정형, 반정형, 비정형 데이터의 이해

1. 정형 데이터(Structured data)

- 의미를 파악하기 쉽고, 규칙적인 데이터 / 정형화된 스키마 구조, DBMS에 내용이 저장될 수 있는 구조 (Gender라는 컬럼 - male, female)

2. 비정형 데이터(Unstructured data)

- 정해진 규칙이 없어서 값의 의미를 파악하기 어려운 데이터 / 수집 데이터 각각이 데이터 객체로 구분, 고정 필드 및 메타데이터가 정의되지 않음
- (텍스트, 음성 ,영상 등)
- 3V : Velocity(속도), Volume(양), Variety(다양) => 비정형 데이터는 Variety에 속한다.

3. 반정형 데이터(Semi-structured data)

- 데이터 내부의 데이터 구조에 대한 메타 정보가 포함된 구조 (HTML, XML같은 포맷)

데이터 전처리

1. 평활화 : 데이터로부터 잡음을 제거하기 위해 데이터 추세에 벗어나는 값들을 변환하는 기법
2. 집계 : 다양한 차원의 방법으로 데이터를 요약하는 기법
3. 정규화 : 데이터를 정해진 구간 내에 들도록 하는 기법
4. 일반화 : 특정 구간에 분포하는 값으로 스케일을 변화시키는 기법

척도(Scale)의 종류

1. 명목 척도 : 단순히 집단의 분류를 목적으로 사용된 척도
2. 서열 척도 : 측정 대상 사이의 대소 관계를 나타내기 위한 척도
3. 구간척도(등간 척도) : 서열과 의미 있는 차이를 가지는 척도
4. 비율척도 : 구간척도의 성질을 가지며 척도 간의 비율(Ratio)도 의미가 있는 척도

R 데이터 유형 이해

1. 벡터

- R에서 데이터를 저장하는 기본적인 단위로 동일한 데이터 타입의 여러 값을 저장하는 객체
- 1차원 배열 형태로 특정 항목의 요소를 사용하려면 벡터명[색인]을 사용한다.
- 각 벡터의 요소에 names()함수를 사용해서 이름을 지정할 수 있다.

2. matrix

- 2차원 데이터 구조를 나타내는 자료형
- 행(row)과 열(column)로 구성되며, 모든 원소는 동일한 데이터 타입을 가진다.

3. list :

- 여러가지 데이터 유형을 포함할 수 있는 유연한 자료구조
- 서로 다른 데이터 타입의 원소들을 하나의 객체로 그룹화 하여 저장하는데 사용된다.

4. data.frame

- 테이블 형태의 데이터 구조를 나타내는 자로형
- 서로 다른 데이터 타입의 열(column)로 구성된 데이터를 저장하며, 각 열은 동일한 길이를 가져야 한다.

시각화 그래프 종류

관찰점들을 표시하는 시각화 방법? => 스캐터(Scatter) 차트 연속형 자료에 대한 도수분포표를 시각화한 그래프? => 히스토그램 범주별 빈도를 요약해서 나타낸 시각화한 그래프? => 막대(bar)그래프

R 언어의 특성

- 데이터 조작, 계산 및 시각화를 지원하는 데이터 분석 환경 제공
- 벨 연구소에서 만들어진 통계 분석 언어인 S에 근간을 두고있다.
- GPL(General Public License)로 배포되는 오픈소스 소프트웨어
- Cross-platform
- Interpreter Language
- 통계, 기계, 학습, 금융, 생물정보학, 그래픽스에 이르는 다양한 12,000개 이상의 통계 분석 관련 패키지
- 데이터 구조 연산을 수행할 수 있다.
- 파이썬에 비해서 강력한 점은 시각화

회귀분석의 다양한 유형

01. 선형 회귀(Linear Regression) 모델의 이해

- 머신러닝의 가장 큰 목적은 실제 데이터를 바탕으로 모델을 생성해서 만약 다른 값을 넣었을 때 발생할 아웃풋을 예측하는 데 있다.
- 가장 직관적이고 간단한 모델은 선(line)이다. 데이터를 가장 잘 설명하는 선을 찾는 분석 방법을 선형 회귀(Linear Regression)분석이라 한다. (최적적합선) (키값 기반으로 몸무게 예측, 공부한 시간 기반으로 점수 예측, 기값 기반으로 발사이즈 예측 등)
- 직선 그래프를 그리는 1차 함수로 표현 y = mx + b
- 선형 회구는 데이터를 가장 잘 설명하는 기울기 m, 절편 b를 찾는 것
- 일반 선형 회귀, 릿지(Ridge),

02. 로지스틱 회귀(Logistic Regression)

- 0 또는 1, 참 또는 거짓, 흑 또는 백, 스팸 또는 스팸 아닌 것 등의 두 가지 값 중 하나만 취할 수 있는 경우 로지스틱 회귀를 사용하여 데이터를 분석할 수 있다.

03. 리지(Ridge Regression)

- 리지 회귀는 정규화 또는 규제화(regularization) 기법으로 알려져 있으며 모델의 복잡성을 줄이는 데 사용된다. 또한 '리지 회귀 페널티'로 알려진 약간의 편향, 즉 바이어스(bias)를 사용하여 모델이 과대적합(overfitting)에 덜 취약하게 만든다.

04. 다항 회귀(Polynomial regression)

- 다항 회귀는 선형 모델을 사용하여 비선형 데이터 집합을 모델링한다.
- 다항 회귀는 독립 변수가 여러 개인 선형 회귀를 뜻하는 다중 선형 회귀와 비슷한 방식으로 작동하지만, 비선형 곡선을 사용한다.

부트 스트래핑을 사용한 앙상블 학습 알고리즘 => 배깅(Bagging)

- 앙상블 학습의 유형 : 보팅(Voting), 배깅(Bagging), 부스팅(Boosting), 스태킹(Stacking)

1. 보팅(Voting)

- 여러 종류의 알고리즘을 사용한 각각의 결과에 대해 투표를 통해 최종 결과를 예측하는 방식

2. 배깅(Bagging) :

- 개별 분류기가 부트 스트래핑 방식으로 샘플링된 데이터 세트에 대해서 학습을 통해 개별적인 예측을 수행한 결과를 보팅을 통해서 최종 예측 결과를 선정하는 방식
- 같은 알고리즘에 대해 데이터 샘플을 다르게 두고 학습을 수행해 보팅을 수행하는 방식

3. 부스팅(Boosting)

- 여러 개의 알고리즘이 순차적으로 학습을 하되, 앞에 학습한 알고리즘 예측이 틀린 데이터에 대해 올바르게 예측할 수 있도록, 그 다음번 알고리즘에 가중치를 부여하여 학습과 예측을 진행하는 방식

4. 스태킹(Stacking)

- 여러 가지 다른 모델의 예측 결과값을 다시 학습 데이터로 만들어 다른 모델(메타 모델)로 재 학습시켜 결과를 예측하는 방법

은닉층(Hidden Layer)에서 가중치 조정이 이루어지지 않아 신경망의 학습이 제대로 이루어지지 않는 현상은? => 그래디언트 소실(vanishing gradient)

- 그래디언트 소실은 역전파 알고리즘으로 심층 신경망을 학습시키는 과정에서, 출력층에서 멀어질수록 신경망의 출력 오차가 반영되지 않는 현상을 말한다.
- 오차가 앞쪽의 레이어까지 전달이 안돼 가중치가 변화가 되지 않는다. 즉 학습되지 않는 현상.

딥러닝 신경망 모델의 특징 이해

1. CNN(Convolutional Neural Networks) ConvNet

- 입력 데이터에 대해 2D 컨벌루션 레이어를 사용으로써 특징을 추출, 이미지와 같은 2차원 데이터 처리에 적합한 아키텍처

2. 순환신경망(RNN. Recurrent Neural Network)

- 순차적 정보가 담긴 데이터에서 규칙적인 패턴을 인식하고, 추상화 된 정보를 추출
- 텍스트, 음성, 음악, 영상 등 순차적 데이터를 다루는 데 적합

3. GAN(Generative Adversarial Network. 생성 대립 신경망)

- 비지도 학습 방법, 훈련으로 학습된 패턴을 이용해 이미지나 음성을 생성할 수 있다.

- GAN은 이미지 및 음성 복원 등에 적용된다.

퍼셉트론(Perceptron)

- 다수의 입력으로부터 하나의 결과를 내보내는 알고리즘 각각의 입력값에 가중치 곱의 힘을 신경망에 보내고 활성화함수(임계치)를 사용하여 임계치를 넘으면 출력 신호로 1을 출력하고. 임계치를 넘지 못하면 0을 출력한다.

다층 퍼셉트론(MLP)에서 기울기 소실의 원인?

- 은닉층을 많이 거칠수록 전달되는 오차가 크게 줄어들어 학습이 되지 않는 현상이 발생하는데, 이를 기울기 소멸 문제라고 한다.
- 기울기 소실 문제 해결을 위한 활성화 함수 : relu, thnh, leekyrelu, sigmoid

출력층에서 사용하는 활성화 함수 종류와 특성 이해

활성화 함수(Activation function)란?

- 출력값을 활성화를 일으키게 할 것인가를 정하고 그 값을 부여하는 함수
- 활서오하 함수를 사용하면 입력값에 대한 출력값이 linear하게 나오지 않으므로 linear system을 non-linear한 systemd으로 바꿀 수 있다.

활성화 함수 종류

1. sigmoid

- 입력값 x 값이 작아질수록 0에 수렴하고, 커질수록 1에 수렴한다. (Logistic 함수, S자형 함수)

2. softmax

- 출력층에서 주로 사용한다.
- 세 가지 이상의(상호 배타적인) 선택지 중 하나를 고르는 다중 클래스 분류 문제에 주로 사용한다.

3. Stop Function

- 출력값이 0이 될지, 1이 될지를 결정
- 계단 모양 함수로, 특정값 이하는 0, 특정값 이상은 1로 출력하도록 만들어진 함수

4. ReLU

- 음수를 입력하면 0을 출력하고, 양수를 입력하면 입력값을 그대로 반환한다.
- x가 0보다 크면 기울기가 1인 직선, 0보다 작으면 함수 값이 0이 된다.

5. Leaky ReLU
6. Hyperbolic tangent function

서포트 벡터 머신 분류 모델 특성 이해

SVM(Support Vector Machine)

- 두 그룹에서 각각의 데이터 간 거리를 측정하여 두 개의 데이터 사이의 중심을 구한 후, 그 가운데서 최적의 초평면을 구함으로써 흰색과 검은색 그룹을 나누는 방법을 학습한다.
- 신경망 포함한 분류 모델은 '오류율을 최소화'하려는 목적으로 설계되었다
- SVM은 두 부류(class) 사이에 존재하는 '여백을 최대화'하려는 목적으로 설계되었다
- SVM의 margin - 분류를 위한 경계선과 이 경계선에 가장 가까운 트레이닝 데이터 사이의 거리를 의미한다.
- 직선 : 선형 분류 모델 적용
- 직선 X : 비선형 분류 모델 적용

군집 분석의 특성 이해

- 동일한 성격을 가진 여러 개의 그룹으로 대상을 분류하는 것
- 군집화 혹은 군집분석이라고 부른다.
- 대상 개체를 유사하거나 서로 관련있는 항목끼리 묶어서 몇 개의 집단으로 그룹화하거나, 각 집단의 성격을 파악함으로써 데이터 전체 구조에 대한 이해를 돕는 탐색적 분석 방법이다.

군집분석 특징

- 종속변수에 대한 독립변수의 영향과 같이 사전에 정의된 특수한 목적이 없으며, 데이터 자체에 의존해서 데이터 구조와 자료를 탐색하고 요약하는 기법이다.
- 동일한 군집 내 개체들을 유사한 성격을 가진다. <-> 서로 다른 군집은 이질적인 성격을 가지도록 군집이 형성된다.

군집 유형

- 상호 배반적 군집 : 각 관찰치가 군집 중 오직 하나에 속한다.
- 계보적 군집 : 한 군집이 다른 군집에 포함, 상하종속 관계
- 중복 군집 : 두 개 이상의 군집에 한 괄찰치가 소속
- 퍼지(Fuzzy) : 각 군집에 속할 확률을 표현하는 방법

이미지 분류 분석에서 신경망을 이용한 이미지의 특성 추출하는 단계? => Convolution

- CNN (합성곱신경망 : Convolution Neural Network)
- Convolution : 데이터의 특징을 추출하는 과정으로 데이터에 각 성분의 인접 성분들을 조사해 특징을 파악하고 파악한 특징을 한장으로 도출시키는 과정

하이퍼파라미터란?

- 최적의 훈련 모델을 구현하기 위해 모델에서 설정하는 변수로 학습률(Learning & Late), 에포크 수(훈련 반복 횟수), 가중치 초기화 등을 결정 할 수 있다.
- 하이퍼파라미터 튜닝 기법으로 훈련 모델의 최적값을 찾을 수 있다.

하이퍼파라미터 특징

- 모델의 매개 변수를 추징하는 데 도움이 되는 프로세스에 사용된다.

- 개발자에 의해 수동으로 설정할 수 있다. (임의 조정 가능)
- 학습 알고리즘의 샘플에 대한 일반화를 위해 조절된다.

하이퍼파라미터 튜닝기법

- 그리드 탐색
- 랜덤 탐색
- 베이지안 최적화
- 휴리스틱 탐색

[정리]

- 모델 파라미터는 새로운 샘플이 주어지면 무엇을 예측할지 결정하기 위해 사용하는 것이며, 학습 모델에 의해 결정된다.
- 하이퍼파라미터는 학습 알고리즘 자체의 파라미터로 모델이 새로운 샘플에 잘 일반화 되도록 최적값을 찾으나, 데이터 분석 결과로 얻어지는 값이 아니므로 절대적인 최적값은 존재하지 않고, 사용자가 직접 설정해애 한다.

최적화 알고리즘(Optimization Algorithm) 함수별 특성 이해

1. Gradient Descent Algorithm (경사하강법)

- 네트워크의 parameter들을 θ (->W,b)라고 했을 때, Loss function J(θ)의 optima(Loss funtion의 최소화)를 찾기 위해 파라미터의 기울기(gradient) ∇θJ(θ) (즉, dW와 db)를 이용하는 방법

2. Momentum Algorithm (모멘텀)

- Momentum 알고리즘을 이해하기 위해서는 지수 가중 평균(Exponentially Weighted Average)(=지수 이동 평균)을 이해해야 한다. 지수 가중 평균(=지수 이동 평균) : 데이터의 이동 평균을 구할 때, 오래된 데이터가 미치는 영향을 지수적으로 감쇠(exponential decay) 하도록 만들어 주는 방법.

3. RMSprop(Root Mean Square) Algorithm

- 기울기 강하의 속도를 증가시키는 알고리즘

4. Adam(Adaptive Moment Estimation) Optimization Algorithm

- 모멘텀과 RMSprop을 섞어놓은 최적화 알고리즘 입기 때문에, 딥러닝에서 가장 흔히 사용되는 최적화 알고리즘

5. Opimization problem solving

- Learning rate decay를 사용하면 처음에는 learning rate α를 하여 빠르게 학습을 진행하고, α값이 점차 작아지면서 optima 부근의 매우 좁은 범위까지 집입할 수 있게 된다.

RNN 신경망 이해와 특징?

- 순차적인 데이터
- 시계열 데이터 분석에 사용되는 신경망 모델
- 은닉층의 출력을 다음 은닉층의 입력으로 보내고, 출력층으로 보낸다. (누적)
- 은닉층의 노드를 'cell'이라고 하며, 기억 기능을 가진다.

혼동 행렬(Confusion Matrix)을 통한 모형의 평가지표에서 계산할 수 있는 정확률, 재현율, 정밀도 공식

혼동행렬(confusion matrix)

- 모델의 성능을 평가할 때 사용되는 지표
- 예측값이 실제 관측값을 얼마나 정확히 예측했는지 보여주는 행렬

1. 정확도(accuracy)

- 모델이 입력된 데이터에 대해 얼마나 정확하게 예측하는지를 나타낸다.
  (공식) 정확도 = 예측값결과와 실제값이 동일한 건수 / 전체 데이터수 = (TP/TN) / (TP + TN + FN + FP)

2. 정밀도(precision)

- 모델의 예측값이 얼마나 정확하게 예측됐는가를 나타내는 지표
- "예" 라고 했을 때의 정답률
  (공식) 정밀도 = TP / (TP + FP)

3. 재현율(recall)

- 실제값 중에서 모델이 검춤할 실제값의 비율을 나타내는 지표
- 실제로 병이 있는 전체 중 참 긍정의 비율
  (공식) 재현율 = TP / (TP + FN)

모델 학습시에 과적합 방지를 위해 적용할 수 있는 방법들?

과적합(Overfitting)

: 모델이 학습 데이터에만 과도하게 최적화 되어, 실제 예측은 다른 데이터로 수행할 경우에는 예측 성능이 떨어지는 것

과소적합(Underfitting)

: 머신러닝 모델이 학습 데이터에 대해 충분히 복잡한 모델을 학습하지 못하여 예측 성능이 낮은 상태를 의미한다. 즉 모델이 학습 데이터에 대한 패턴을 제대로 학습하지 못하고 단순한 모델이나 너무 제한적인 모델을 만드는 경우에 발생할 수 있다.

Overfitting 해결 방법

- 더 많은 데이터 수집
- 데이터 확장(Data Augmentation)
- 모델 복잡도 감소
- 가중치 규제(Ragularization)
- 교차 검증(Cross Vaildation)
- 조기 종료(Early Stopping)
- 앙상블(Ensemble)
- 드롭아웃(Dropout) : 학습 과정에서 신경망 일부 사용 X