 따라 그 기준이 다르다.

5. 아래 상자에 제시된 자동화된 주식매매의사결정지원 모형을 개발 중인 데이터 분석가의 생각을 읽고, 이 분석가가 잘못 생각하고 있는 점은 무엇인지 비판하시오.

각 기업의 주가가 오늘과 비교했을 때 내일 오를지(Y=1) 내릴지(Y=0)를 기준으로 종속변수를 설정하여 이를 예측할 수 있는 이분류(binary classification)모형을 구축할 경우, 익일 주가의 등락여부만 예측할 수 있다. 하지만, 만약 해당 기업의 주가가 내일 얼마의 값을 갖게

될 지를 예측(regression)해도, 오늘의 주가와 비교해 보면 상대적으로 오른 것인지 내린 것인지 판단할 수 있다. 오히려 후자의 방식으로 접근하게 되면, 오늘 대비 크게 오르거나 내릴지, 혹은 적게 오르거나 내릴지 까지도 예측이 가능하므로, 더 많은 예측 정보를 얻을 수 있을 것이다. 따라서 자동화된 주식매매 의사결정지원 모형을 개발하려면 익일 주가의 등락을 예측하기 보다는 익일 주가 자체를 예측하도록 하는 것이 더 효과적이다.

데이터 분석가가 잘못 생각하고 있는 점은 회귀모델과 분류모델의 평가가 동일 할 것이라는 생각이다. 회귀 관점에서는 실제값과 오차값의 차이를 기준으로 평가하고, 분류 관점에서는 실제 결과와 예측 결과의 정확도를 기준으로 평가한다. 회귀모델의 예측치에 따라 분류 관점에서는 분류 정확도가 높은 것이 중요한 반면, 회귀 관점에서는 분류와 별개로 실제값과 예측값의 차이가 작은 것이 중요하다. 즉, 회귀모델의 예측치에 따라 분류 관점에서의 우수한 모델과 회귀 관점에서의 우수한 모델이 다를 수 있다는 것인데 데이터 분석가는 이에 대해 회귀모델과 분류모델의 평가가 동일 할 것이라는 잘못된 생각을 가지고 있다.

6. **분류(classification)와 군집화(clustering)은 모두 전체 샘플들을 여러 개의 소집단으로 나누는 것을 의미한다. 그렇다면, 분류와 군집화의 차이점은 무엇인가?**

분류(classification)와 군집화(clustering)의 가장 큰 차이는 사전에 정해진 기준의 유무에 따른 차이이다. 분류(classification)의 경우, 사전에 정해진 기준에 따라 서로 동질한 데이터들을 같은 그룹으로 묶어준다면, 군집화(clustering)의 경우, 사전에 정해진 기준 없이, 서로 동질한 데이터들을 같은 그룹으로 묶어주는 기법을 의미한다.

7. **전처리 기법으로 변수 이산화(feature discretization)과 퍼지 이론(fuzzy theory)을 적용할 때, 각각 어떻게 정보 이득(information gain) 효과를 누릴 수 있는지 예를 들어 설명하시오.**

변수 이산화를 통해 이산화 했을 때, 오히려 정보의 양이 더 늘어나게 만들 수도 있다. 예시로, 설문을 통해 나이를 조사하였을 때, 연속형 변수인 나이를 이산형 변수로 변수 이산화를 하여, 오히려 정보의 양을 더 늘어나게 만들 수 있다.

퍼지 이론은 불확실한 개념을 수학적으로 모델링하는 기법으로서, 이것을 역으로 활용하면 그 과정에서 정보 이득 효과를 얻을 수 있다. 예시로 에어컨 제어 모듈에 퍼지 이론을 적용할 경우, 1 차원인 기온 변수가 Cold, Warm, Hot, Very Hot 4 차원 변수로 늘어나게 할 수 있다.

8. 분류 모델 구축을 위한 데이터 전처리 과정을 단계별로 간략히 설명하시오.

첫 번재로 분석 단위(Unit of Analysis)에 맞게 데이터 정리 작업을 수행한다. 첫 번째 단계에서는 분석 목적에 따라 분석 단위를 설정하고, 설정된 분석 단위에 따라 1 개의 레코드가 형성 될 수 있도록 원시 데이터를 정리한다. 예시로, 분석 단위에 따라 중복 항목 제거를 하는 등의 작업을 한다.

두 번째로 결측값(Missing Value)를 처리한다. 두 번째 단계에서는 너무 많은 항목이 비어있는 변수나 너무 많은 항목이 비어있는 레코드를 삭제하고, 여전히 남아있는 단계에서는 일반적으로 평균, 중앙값, 최빈치 등 대푯값으로 대체한다.

세 번째로 다중값 속성 및 명목 변수에 대한 전처리를 한다. 세 번재 단계에서는 각 속성은 단일변수값을 갖게끔 수정하며, 명목척도로 측정된 변수의 경우, 0 또는 1 의 값을 갖는 이진 더미(dummy) 변수로 변환해야 추후 해석이 가능하다.

네 번째로 이상치(Outlier)를 제거한다. 네 번째 단계에서는 정상적이지 않은 극단적인 변수값을 조정하여, 모델의 학습이 왜곡되는 것을 예방한다.

다섯 번째로 새로운 파생변수 개발을 한다. 다섯 번째 단계에서는 모델 구축 목적에 부합하는 다양한 독립변수들을 개발하는 등의 작업을 한다.

여섯 번째로 정규화(Normalization) 또는 변수 스케일링 작업을 한다. 여섯 번째 단계에서는 변수가 Range 가 클 때, 정규화 또는 변수 스케일링 작업을 한다.

일곱 번째로 데이터셋을 구분한다. 일곱 번째 단계에서는 과적합화가 되지 않도록 학습 데이터와 테스트 데이터를 나눈 뒤, 테스트 데이터셋을 활용하여 검증한다.

여덟 번째로 분류 모델에 사용될 후보 입력변수를 선정한다. 여덟 번째 단계에서는 카이제

곱 검정, 독립표본 t 검정, 일원배치 분산분석을 통해 분류 모델에 사용될 후보 입력변수를 선정한다.

9. **이상치(outlier)는 세상에 존재할 수 없는 값을 의미하는가? 이상치의 의미를 간략히 서술하고, 이상치를 새로운 값으로 대체할 때 적용할 수 있는 방법들에 대해 설명하시오.**

   이상치는 이상한 값이지, 있을 수 없는 값이 아니다. 이상치란 정상적이지 않은 극단적인 변수값을 조정하여, 모델의 학습이 왜곡되는 것을 예방하는 것이다. 이상치의 경우, 3-Sigma, IQR, Isolation Forest 등의 이상치 감지 기준을 설정 한 뒤, 제거하거나 대체한다.

   이상치를 새로운 값으로 대체 할 때 적용할 수 있는 방법 중 가장 간단한 방법은 평균 또는 중앙값으로 이상치를 대체하는 방법이며, 이 외에는 회귀 모델을 사용하여 이상치를 예측하고 대체하거나, 보간법을 통해 이상치 주변의 데이터를 사용하여 누락된 값을 추정 및 대체하는 방법 등이 있다.

10. **다음날의 주가지수를 예측하는 모형 A와 B가 있다. A는 모형을 구축한 날까지의 주가(과거주가)는 98% 맞춘다. 그런데, 그 다음날부터 주가지수를 예측시켜보니, 70%를 맞추었다. 반면 모형 B는 과거주가는 83% 맞추는데, 미래주가는 78% 맞춘다. 더 잘 구축된 모형은 무엇이겠는가? 그 이유는?**

    모형 B가 더 잘 구축된 모형이라고 할 수 있다. 모형 A는 훈련 데이터에서는 98%를 맞췄지만, 과거 데이터에 지나치게 적합되어 다음날부터 70% 밖에 못 맞추는 반면, 모형 B는 훈련 데이터에서는 비록 83%를 맞췄지만, 일반적인 패턴을 잘 학습하여 미래주가에서 78%를 맞췄다.

    이를 통해, 모형 A의 경우, 모형 B 보다 과적합화(overfitting) 되었다고 판단할 수 있다. 과적합화는 훈련 데이터에 지나치게 적합되어 학습 환경에서는 잘 맞추나, 실제 환경(테스트)에서는 잘 못 맞추므로, 잘 구축된 모형이라고 할 수 없다.

    따라서, 모형 B가 모형 A에 비해 더 잘 구축된 모형이라고 할 수 있다.

11. **이분류 예측모형 (binary classification model) 개발을 위한 학습용 데이터를 준비할 때, 두 집단 간 비중이 1:1 이 되도록 준비해야 하는 이유는 무엇인가? 또한 학습용 데이터(training dataset)와 시험용 데이터(test dataset)를 구분하여 준비해야 하는 이유를 설명하고, 통상 이 두 종류의 데이터는 어떤 비중으로 준비하는지 설명하시오.**

이분류 예측모형 개발을 위한 학습용 데이터를 준비할 때, 두 집단 간 비중이 1:1 이 되도록 준비해야 하는 이유는 데이터 불균형(Data Imbalance) 문제 때문이다. 예시로, 기업부도 예측 데이터의 경우 정상보다 부도의 데이터 양이 훨씬 적을 것인데, 모델 학습없이도 무조건 정상으로 판별하면 정확도가 매우 높게 평가 될 것이다. 이럴 경우, 적절한 모델 학습이 어렵다.

학습용 데이터와 시험용 데이터를 구분하여 준비해야 하는 이유는 과적합화를 방지하기 위해서이다. 통상 적으로 모형 구축 시 과적합화를 예방하기 위해, 모형 구축 시, 테스트 데이터 셋을 활용하여 검증하며, 이 때 전체 데이터가 100 이라면 통상 적으로 학습:테스트=80:20 혹은 70:30 의 비중으로 자료를 미리 나누어 둔다.

12. 정규화(normalization)은 어떤 경우에 필요한 지 설명하고, 가장 많이 사용되는 2 가지 정규화 기법에 대해 설명하시오.

정규화는 변수 간 Range 차이가 클 때 필요하다고 볼 수 있다. 예시로 독립변수 중에 하나는 값의 차이가 크고, 다른 하나는 작다면, 값의 차이가 큰 변수인 경우 해당 가중치가 살짝만 변해도 종속변수(Y)에 막대한 영향을 미치는 반면, 값의 차이가 작은 변수인 경우 상대적으로 종속변수(Y)에 미치는 영향이 작아져, 모델 학습과정에서 값의 차이가 큰 변수가 더 영향을 미치게 될 것이다. 이럴 경우, 정규화 적용을 고려해야 한다.

정규화로 가장 많이 사용되는 정규화 기법은 최소-최대 정규화(Min-Max Normalization)와 Z-Score 정규화가 있다.

최소-최대 정규화(Min-Max Normalization)에서는 최소값이 0, 최대값이 1 이되도록 정규화하며 수식은 아래와 같다.

$$\frac{X - \text{최소값}}{\text{최대값} - \text{최소값}}$$

한 편, Z-Score 정규화는 평균이 0, 표준편차가 1 인 표준정규분포를 갖는 Z-Score 로 정규화하며, 수식은 아래와 같다.

$$\frac{X - \text{평균}}{\text{표준편차}}$$

■ Week03. 비즈니스 애널리틱스 개요 (II)

13. 모집단(population), 모수(parameter), 표본(sample) 및 통계량(statistics)의 관계에 대하여 설명하시오.

모집단은 관심의 대상이 되는 집단 전체를 의미한다.

모수는 모집단이 가지고 있는 특징을 나타내는 수치를 의미한다.

표본은 모수의 값을 알아내기 위해 추출된 모집단의 일부분을 의미한다.

통계량은 모집단의 모수에 대응되는 표본의 특징을 대표하는 값이다.

예시로, 인구주택총조사처럼 전수조사를 할 경우, 대한민국 국민이 모집단이 되며, 모수는 모집단(대한민국 국민)이 가지고 있는 특징을 나타내는 수치를 의미한다.

그러나, 일반적으로 통계분석의 경우, 시간과 비용 등의 문제로 모집단 전체를 전수조사하기 어려우므로, 표본을 선정하여 표본에서 모집단의 모수에 대응되는 표본의 특징을 대표하는 통계량을 토대로 모수를 추정한다.

14. 다음 통계분석들은 두 변수 사이에 통계적으로 유의한 관계가 있는지를 검정할 때 사용하는 방법들이다. 어떤 상황에서 다음 통계분석들을 적용해야 하는지 예를 들어 설명하시오.

⇨ 카이제곱분석 : 대상변수 A 와 B 가 둘 다 이산형인 경우, 두 이산형 변수가 서로 상관관계가 있는지, 없는지를 통계적으로 검정할 때 사용한다.

예시) 성별과 결혼유무 사이에 유의한 관계가 있는가?

⇨ 독립표본 t-검정 : 대상변수 A 가 이산형(2 그룹/독립)인 반면, 대상변수 B 가 연속형인 경우, 서로 독립된 두 집단에 대해 각 집단별 특정 연속형 변수 평균값이 서로 차이가 있는지, 없는지를 통계적으로 검정할 때 사용한다.

예시) 성별에 따른 평균 취업률의 차이가 있는가?

⇨ 대응표본 t-검정 : 대상변수 A 가 이산형(2 그룹/Pair)인 반면, 대상변수 B 가 연속형인 경우, 서로 동일한 모집단에서 추출된 두 표본에 대해 특정 연속형 변수 평균값이 서로 차이가 있는지, 없는지 통계적으로 검정할 때 사용한다.

예시) 보충수업 후, 3 학년 1 반 수학성적의 향상이 있는가?

⇨ 일원배치 분산분석(One-way ANOVA) : 대상변수 A 가 이산형(3 그룹 이상)인 반면, 대

상변수 B 는 연속형인 경우, 셋 이상의 집단에 대해 각 집단별 특정 연속형 변수 평균값이 서로 차이가 있는지, 없는지를 통계적으로 검정할 때 사용한다.

예시) 거주지역에 따른 평균 소득액의 차이가 있는가?

⇨ 회귀분석 : 대상변수 A 와 B 가 둘 다 연속형인 경우, 독립변수(X)와 종속변수(Y)의 관계식을 구하는 기법으로 종속변수의 값을 예측하고, 독립 변수들의 영향력과 관계를 추정하는데 사용한다.

예시) 가계 수입과 사교육비 지출 사이에 유의한 관계가 있는가?

■ Week04. 예측/분류 (I): LOGIT/MDA

15. 다중공선성(multicollinearity)이란 무엇인지 설명하고, 중회귀분석에서 다중공선성의 발생 여부를 확인하는 지표에 대해 설명하시오.

다중공선성은 일부 설명 변수가 다른 설명 변수와 상관관계 정도가 높아 회귀 분석 시 부정적인 영향을 미치는 현상이다. 중회귀분석에서 모든 독립변수가 서로 독립이지 않을 경우, 나타나는 문제이며, 독립변수 사이의 강한 상관관계 때문에 독립변수가 종속변수를 설명하는 변동성이 겹쳐서 발생한다. (잘못된 변수 해석으로 인한 예측 정확도 하락)

중회귀분석에서 다중공선성 발생 여부를 확인하는 지표로 VIF(Variance Inflation Factor)가 있다. VIF 는 다른 변수의 선형 결합을 통해 특정 설명 변수를 설명할 수 있는 정도를 나타내며 식은 아래와 같다.

$$VIF_i = \frac{1}{1 - R_i^2}$$

- x_i를 종속변수, 나머지 변수들을 독립변수로 하여 회귀 모형 적합 가정하였을 때, 회귀 모형의 결정계수 R_i^2를 기반으로 VIF_i 산출
- R_i^2가 커질수록, 분모가 작아져 VIF_i가 커진다.

VIF_i는 i 번째 독립 변수와 다른 모든 변수들 간의 상관관계 제곱으로, i 번째 독립 변수가 다른 독립 변수들에 의해 설명되는 비율을 나타낸다.

일반적으로 VIF 가 10 이상인 경우, 다중공선성이 있는 변수라고 판단한다. (독립적인 변수로서 역할을 못한다고 해석)

16. 통계 기반 분류(예측) 기법인 로지스틱 회귀분석(LOGIT)과 다중판별분석(MDA)의 원리를 간략히 설명하고, 이러한 통계 기반 분류(예측) 기법들의 장·단점에 대해 약술하시오.

로지스틱 회귀분석(LOGIT)의 원리는 일반 회귀분석의 결과를 '로지스틱 변환'하여, 마치 정규분포의 누적확률분포와 유사하게 0 과 1 사이의 값을 갖도록 설계한 것이다.

다중판별분석(MDA)의 원리는 미리 정해진 그룹 간 차이를 잘 설명하여 줄 수 있는 '독립변수들의 선형 결합'을 찾고 함수식(선형판별함수)에 따라 새로운 개체를 분류하는 것이다.

로지스틱 회귀분석과 다중판별분석과 같은 통계 기반 분류(예측) 기법들의 경우, 각 변수를 통계적으로 설명하기에 적합하며, 연산이 빠르다는 장점이 있으나 이러한 통계 기반 분류(예측) 기법은 선형(linear)식에 기반하고 있어, 비선형의 복잡한 현실세계를 충분히 설명하기에 한계가 있다는 단점이 있다.

17. 로지스틱 회귀분석에서 오즈(odds)와 오즈비(odds ratio)에 대해 설명하시오.

오즈(odds)는 특정 사전이 발생하지 않을 확률 대비 특정 사건이 발생할 확률을 의미한다.

$$\left(\frac{p}{1-p}\right) = e^{z}$$

- p는 예측확률을 의미

오즈비(odds ratio)는 독립변수 x 가 한 단위 증가할 때, 종속변수가 0 이 될 확률 대비 1 이 될 확률(즉, 오즈)가 몇 배 증가하는지 나타내는 값으로 e^{β}를 산출한다.

만약, x -> x+1 이 될 때, 특정 사건이 발생할 확률을 p'이라고 한다면,

$$\left(\frac{p'}{1-p'}\right) = e^{\alpha+\beta(x+1)+\epsilon} = e^{\alpha+\beta x+\beta+\epsilon} = e^{\alpha+\beta x+\epsilon} \times e^{\beta}$$

$$\left(\frac{p}{1-p}\right) = e^{z} = e^{\alpha+\beta x+\epsilon}$$

상기 두 식에 의하여, 다음과 같이 오즈 비율 도출

$$e^{\beta} = \frac{\left(\frac{p'}{1-p'}\right)}{\left(\frac{p}{1-p}\right)}$$

■ Week05. 예측/분류 (II): ANN

18. 인공신경망의 원리를 간략히 설명하고, 인공신경망 기법의 장·단점에 대해 약술하시오.

인공신경망의 원리는 인간의 두뇌(neuron)을 모방한 퍼셉트론을 여러 층으로 구성하여, 입력층, 은닉층, 출력층으로 구성된 여러 개의 뉴런들이 서로 연결되어 정보를 처리하고 학습하는 것이다.

인공신경망의 장점으로는 비선형 패턴의 학습이 가능하며, 상대적으로 높은 예측 정확도를 산출하고, 새로운 데이터가 추가 되었을 때, 기존에 학습되었던 결과를 재활용 할 수 있다는 것이 있다.

인공신경망의 단점으로는 많은 양의 학습 데이터를 요구하고, 과적합의 위험성이 매우 높으며, 모델러가 직관적으로 값을 정해야 할 파라미터가 많고, 학습 시 많은 컴퓨팅 자원을 필요로 한다는 것이 있다.

19. 역전파 알고리즘이 어떻게 작동하는지 설명하시오.

1 단계: 초기 연결 가중치 결정

가) 연결가중치를 임의의 아주 작은 값(보통 -1 ~ 1 사이)으로 초기화 한다.

2 단계: 전방향 계산

가) 은닉층 및 출력층에서 입력값에 연결가중치를 곱하여 각 처리요소들의 출력값을 계산한다.

나) 전이함수를 사용하여 출력값을 결정한다.

3 단계: 역방향 계산

가) 출력층의 출력값과 목표출력값 사이의 오류치를 계산한다.

나) 출력층과 은닉층 사이의 연결 가중치를 수정한다.

다) 은닉층과 입력층 사이의 연결 가중치를 수정한다.

4 단계: Epoch(2,3 단계)의 반복

즉, 정리하면, 가중치를 초기화 하고 출력값을 계산한 다음, 출력값과 목표값을 비교하고 가중치를 조정하는 과정의 반복을 통해 작동한다.

20. 인공신경망을 학습할 때 학습용/시험용 데이터셋 외에 검증용(validation) 데이터셋을 추가로 활용해야 하는 이유는 무엇이며, 일반적으로 학습용/검증용/시험용 데이터셋은 어떤 비중으로 구성하는지 설명하시오.

검증용 데이터셋을 추가로 활용해야 하는 이유는 학습된 모델의 성능을 평가하고, 하이퍼파라미터를 조정하는데 사용하기 위해서이다. 인공신경망의 경우, 과적합의 위험으로 인해 학습중지점 선택이 매우 중요하다. 이럴 경우, 최소 오차 도달 뒤, 얼마 이상 학습을 진행했음에도 불구하고 검증용 데이터셋 기준으로 더 이상 개선이 없으면 학습을 중지하는 방식으로 설정한다.

일반적으로, 학습용/검증용/시험용 데이터 셋은 학습용 60%, 검증용 20%, 테스트용 20%으로 비중으로 자료를 미리 나누어 둔다.

■ Week06. 예측/분류 (III): CBR/DT

21. 사례기반추론(CBR, Case Based Reasoning)의 원리를 간략히 설명하고, 사례기반추론 기법의 장·단점에 대해 약술하시오.

사례기반추론은 인간의 학습 방식에서 영감을 받아 탄생하게 된 인공지능 기술로 사례기반 추론의 원리는 4R 로 구성된다. 사례기반추론의 구동 원리인 4R 은 Retrieve / Reuse / Revise / Retain 단계로 구성되며, 보통은 Retrieve 와 Reuse 단계까지만 활용한다. Retrieve 단계가 중요한데, 기존의 경험과 지식을 활용하여 문제를 해결하기 위해 과거 사례 검색 시, 유사하다는 것을 판단을 잘 해야 한다.

사례기반추론기법의 장점은 데이터가 가장 없을 때, 사용하기 좋은 알고리즘으로 둘 중에 하나가 아니라 여러 가지 중에 하나 고를 때, 효과가 좋다. 또한 학습 단계가 사실상 없어 Lazy Learning 이라고 불리기도 하며, 기존 사례를 활용하여 문제를 해결하기 때문에 처음부터 학습할 필요 없이 지연된 계산의 특성을 가지고 있다.

그러나, 사례기반추론기법의 단점은 개별 사례를 통해 매번 사례를 검색해야 하는데 사례의 수가 많을 경우, 검색 시간이 증가 할 수 있으며, 사례 일반화의 어려움 등으로 인해 잘 못 맞춘다는 단점이 있다.

22. **의사결정나무(DT, Decision Tree)의 핵심 작동 원리 2 가지를 설명하고, 의사결정나무 기법의 장·단점에 대해 약술하시오.**

의사결정나무의 핵심 작동 원리는 반복적 분할과 가지치기이다. 반복적 분할의 경우, 학습용 데이터들을 반복해서 분할하여, 세분된 영역 내의 동질성(homogeneity)이 최대가 되도록 한다. 이를 통해 계속 분할해가다 보면, 모든 학습용 데이터를 100% 정확하게 분류해 낼 수 있을 만큼까지 세분화해 나갈 수 있다.

한편, 가지치기의 경우, 반복적 분할을 반복하다 보면, 과적합화는 피할 수 없이 나타나게 되는데, 이러한 과적합화를 피하기 위해 불필요한 가지(정보 제공이 그리 많은 가지)를 제거함으로써 나무를 단순화하는 작업이 이후 이루어지게 된다.

의사결정나무 기법의 장점은 IF-THEN 의 규칙으로 모형을 표현할 수 있으므로, 해석력이 우수하며, SW 로 구현하기 용이하다는 것이다. 또한, 모형이 가볍고 처리속도가 빨라서 대용량 데이터를 처리하는데 적합하다는 장점이 있다.

의사결정나무 기법의 단점은 과적합 문제의 발생 가능성이 높다는 것이다. 즉, 규칙이 많아질 수록, 과적합의 가능성이 높아진다. 또한, 단일 의사결정나무 모델만으로는 높은 정확도의 모델을 만드는데 한계가 있다.

23. **K-NN(K-최근접 이웃) 기법에서 K 를 너무 작게 설정할 경우와 너무 크게 설정할 경우, 각각 어떤 문제가 발생하게 되는지 설명하시오.**

K-NN 에서 K 는 최근접 이웃의 수를 의미한다. K 를 너무 작게 설정할 경우, 과적합(Overfitting) 문제가 있다. 이 경우, 이상치에 민감하게 반응하며, 새로운 데이터에 대한 일반화 능력이 저하 될 수 있다.

반대로, K 를 너무 크게 설정할 경우, 이웃들이 더 많이 포함되므로 전체적인 패턴을 잡아내는 능력이 감소하는 문제가 있다. 이 경우, 모델의 복잡성이 낮아져서 데이터의 다양한 특징을 반영하지 못 할 수 있다.

■ Week07. 예측/분류 (Ⅳ): SVM

24. SVM(Support Vector Machines)의 기본 아이디어 3가지에 대해 설명하고, SVM 기법의 장·단점에 대해 약술하시오.

SVM 의 기본 아이디어는 최대 마진 분류(Maximum Margin Classification), 소프트 마진 분류(Soft Margin Classification), 고차원으로의 사상(Mapping into Higher Dimension)이며, 설명은 아래와 같다.

1) 최대 마진 분류: 두 집단을 구분하는 선형 구분자와 각 집단 경계에 잇는 점 사이의 거리, 즉 마진을 최대화하는 구분자를 선택하면, 두 집단을 가장 효과적으로 분류하는 분류기가 될 것이라는 논리이다.

2) 소프트 마진 분류: 실제로는 데이터가 완벽하게 구분되는 경우가 없어서, 여유변수(Slack variable) 도입이 필요한데, 모델에 여유변수가 추가됨으로써, 잡음이나 분류 난이도가 높아 틀릴 수 밖에 없는 사례들을 적절하게 처리 가능하다.

3) 고차원으로의 사상: 저차원에서 선형 분리가 잘 안되는 분류 문제가 있다면, 커널 함수를 이용해 데이터를 고차원으로 사상(mapping)하는 기술이다.

SVM 기법의 장단점은 아래와 같다.

SVM 기법의 장점으로는 선형 분류와 비선형 분류 모두 가능하며, 구조적 위험 최소화를 구현하고, 통제해야 하는 파라미터의 수가 적으며, 적은 양의 데이터로도 높은 정확도를 보인다는 것이다. 이러한 장점들을 통해 전통적인 인공신경망에 비해 컴퓨팅 소요가 적으면서도 비슷하거나 더 우수한 예측 정확도를 제공하고 일반화 성능이 우수하여 과적합의 위험을 낮으며, 학습용 데이터의 양에 모형이 받는 영향이 적다.

SVM 의 단점으로는 계산 비용이 클 수 있으며, 입력 데이터의 스케일에 민감하며, 분류 문제의 경우, 기본적으로 이분류(binary classification)만 가능하다는 것이 있다. 따라서, 입력 데이터의 스케일이 클 경우, 적용 시 적절한 전처리(정규화)가 필요하며, 다분류로 확장하기 위해서는 특별한 접근법이 필요하다.

25. **이분류 모형의 성능을 평가하는 지표인 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F1 점수에 대해 설명하시오.**

		예측 (Predicted)	
		1(+)	**0(-)**
실제 (Actual)	**1(+)**	**True Positive (TP)** *Sensitivity*	**False Negative (FN)** *Type II Error*
	0(-)	**False Positive (FP)** *Type I Error*	**True Negative (TN)** *Specificity*

정확도(Accuracy)= 혼동행렬 전체 중에 올바르게 예측한 비율

$$\frac{(TP + TN)}{(TP + FP + TN + FN)}$$

정밀도(Precision) = 1(+)로 예측한 표본 중 실제 얼마나 정확하게 맞추었는지에 대한 비율

$$\frac{TP}{(TP + FP)}$$

재현율(Recall) = 실제 1(+)의 표본 중 얼마나 정확하게 예측했는지에 대한 비율

$$\frac{TP}{(TP + FN)}$$

F1 점수= Precision 과 Recall 의 조화평균

$$F_1 = \frac{2}{\frac{1}{recall} + \frac{1}{precision}} = 2 \times \frac{precision \times recall}{precision + recall}$$

26. **이분류 모형의 성능을 평가하는 지표인 민감도(Sensitivity), 특이도(Specificity), ROC 곡선과 AUROC 에 대해 설명하시오.**

민감도= 실제 1(+)의 표본 중 얼마나 정확하게 예측했는지에 대한 비율로 재현율(Recall)과 같다.

$$\frac{TP}{(TP+FN)}$$

특이도=실제 0(-)의 표본 중 얼마나 정확하게 예측했는지에 대한 비율이다.

$$\frac{TN}{(FP+TN)}$$

ROC(Receiver Operating Characteristic) 곡선= 구축한 모형의 성능을 민감도와 1-특이도에 의해 판단하고자 할 때, 이를 시각적으로 식별할 수 있도록 도와주기 위하여 만들어진 곡선

AUROC(AUC, Area Under ROC)= ROC 곡선 아래의 면적을 수치화 한 값 (0.5 ~ 1 사이)

27. 이분류 모형의 0 과 1 을 판별하는 기준 임계값을 0.5 에서 0.8 로 올릴 경우, 정밀도(Precision)와 재현율(Recall)은 어떻게 변화하는지 기술하고, 그 이유를 설명하시오.

정밀도는 증가하고 재현율은 감소한다.

이유:

기준 임계값을 높일 경우, 1(+)으로 예측하는 표본들이 적어지면서, 정밀도 공식에서 분모 구성요소인 FP 가 작아지기 때문에 정밀도는 증가한다.

반면에, 1(+)으로 예측하는 표본들이 적어지고, 0(-)으로 예측하는 표본이 많아지면서, 재현율 공식에서 분모 구성요소인 FN 이 커지기 때문에, 재현율은 감소한다.

28. 이분류 모형의 0 과 1 을 판별하는 기준 임계값을 0.5 에서 0.2 로 낮출 경우, 민감도(Sensitivity)와 특이도(Specificity)는 어떻게 변화하는지 기술하고, 그 이유를 설명하시오.

민감도는 상승하고 특이도는 감소한다.

이유:

기준 임계값을 낮출 경우, 1(+)으로 예측한 표본들이 많아지면서, 민감도 공식에서 분자인 TP 값이 커지기 때문에 민감도는 상승한다.

반면에, 1(+)으로 예측한 표본들이 많아지면서, 특이도 공식에서 분모 중에 FP 가 커지므로, 특이도는 감소하게 된다.

■ Week08. 예측/분류 (V): MSVM

29. Multiclass SVM 을 구현하는 전통적인 6 가지 접근법에 대해 간략히 설명하시오.

Multiclass SVM 을 구현하는 전통적인 접근법에는 다수의 이분류 모델들을 적절하게 결합하

여 다분류를 수행하는 방식의 One-Against-All, One-Against-One, DAGSVM(Directed Acyclic Graph SVM), ECOC(Error-Correcting Output Code) 접근법과 SVM 의 원리를 반영한 단일 다분류 모델을 생성하는 방식의 Method by Weston & Watkins, Method By Crammer & Singer 접근법이 있다. 각각의 설명은 아래와 같다.

1) One-Against-All: k 개의 클래스로 분류해야 하는 문제인 경우, 총 k 개의 이분류기(binary classifier)를 구축하는 접근법

2) One-Against-One: k 개의 클래스로 분류해야 하는 문제인 경우, 총 $C_2^k = \frac{k!}{(k-2)!2!} = \frac{k(k-1)}{2}$개의 이분류(binary classifier)를 구축하는 접근법

3) DAGSVM(Directed Acyclic Graph SVM): 학습 단계에 있어서는 One-Against-One 방법과 동일하나, 시험(검증) 단계에서 적용할 때에는 그래프 방문(graph-visiting) 전략을 사용하여 보다 효율적으로 적용 가능한 접근법

4) ECOC(Error Correcting Output Coding): 두 집단을 구분하는 개별 이분류기들의 판단을 적절하게 융합하여 활용(새로운 데이터가 들어올 시, Hamming Distance 를 가장 최소화하는 Class 선정)하는 접근법

5) Method by Weston & Watkins: 이분류 SVM 모델을 자연스럽게 다분류 문제로 확장하여 OAO 모델들을 종합적으로 고려하여, 점수를 산출 한 뒤, 가장 높은 점수를 출력하는 클래스로 분류를 수행하는 접근법

6) Method by Crammer & Singer: 기본 원리는 Weston & Watkins 와 동일하나, 여유변수를 덜 사용함으로써 보다 효율적으로 다분류 SVM 을 수행할 수 있도록 설계된 접근법

30. Ordinal Multiclass SVM 은 어떤 원리로 작동하는지 설명하고, 이 기법의 장점을 설명하시오.

클래스의 순서 정보(order information)를 활용하여, 다수의 이분류 기계학습 모델을 구축하는데, 이 때, Fusing Method 와 기준으로 Forward, Backward 와 Partitioning Method 기준으로 One-Against-TheNext, One-Against-Followers 으로 나눈다. Ordinal Multiclass SVM 의 4 가지 하위 유형과 작동원리는 1vsFollowers - Forward, 1vsFollowers - Backward, 1vsTheNext - Forward, 1vsTheNext - Backward 4 가지 원리로 작동한다.

Oridinal Multiclass SVM 기법의 경우, 클래스의 순서 정보를 활용하여, 더 적은 분류기를 사용하면서도 더 우수한 예측 성능을 산출할 수 있다는 장점이 있다.

31. 예측(분류)를 위해 다수의 모형을 결합하고자 할 때 적용할 수 있는 모형 결합 기법의 4 가지 유형(접근법)을 설명하시오.

1) Preprocessor: 다양한 전처리 기법을 사용하여 입력 데이터를 변환한 후 개별 모델에 적용하는 방식으로, Feature Discretization, Fuzzy theory 등이 있다.
2) Embedded: 모델 자체 내에서 모형 결합을 구현하는 방식으로 GA-ANN, GA-CBR, GA-SVM 등이 있다.
3) Decomposition: 입력 데이터를 여러 개의 부분으로 분해한 후, 각 부분에 다수의 모형을 적용하는 방식으로 Hybrid Recommender System (CB+CF)이 있다.
4) Ensemble: 다수의 약한(weak) 분류기들을 결합하여 하나의 강한(strong) 분류기를 생성하는 접근법으로, Bagging, Boosting, Random Forest 등이 있다.

32. 앙상블(Ensemble) 기법의 2 가지 유형을 설명하고, 각각의 사례를 제시하시오.

앙상블 기법의 2 가지 유형은 Homogeneous Ensemble Models 와 Heterogeneous Ensemble Models (Stacking)이 있다.

Homogeneous Ensemble Models 은 무작위로 학습용 데이터를 복원추출한 다수의 모델을 병렬적으로 학습한 뒤, voting 으로 결과를 산출하는 Bagging(Bootstrap aggregating)과 새로운 모델을 학습할 때, 기존 모델에서 틀린 학습용 표본을 더 강조하여 학습하는 Boosting 이 있다.

Heterogeneous Ensemble Models (Stacking)은 서로 다른 유형의 모델을 약한 분류기로 사용한 뒤 결합하며, 경우에 따라서는 이종의 모델들이 산출한 결과를 다시 한 번 학습하여 최종 예측결과를 산출하는 메타모델(Meta-model)을 개발하여 사용할 수도 있다.

33. 기업부도예측을 위한 3 개의 모형들이 아래와 같이 예측값(0: 정상, 1: 부도)을 산출하였다. ① Simple Average, ⓦeighted Average, ⓢelecting the Best Predictor, ⓜajority Voting 으로 모형 간 결합을 할 경우, 각각 예측결과가 어떻게 산출될 지 설명하시오.

모형 1: 로지스틱회귀모형 (예측 정확도: 20%)	모형 2: 인공신경망 (예측정확도: 60%)	모형 3: 의사결정나무 (예측정확도: 40%)
0.6	0.3	0.9

① Simple Average: $(0.6 + 0.3 + 0.9) \div 3 = 0.6$ 결과값이 나왔을 때, 0.5 보다 크므로 1 로 분류하여 예측결과를 부도로 산출한다.

② Weighted Average: $(0.6 \times 0.2) + (0.3 \times 0.7) + (0.9 \times 0.4) \div (0.2 + 0.6 + 0.4) = 0.575$ 결과값이 나왔을 때, 0.5 보다 크므로 1 로 분류하여 예측결과를 부도로 산출한다.

③ Selecting the Best Predictor: 0.9 ($\because |0.9 - 0.5| > |0.3 - 0.5| > |0.6 - 0.5|$) 결과값이 나왔을 때, 0,5 보다 크므로 1 로 분류하여 예측결과를 부도로 산출한다.

④ Majority Voting: 1 ($\because 0 \rightarrow 0$ 표, $1 \rightarrow 3$ 표)이므로 예측결과를 부도로 산출한다.

■ Week09. 최적화 (I): 선형계획법/정수계획법

34. 아래 상자에 설명된 시나리오를 잘 읽고, 해당 시나리오의 의사결정문제를 해결하기 위한 선형계획모형을 도출하시오.

> 국민농장은 500 에이커의 땅에 콩과 옥수수를 재배하고 있다. 콩은 한 에이커에서 10 만원의 이익을 보고, 옥수수는 한 에이커에서 20 만원의 이익을 보고 있다. 정부의 정책으로 인하여, 콩은 200 에이커 이상 심지 않으려 하고 있다. 작물재배에 동원할 수 있는 노동력은 총 1,200 시간인데, 콩 한 에이커 경작에는 2 시간, 옥수수 한 에이커 경작에는 6 시간이 소요된다. 이러한 상황에서, 국민농장은 자사의 이익을 최대화 하는 콩과 옥수수의 재배계획을 수립하고자 한다.

$$Let.$$

$$x_1 = Number\ of\ acres\ for\ soybean\ cultivation$$

$$x_2 = Number\ of\ acres\ for\ corn\ cultivation$$

$$Max.$$

$$z = 100000\ \times x_1 + 200000\ \times\ x_2$$

$$s.t.$$

$$x_1 + x_2 \leq 500$$

$$x_1\ <\ 200$$

$$2 \times x_1 + 6 \times x_2 \leq 1200$$

35. 아래 상자에 설명된 시나리오를 잘 읽고, 해당 시나리오의 의사결정문제를 해결하기 위한 선형계획모형을 도출하시오.

경제난에 시달리고 있는 기택씨네 장남 기우군은 최소의 비용으로 체내에 필수적인 단백질과 지방을 섭취하고자 한다. 하나에 40 원인 달걀 1 개를 섭취할 경우, 단백질과 지방 모두 4mg 씩 섭취할 수 있다. 반면 개당 60 원씩 판매되는 베이컨을 1 개 섭취할 경우, 단백질은 3mg, 지방은 5mg 을 섭취할 수 있다. 성인에게 필요로 되는 하루 최소 단백질 요구량은 20mg, 지방 요구량은 28mg 이라고 할 때, 기우군이 목적을 달성하기 위해서는 매일 달걀과 베이컨을 각각 몇 개씩 섭취해야 하는가?

$$Let.$$

$$x_1 = Number\ of\ eggs\ consumed$$

$$x_2 = Number\ of\ bacons\ consumed$$

$$Min.$$

$$z = 40\ \times x_1 + 60\ \times\ x_2$$

$$s.t.$$

$$4\ \times\ x_1 +\ 3\ \times x_2\ \geq\ 20$$

$$4 \times x_1 + 5 \times x_2\ \geq\ 28$$

36. 아래 상자에 설명된 시나리오를 잘 읽고, 해당 시나리오의 의사결정문제를 해결하기 위한 선형계획모형을 도출하시오.

작은 가구 공장을 운영 중인 김쿠민씨는 매출의 최대화를 원한다. 김쿠민씨는 매일 50 ㎏의 원목을 제공받는다. 공장에는 노동자가 10 명 있으며, 이들은 하루에 8 시간 근무한다. 책상을 만들려면 3 시간의 노동시간과 5 ㎏의 원목이 필요하고, 의자를 만들려면 4 시간의 노동시간과 2 ㎏의 원목이 필요하다. 책상과 의자는 각각 22 만원, 20 만원에 팔린다. 김쿠민씨는 책상과 의자를 매일 몇 개씩 생산해야 하는가?

$$Let.$$

$$x_1 = Number\ of\ desks\ produced$$

$$x_2 = Number\ of\ chairs\ produced$$

$$Max.$$

$$z = 220000\ \times x_1 + 200000\ \times\ x_2$$

$$s.t.$$

$$3\ \times\ x_1 +\ \ 4\ \times x_2\ \leq\ 80$$

$$5 \times x_1 + 2 \times x_2\ \leq\ 50$$

■ Week10. 최적화 (II): 유전자 알고리즘(GA)

37. 다음 상자 안에 제시된 용어들의 의미를 포함하여, 유전자 알고리즘(genetic algorithm)의 구동 절차에 대해 설명하시오.

염색체(chromosome)
모집단(population)
적합도(fitness) 및 적합도 함수(fitness function)
유전 연산(genetic operators)
Roulette Wheel 방식 선택
Rank-based 방식 선택
Single-point 교배
Two-point 교배
Uniform 교배
돌연변이

유전자 알고리즘의 절차는 크게 4 단계로 나눌 수 있다.

Coding or Representation:

✔ 후보 해를 구성하는 모든 의사결정 변수값들을 하나의 문자열로 표현한다.(보통 이진 배열로 표현)

- 해(solution)을 구성하는 요소를 유전자라고 하는데, 유전자의 조합으로 후보 해(candidate solution)을 구성한다.

✔ 초기 모집단(population) 형성(랜덤하게 생성)

- 모집단은 염색체의 집합체로서, 후보 해들의 집합이다.

Fitness function:

✔ 각 염색체(chromosome)의 적합도 산출한다.

- 염색체는 주어진 문제에서 하나의 후보 해를 나타내는 자료구조로 주로 유전자(gene)의 스트링이거나 이진배열을 의미한다.

✔ 그 중 가장 적합도가 높은 염색체들을 부모로 선택

- 적합도는 염색체(후보 해)의 우성 또는 열성 수준을 판단하는데 사용되는 평가 기준으로 적합도 함수(fitness function)에 의해 평가되며, 이 값이 높을수록 해로서의 가치가 높다.
- 모집단의 다음세대를 생성하기 위해 적용되는 유전 연산(Genetic Operators)은 선택, 교차, 돌연변이 등이 해당된다.
- 부모를 선택(Selection)하는 경우, 우성인 성질을 포함하고 있는 염색체를 부모로 선택하는데, 주로 Roulette Wheel 방식을 적용한다. 단, Evolver 는 Rank-based 방식을 사용한다. (상대적으로 더 smooth 하게 선택)
- Roulette Wheel 방식의 경우, 염색체의 적합도를 기반으로 선택 확률을 계산하여 염색체를 선택하고, Rank-based 방식의 경우, 염색체들을 적합도 순서대로 정렬한 후, 순위에 따라 선택하는 방식이다.

Reproduction

✔ 선택된 부모들을 교배(Crossover), 돌연변이(Mutation)하여 후손 생성

- 교배의 경우, 선택된 두 염색체 간 교배를 의미하며, 두 후손 산출한다.
- Single-point 교배의 경우, 두 개의 부모 염색체를 한 지점에서 분할하여 교배하는 방식으로 분할된 지점을 기준으로 부모 염색체의 유전자를 교환한다.
- Two-point 교배의 경우, 두 개의 부모 염색체를 두 지점에서 분할하여 교배하는 방식이다. 두 분할 지점 사이에 있는 부분을 서로 교환한다.
- Uniform 교배의 경우, 사전 설정된 교배율에 따라 교배할 위치를 랜덤하게 선택하고, 선택되지 않은 부분에 대해서는 기존 유전자를 유지한다.
- 돌연변이의 경우, 하나의 염색체를 입력 받아, 확률에 따라 무작위로 변형한다. 유전자 알고리즘이 국지 최적해에 빠지지 않도록 막아주는 예방장치의 역할을 수행한다.

Convergence

✔ 종료 조건을 충족할 경우, 진화 과정을 중지하고 그렇지 않다면, Fitness function 과 Reproduction 단계를 반복

38. 선형계획법(linear programming)과 유전자 알고리즘(genetic algorithm)은 모두 어떤 문제의 최적해를 찾아내는데 사용되는 기법들이다. 그렇다면, 두 기법의 차이점은 무엇이며, 현실세계의 경영 문제를 해결하는데 있어 적용가능성(applicability)이 더 우수한 기법은 어떤 것이라 생각하는가?

선형계획법은 선형 조건을 만족시키면서 선형 목적함수를 최대화하거나 최소화하는 변수들의 최적값을 도출하는 기법이다. 심플렉스 해법을 통해 확실한 최적해를 수학적으로 산출할 수 있으나, 불확실성이 존재하는 확률적 상황에서 사용할 수 없다. 특히, 목적함수와 제약식 모두 반드시 선형식이어야 하는 제약으로 인해 비현실적인 측면이 있다.

한 편, 유전자 알고리즘은 수리적인 모형으로 해결이 불가능한 비선형 최적화 문제를 탐색을 통해, 풀 수 없는 문제에 대한 (유사)최적해를 효율적으로 탐색하는 기법이다. 비선형 제약조건 및 비선형 목적함수 환경 하에서도 최적화가 가능하며, 기울기 정보를 사용하지 않아 함수의 연속성이나 미분가능성 등의 제약 없이 최적화가 가능하고, 비교적 빠른 시간 내에 적당한 해를 찾을 수 있다.

물론, 선형계획법이 목적함수와 제약식 모두 반드시 선형식이어야 한다는 측면에서 비현실적인 측면이 있지만 그럼에도 불구하고 현실세계의 경영 문제의 경우, 선형계획법이 더 우수한 기법이라고 생각한다.

이유는 현실세계의 경영 문제의 경우, 생산 계획, 자원 할당, 운송 문제 등의 선형적인 형태로 모델링 가능한 문제들이 많으며, 효율성 및 해석 가능성 측면에서 확실한 최적해를 산출하고 해석이 용이해야 하는 경우가 많기 때문이다.

그렇다고해서 선형계획법이 항상 우수하다고만은 생각하지 않는다. 예시로 현실세계의 경영 문제 중 마케팅 분야에 있어서 다양한 제약 조건이 제약 조건으로 작용하며 제약 조건들이 비선형적이거나 상호작용을 가지는 경우, 혹은 다중 목적을 동시에 고려하여 다양한 목표를 동시에 최적화해야 하는 경우 등등 에서는 오히려 유전자 알고리즘이 더 우수한 기법이 될 수 있다.

따라서, 현실세계의 경영 문제를 해결하는 데 있어서 선형계획법이 적용가능성이 더 우수한 기법이라고 생각하나, 그렇다고 해서 항상 선형계획법이 무조건 우수한 기법이라고는 할 수 없다고 생각한다.

■ Week11. 가치평가: 분석적 계층화 과정(AHP)과 자료포락분석(DEA)

39. 분석적 계층화 과정(AHP, Analytic Hierarchy Process)의 적용 과정을 단계별로 설명하시오.

1 단계: 의사결정을 위한 계층(Hierarchy) 모델 설계

- 모든 연관된 평가기준들이 포함되어야 한다.
- 모든 평가기준들은 계층적 관계를 갖도록 설계한다.

  ex) 1 단계 기준 하위 요소 -> 2 단계, 2 단계 기준 하위요소 -> 3 단계…
- 모든 계층의 구성요소들은 서로 독립이어야 한다.

2 단계: 각 기준의 상대적 중요도 산출

- 비율척도로 쌍대비교를 수행하여 쌍대비교 결과를 쌍비교행렬로 정리한다.
- 고유벡터의 개념을 활용하여 각 기준의 상대적 중요도를 산출한다.

3 단계: 상기 도출된 가중치에 의해, 대안 평가 및 대안별 최종 점수를 산출

4 단계: 민감도 분석 수행

- 일관성 비율$\left(\textit{일관성지수} \div \textit{랜덤지수}\right)$ 계산한다.
- 일관성 지수는 응답자가 일관성 있게 응답할수록 작은 값 도출된다.
- 랜덤지수는 비교대상이 많아질 때 생기는 일관성 없는 답변을 보정하기 위한 수단이다.

5 단계: 다수의 전문가들의 응답을 종합

- 순위역전 현상 발생 방지를 위해 기하평균을 사용하는 것이 바람직하다.

40. 분석적 계층화 과정(AHP, Analytic Hierarchy Process)에서 일관성 비율(consistency ratio)은 무엇며, 이것을 왜 고려해야 하는지 예를 들어 설명하시오. 아울러, 보통 이 값이 어느 수준의 조건을 충족해야 유효한 설문으로 인정되는지 설명하시오.

일관성 비율은 일관성 지수를 랜덤지수로 나눈 것이다.

- 일관성 지수는 응답자가 일관성 있게 응답할수록 작은 값 도출된다.
- 랜덤지수는 비교대상이 많아질 때 생기는 일관성 없는 답변을 보정하기 위한

수단이다.

일관성 비율을 고려하는 이유는 설문을 신뢰할 수 있는지를 확인하기 위해서이다. 예시로, 평생 함께할 반려자(배우자) 후보의 평가모형에서 "성격보다 가치관이 중요하다."를 선택하고, "가치관보다 외모가 더 중요하다."를 선택하고 이어서, "외모보다 성격이 더 중요하다."를 선택한다면 논리상 맞지 않는다.

일관성 비율은 응답자가 일관성 있게 응답했는지를 확인하여 설문을 신뢰할 수 있는지 확인할 수 있으며, Saaty 는 일관성 비율이 0.1 보다 작아야 그 설문을 신뢰할 수 있다고 하였지만, 일반적으로 실무에서 일관성 비율이 0.1 보다 작은 설문이 많지 않아서, 0.2 를 적용하는 것이 일반적이다.

41. 분석적 계층화 과정(AHP, Analytic Hierarchy Process)을 적용할 때 각별히 유의해야 할 사항들로는 어떤 것들이 있는가?

분석적 계층화 과정을 적용할 때 각별히 유의해야 할 사항들로는 크게 3 가지 정도 있다.

1) 여러 응답자들의 서로 상반되는 의견들을 어떻게 종합할 것인가?
   - 응답자들이 충분한 컨센서스를 이룬 상태에서 응답에 참여하게끔 해야 함.
2) 계층화 모델을 구성하는 기준들이 확실하게 MECE(Mutually Exclusive, Collectively Exhaustive)하게 설계되어 있는가?
   - 사후적으로 검증할 방법이 전무하므로, 설계를 잘해야 한다.
3) 응답자들의 중심화 경향을 어떻게 해결할 것인가?
   - 강제로 후처리(post-processing)하여, 자료를 분석하는 것이 더 효과적일 수 있다.

42. 자료포락분석(DEA, Data Envelopment Analysis)의 원리와 사용처, 특징(장점) 등에 대해 설명하시오.

자료포락분석은 의사결정단위의 상대적 효율성을 평가하는데 사용되는 성과측정 기법이다.

자료포락분석의 원리는 아래 공식을 통해 상대적 효율성을 측정할 수 있는 것이다.

$$100 \times \frac{Length\ of\ line\ from\ origin\ to\ [DMU\ X]}{Length\ of\ line\ from\ origin\ through\ [DMU\ X]\ to\ efficient\ frontier}$$

자료포락분석은 화폐 단위의 평가를 하지 않는다. 그러나, 자료포락분석은 성과평가 수행에 효과적이므로, 사용처로 공공기관이나 NGO 같이 비영리단체 같은 조직의 성과를 평가하고자 할 때 , 성과평가를 수행할 수 있어서 효과적으로 적용될 수 있다.

자료포락분석은 상대적 효율성을 평가할 수 있다는 특징을 가지고 있으며, 입력과 산출물에 대한 데이터만 제공되면, 성과평가를 수행할 수 있어서 유리하다는 장점을 가지고 있다.

■ Week12. 군집화(Clustering)

43. 계층적 군집분석과 비계층적 군집분석의 특징을 설명하고, 이 중 현실 문제해결에 더 많이 사용되는 군집분석은 어떤 것인지 설명하시오.

계층적 군집분석은 군집과 군집 간 포함관계가 형성하고 군집의 개수를 사후적으로 결정 가능하다는 특징이 있다

비계층적 군집분석은 상호배타적 군집 형성이며 일반적으로 미리 군집의 개수를 정하고 시작해야한다는 특징이 있다.

44. 계층적 군집분석 수행 시, 개체와 군집 혹은 군집과 군집 사이의 거리를 계산하는 방법들로는 어떤 것들이 있는지 설명하시오.

계층적 군집분석 수행 시, 개체와 군집 혹은 군집과 군집 사이의 거리를 계산하는 방법은 평균 연결법(집단-간), 평균 연결법(집단-내), 최단 연결법, 최장 연결법, 중심 연결법, 중위수 연결법, Ward 의 방법이 있다. 각각의 방법에 대한 설명은 아래와 같다.

1) 평균 연결법(집단-간): 각 개체와 군집 사이의 평균 거리를 사용하여 군집 간 거리를 계산하는 방법.
2) 평균 연결법(집단-내): 군집 내의 모든 개체 쌍 간의 평균 거리를 사용하여 군집 간 거리를 계산하는 방법.
3) 최단 연결법: 두 군집 내에서 가장 가까운 개체 쌍 간의 거리를 사용하여 군집 간 거리를 계산하는 방법.
4) 최장 연결법: 두 군집 내에서 가장 먼 개체 쌍 간의 거리를 사용하여 군집 간 거리를 계산하는 방법.
5) 중심 연결법: 두 군집의 중심(평균) 사이의 거리를 사용하여 군집 간 거리를 계산하는 방법.
6) 중위수 연결법: 두 군집 내에서 중위수(중간값)에 해당하는 개체 쌍 간의 거리를 사용하여 군집 간

거리를 계산하는 방법.

7) Ward 의 방법: 군집을 키웠을 때 발생하는 군집 내의 분산 증가를 최소화하는 방식으로 군집 간 거리를 계산하는 방법.

45. K-Means 알고리즘의 작동 과정과 장·단점에 대해 간략히 설명하시오.

K-Means 알고리즘은 사전에 정해진 군집의 개수(K)에 따라 데이터를 군집으로 그룹화 하는 비지도학습으로, K-Means 알고리즘의 동작과정은 아래와 같다.

초기 중심 선택: K 개의 군집 중심점(임시)을 설정

할당 단계(Assign Step): K 개의 군집 중심점을 기준으로 각 데이터 포인트를 가장 가까운 중심에 할당하여 군집화 수행

업데이트 단계(Update Step): 각 군집별 중심점을 계산하여, 새로운 중심점 설정

할당과 업데이터의 반복: 할당 단계와 업데이트 단계를 반복하여 군집의 중심과 할당을 최적화하는 단계(중심의 변화가 없거나 미리 정한 반복 횟수에 도달 할 때까지 계속 반복)

46. 자기조직화지도(SOM, Self-Organizing Maps) 알고리즘의 작동 과정과 장·단점에 대해 간략히 설명하시오.

SOM 은 비지도 학습 기법 중 하나로 고차원의 데이터를 저차원(2 차원)의 격자 형태로 매핑하는 알고리즘으로, SOM 의 구동 절차는 아래와 같다.

1) 2 차원 격자로 된 격자 셀(Cell)을 생성하고, 각 격자 셀의 가중치 벡터(Weight Vector)를 적절하기 초기화 한다. (보통 0~1 사이의 작은 임의의 값(random value)으로 설정)

2) 학습데이터를 하나$\left(X_i = (x_{i1}, x_{i2}, x_{i3}, \ldots\ldots, x_{in})\right)$씩 불러서, 학습 데이터와 거리가 가장 가까운 가중치 벡터를 가진 뉴런$\left(W_j = \left(w_{j1}, w_{j2}, w_{j3}, w_{jn}\right)\right)$을 찾음

$$Min\ D(j) = \sqrt{\sum_{k=1}^{n}\left(x_{ik} - w_{jk}\right)^2}$$

3) 학습률을 α라고 할 때, 가장 가까운 벡터를 가진 선택된 뉴런(Winner)의 가중치 벡터를 다음 산식에 의해 업데이트

$$W_j^{NEW} = W_j + \alpha \mid\mid X_i - W_j \mid\mid$$

4) Winner 주변의 노드들(Neighborhood)의 가중치 벡터 역시 업데이트 하나, 적용할 때 Neighborhood 는 Winner 로부터 멀리 떨어져 있을수록 보다 더 작은 값을 적용

5) 2 단계 ~4 단계까지를 종료 조건이 충족될 때까지 반복 수행

자기조직화지도(SOM)의 장점은 아래와 같다.

1) 시각화와 해석가능성: 고차원 데이터를 저차원의 격자 형태로 매핑하여 시각화 함으로써 시각적으로 이해하기 용이하다.
2) 사전에 군집의 개수를 정하지 않고, 어떻게 군집들이 묶이는지 확인 가능하다.
3) 차원 축소 도구로도 활용 가능하다. (고차원 → 저차원의 격자)

자가조직화지도(SOM)의 단점은 아래와 같다.

1) 계산 복잡성: 큰 규모의 데이터셋에 대해서 SOM 을 적용하는 것은 계산적으로 복잡하다.
2) 과적합 가능성: 기본적으로 신경망의 학습 원리를 따르고 있어서, 이상치에 과적화될 위험이 존재한다.
3) 편리하게 SOM 을 구현할 수 있는 도구의 부재: 격자 형태의 매핑과 시각화는 지원하지 않지만, Neuroshell 2 를 이용하면 SOM 군집분석 가능

■ Week13. 추천시스템과 딥러닝

47. 내용기반(CB, Content-Based) 추천기법의 원리와 장·단점에 대해 간략히 설명하시오.

내용 기반 추천 기법은 목표 고객이 과거에 어떤 특성을 가진 상품을 선호했는지를 기준으로 유사한 특성을 가진 상품을 추천하는 방식이다.

내용 기반 추천 기법의 원리는 모든 아이템 쌍 간의 유사도(ex: Cosine Similarity)를 계산하여 유사도 값을 아이템 간의 상관 관계 또는 상호작용을 나타내는 Item-to-Item Correlation 으로 사용하는 것이다.

내용 기반 추천 기법의 장점은 아래와 같다.

1) 직접적이고 단순하여, 구축과 적용이 간단하다.
2) 새로 출시된 상품도 상품의 특성만 모델링되면 바로 추천 대상이 될 수 있다.
3) 추천을 하는데 사용되는 기준이 명확하여, 왜 그런 추천을 하게 되었는지 설명이 가능하다.

내용 기반 추천 기법의 단점은 아래와 같다.

1) 추천결과가 대체로 잘 안맞는다.
2) 사람이 개입해야 하는 상품 특성 추출이 어렵다.
3) 과거에 입력된 정보에 너무 의존하는 경향이 있다.

48. 협업필터링(CF, Collaborative Filtering) 추천기법의 원리와 장·단점에 대해 간략히 설명하시오.

협업 필터링 추천 기법은 목표 고객과 유사한 선호도를 보인 다른 고객이 구매한 결과를 바탕으로 추천결과를 생성하거나, 목표 고객이 구매한(혹은 선호한) 상품과 유사한 상품을 바탕으로 추천결과를 생성하는 기법이다.

협업필터링 추천기법의 원리는 사용자-아이템 행렬을 생성하고 사용자 혹은 아이템 간의 유사도를 계산하여, 유사도를 기반으로 가장 유사한 이웃 사용자들이나 아이템을 선택한 뒤, 선택된 이웃들을 기반으로 사용자가 구매하지 않은 아이템에 대한 예측을 생성하는 것이다.

협업 필터링의 추천기법의 장점은 아래와 같다.
1) 상대적으로 높은 추천 정확도
2) 고객의 인구통계 특성이나 상품의 특성을 파악 없이, 구매여부나 선호도만 활용

협업 필터링 추천기법의 단점은 아래와 같다.
1) 고객의 구매 정보가 많아야 적용 가능(Sparsity Problem)
2) 신규 고객이나 새로운 상품에 대해서는 추천하지 못함(Cold Start Problem)
3) 고객수가 늘어날수록, 요구되는 계산 시간이 급속하게 증대(Scalability Problem)
4) 고객들이 다들 유사하게 상품을 구매하면, 추천 결과 생성 못함

49. 최신의 추천시스템 연구동향인 다기준(multi-criteria) 추천과 세렌디피티(serendipity) 추천, 그리고 그레이쉽(grey sheep) 추천에 대해 간략히 설명하시오.

1) 다기준 추천: 사용자의 니즈를 보다 정확하게 파악하면, 보다 정확한 추천이 가능하다. 각 기준에 가중치를 부여하거나 선호도를 설정하여 사용자의 선호도에 맞게 추천을 조정
2) 세렌디피티 추천: 사용자의 기존 취향이나 선호도와는 다소 다른 항목을 추천하여, 새로운 경험을 제공하고 다양성을 확장시키는 추천 기술로, 필터 버블(자신의 관점에 동의하지 않는 정보로부터 분리되어 다양성이 제한되는 현상)문제를 해결하는데 중요한 역할을 함

3) 그레이쉽 추천: 사용자가 다수의 일반적인 그룹이나 카테고리에 속하지 않는 중립적인 위치에 있는 경우, 해당 사용자를 Grey Sheep 이라고 부른다. 그레이쉽 추천은 사용자들 사이에서 중립적인 위치(Grey Sheep)를 가지고 있는 사용자들을 고려하여 추천을 수행하고 이를 위해 사용자 간 유사성이 아닌 사용자의 특이한 취향이나 관심사를 고려하는 알고리즘을 사용

50. 제프리 힌턴이 장시간에 걸쳐 찾아낸 기존 인공신경망 접근법의 4 가지 문제점은 무엇이며, 이것이 딥러닝 출현에 앞서 어떻게 해결되었는지 설명하시오.

제프리 힌턴이 20 년간 찾아낸 기존 인공신경망 접근법의 4 가지 문제점

1. 데이터 부족의 문제 :우리가 사용한 학습 데이터셋이 약 수천배 작은 규모였던 것 같다.
2. 컴퓨팅 파워의 문제: 우리가 사용한 컴퓨터가 수백만배 느렸던 것 같다.
3. 초기 가중치 설정 문제: 우리가 학습을 위한 초기 가중치를 너무 대충 설정했던 것 같다.
4. 알고리즘의 문제: 우리가 잘못된 활성화 함수(Sigmoid)를 사용해 왔던 것 같다.

딥러닝 출현에 앞서 기존 인공신경망 접근법의 4 가지 문제점에 대한 해결방안

1. 데이터 부족의 문제: 모델의 크기 대비 학습 데이터의 부족은 필연적으로 과적합(overfitting) 문제를 야기하는데, Big Data 가 해결하였다.
2. 컴퓨팅 파워의 문제: 고성능 그래픽의 게임이 해결하였다.(GPU 사용으로 획기적인 학습시간 개선. 벡터, 행렬 연산에 특화)

3 & 4. 초기 가중치 설정 문제, 알고리즘의 문제:

딥러닝 모델 학습에 있어서 가중치에 설정에 따라 기울기 소실 문제(Vanishing Gradient)이 발생하기 때문에 초기 가중치 설정은 중요하다. 특히, Sigmoid 활성화 함수의 경우, 미분을 거듭할수록 기울기 소실 문제가 생겼는데, 이를 새로운 활성화 함수(ReLu)를 적용하여 해결하였다.

추가적으로 가중치 초기화로 사용하는 Default Option 으로 Random Initialization 을 사용할 경우, Back Propagation 과정에서 Expoloding Gradient 혹은 Vanashinig Gradient 문제에 빠질 위험성이 큰데, Xavier Initialization, He Initialization 등의 방법으로 활성화 함수의 특성을 고려하여 가중치를 초기화 하는

초기 가중치 설정 문제를 해결하였다.

- **3&4 번 中 Sigmoid 활성화 함수 사용시, 기울기 소실 문제가 생기는 이유와 해결 방안

$$sigmoid(x) = \frac{1}{1+e^{-x}}$$

(Sigmoid 함수 정의)

$$\frac{d}{dx}sigmoid(x) = \frac{1}{1+e^{-x}}\left(1-\frac{1}{1+e^{-x}}\right)$$

(Sigmoid 함수 미분 과정)

딥러닝의 Back Propagation 에서 Chain rule(미분 연쇄 법칙) 적용시, 미분 결과값이 소실되어 버리는 문제 발생(결과 값이 0 으로 수렴)

$$W \leftarrow W - \eta\frac{\partial L}{\partial W}$$

미분을 거듭할수록 $-\eta\frac{\partial L}{\partial W}$에 해당하는 Gradient 항(가중치 업데이트 수식)이 사라져 버리면서 기울기 소실 문제가 발생한다. ($\frac{\partial L}{\partial W}$ 가 0 에 가까워 지는 문제)

새로운 활성화 함수(ReLu)를 적용을 통한 Vanishing Gradient Problem 보완 할 수 있다.

$$f(x) = max(0, x)$$

0 보다 작은 값은 0 으로 반환하고, 0 보다 큰 값이 나올 경우, 그대로 반환한다. ReLu 함수의 경우, $max(0, x)$에서 0 으로 결과값이 도출된다면 이후 출력값이 모두 0 이 되므로, 이후에 학습이 이루어 지지 않는다는 한계가 있는데, 이를 보완한 LeakyReLU 로 0 대신 아주 작은 값을 곱해주어 문제점을 보완할 수 있다.

$$f(x) = max(0.01x, x)$$

딥러닝 출현에 앞서, 알고리즘 문제를 해결하면서, ReLU 와 같은 활성화 함수는 Resnet, YOLO 모델 등에 활용되고 있다.