1 회귀 분석

1 회귀 분석 개념

- 회귀 분석은 지도 학습의 하나로 학습 데이터에 부합되는 출력 값이 실수인 함수를 찾는 문제

예

- 나이와 키의 상관 관계를 통해 임의의 나이를 입력했을 때의 키 값 구하기
- 중고차의 주행 거리와 가격과의 관계를 통해 임의의 주행 거리 입력 시 중고차 가격 구하기

2 초기 가설

- 회귀 분석은 학습 데이터에 따른 결과값이 주어지면 임의의 데이터를 입력했을 때 실수인 결과값을 찾을 수 있는 함수를 구하는 문제
- 회귀 분석의 가설은 학습 데이터를 이용한 학습 전 함수를 가정하는 것

예

- 나이와 몸무게 추정 데이터
- 16명의 데이터를 수집하여 학습

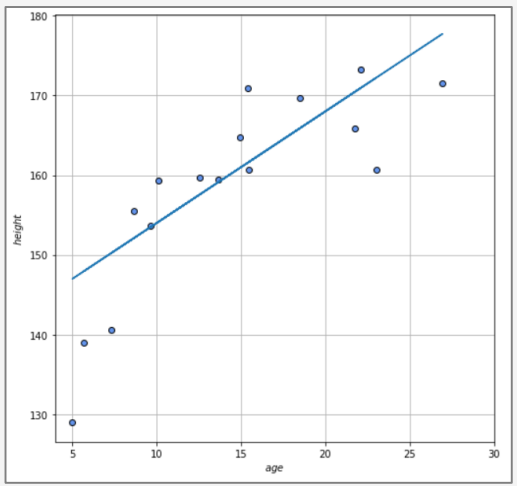
나이	15.43	23.01	5	12.56	8.67	7.31	9.66	13.64	14.92	18.47	15.48	22.13	10.11	26.95	5.68	21.76
키	170.91	160.68	129	159.7	155.46	140.56	153.65	159.43	164.7	169.65	160.71	173.29	159.31	171.52	138.96	165.87

2 초기 가설

- 초기 가설 함수는 1차 함수로 선택
- 주어진 학습 데이터를 통해 가장 쉽게 추정할 수 있는 방법
- 단순하므로 정확한 값 산출은 어려움
- 가설
  - $h(x) = w_0 x + w_1$
  - 우리가 구하고자 하는 것은 w_0, w_1인 파라미터
  - w_0: 기울기, w_1: 절편
- 기울기 w_0와 절편 w_1의 값을 이용하면 다양한 위치와 기울기의 직선을 생성
- $h(x)$는 x에 대한 예측치

2 초기 가설

- 임의의 기울기와 절편으로 구성한 1차 가설 함수
- 기울기와 절편을 어떻게 설정해야 가설 함수가 데이터와 부합할 것인지를 학습



[데이터와 가설 함수]

3 오차 함수

- 가설로 설립한 함수와 최종 학습된 함수의 차이를 데이터를 학습하면서 줄이도록 하는 것이 학습 목표
- 오차 함수란 주어진 임의의 데이터 입력값으로 표현되는 실제 출력값과 가설 함수를 통해 출력된 시뮬레이션 출력값과의 차이를 계산하는 함수
- 최종 학습 목표는 모든 데이터에 대해 실제 출력값과 시뮬레이션으로 출력되는 출력값과의 오차가 최소가 되도록 하는 것

3 오차 함수

- 제곱 오차 함수(Sum of Square Error; SSE) 계열 사용

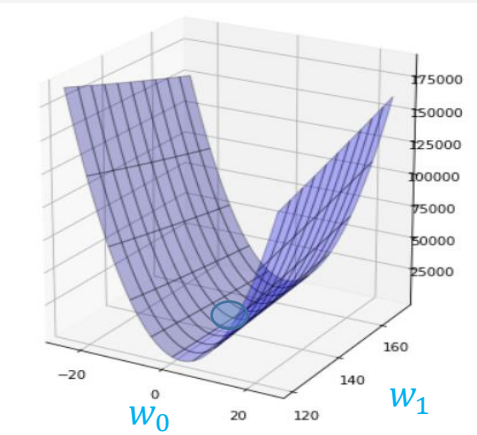
$$E = \frac{1}{N}\sum_{n=0}^{N-1}(h_n - t_n)^2$$

평균 제곱 오차 함수(Mean Square Error)

- 제곱 오차의 경우 N(데이터 개수)에 영향을 받을 수 있으므로 평균 제곱 오차(Mean Square Error; MSE)를 사용
- w_0와 w_1을 결정하면 평균 제곱 오차 E 계산이 가능
- 데이터가 직선 상에 나란히 있지 않기 때문에 E가 완전히 0이 되지는 않음

3 오차 함수

- 1차 선형 회귀는 평균 제곱 오차를 사용할 경우 2차 함수로 오차 함수를 표현할 수 있으며 이 때 오차 함수의 최소값을 구하면 그 때의 파라미터가 최종 학습 목표인 회귀 함수의 파라미터로 결정



[평균 제곱 오차 함수 시각화]

4 경사 하강법(Gradient Descent Method)

- 평균 제곱 오차 함수가 최소가 되는 위치를 근사화 하기 위한 학습 방법
- E가 가장 작아지는 w_0과 w_1를 구하는 방법 중 하나
- 초기 w_0과 w_1를 결정하고 이 점에서의 기울기를 확인하여 E가 감소하는 방향으로 w_0과 w_1를 조금만 이동
- 이 절차를 여러 번 반복하여 최종적으로 E가 가장 작아지는 점인 w_0과 w_1에 도착

4 경사 하강법(Gradient Descent Method)

- At (w_0, w_1), E를 w_0과 w_1로 편미분하여 벡터로 표현:
  $\boldsymbol{E}$ 의 기울기

$$\nabla \boldsymbol{w} E = \left[\frac{\partial E}{\partial w_0} \quad \frac{\partial E}{\partial w_1} \right]$$

- E를 최소화하기 위해서는 E의 기울기의 반대 방향인
  - $\nabla \boldsymbol{w} E$ 로 진행

4 경사 하강법(Gradient Descent Method)

- 각 w_0과 w_1에 대해 편미분

$$E = \frac{1}{N}\sum_{n=0}^{N-1}(h_n - t_n)^2 = \frac{1}{N}\sum_{n=0}^{N-1}(w_0 x_n + w_1 - t_n)^2$$

  - w_0에 대해 편미분

$$\frac{\partial E}{\partial w_0} = \frac{2}{N}\sum_{n=0}^{N-1}(w_0 x_n + w_1 - t_n)x_n = \frac{2}{N}\sum_{n=0}^{N-1}(h_n - t_n)x_n$$

  - w_1에 대해 편미분

$$\frac{\partial E}{\partial w_1} = \frac{2}{N}\sum_{\gamma=0}^{N-1}(w_0 x_n + w_1 - t_n) = \frac{2}{N}\sum_{n=0}^{N-1}(h_n - t_n)$$

4 경사 하강법(Gradient Descent Method)

- 각 w_0과 w_1에 대해 편미분

$$w_0(t+1) = w_0(t) - \alpha \frac{2}{N} \sum_{n=0}^{N-1} (h_n - t_n) x_n$$

$$w_1(t+1) = w_1(t) - \alpha \frac{2}{N} \sum_{n=0}^{N-1} (h_n - t_n)$$

4 경사 하강법(Gradient Descent Method)

- 학습 규칙을 이용한 매개 변수 산출
  - 초기 값 w_0=10, w_1=165로 가정할 경우 기울기 값은 각각

    $dw_0 = 5046.3, dw_1 = 301.8$

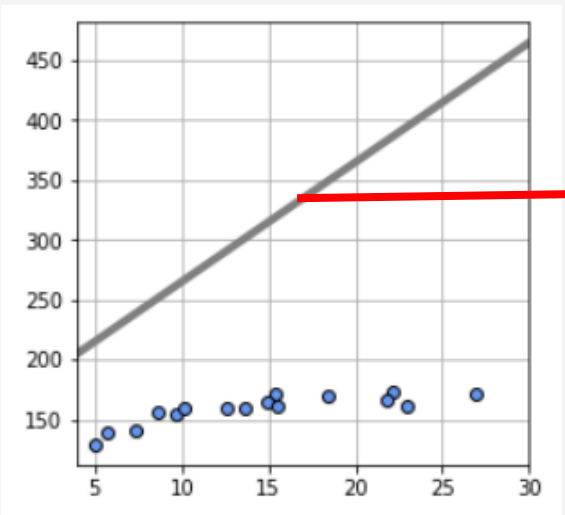
  - 초기 매개 변수 값을 이용하여 각 매개 변수 범위 내에서 학습을 수행

    w_0=10, w_1=165, α=0.001, threshold=0.1

4 경사 하강법(Gradient Descent Method)

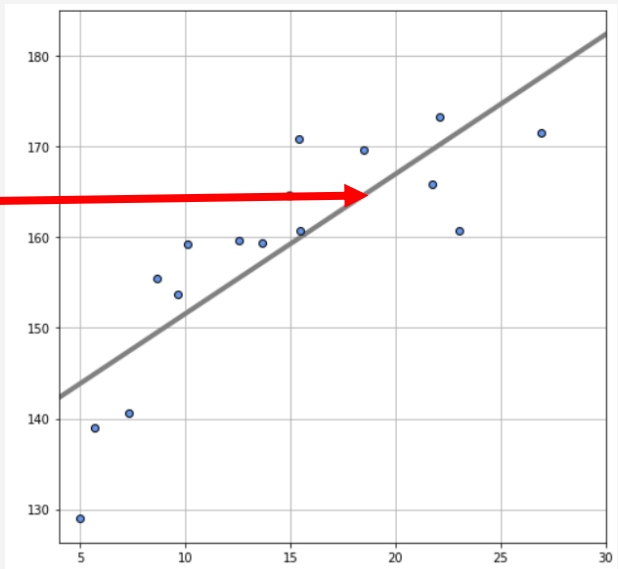
[초기 가설]

$w_0 = 10,\ w_1 = 165$



[학습 후]

$w_0 = 1.540,\ w_1 = 136.17,$ SD: 7.02cm



4 경사 하강법(Gradient Descent Method)

- 그래프의 해석
  - 평균 제곱 오차(MSE): 49.03 cm^2
  - 표준 편차(Standard Deviation): 약 7.0 cm

  → 오차가 정규 분포를 따른다는 가정하에,
  전체 68%의 데이터 점에서 오차가 7.0 이하라는 의미
- 일반적인 경사 하강법
  - 산출된 해는 부분적인 극소값(Local Minima)
  - E가 복잡한 다차원일 경우 가장 최소가 되는 극소값(Global Minima)를 구하는 것은 어려운 문제

  → 다양한 초기값에서 경사 하강법을 시도하여
  E가 작아진 지점을 최소값으로 선택

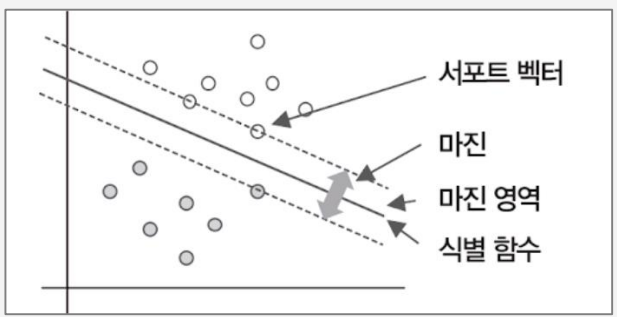
2 서포트 벡터머신

1 서포트 벡터 머신(Support Vector Machine; SVM)

- 데이터 분포를 나누는 기준을 결정하는 지도 학습 모델 중 하나
- 회귀 분석의 관점
  - 데이터에 맞춘 직선과 곡선의 특징을 분석
- SVM의 관점
  - 어떤 패턴으로 데이터를 분류한 후 데이터 사이의 거리에 따라 어떤 카테고리에 속할 것인지 판단
  - 다층 퍼셉트론 같은 신경망을 이용한 데이터 분류와 유사

1 서포트 벡터 머신(Support Vector Machine; SVM)

- 2개의 클래스에 속한 데이터 사이의 거리를 최대화 하면서 가운데를 통과하는 식별 함수를 구함
- 서포트 벡터
  - 마진 영역의 가장 자리에 해당하는 위치에 있는 데이터
- 마진 영역
  - 서포트 벡터와 식별 함수 사이의 공간



[SVM 예]

1 서포트 벡터 머신(Support Vector Machine; SVM)

- 커널 트릭 방법을 이용하여 비선형 식별 함수를 결정할 수도 있음
- 선형 식별 함수
  - 마진 최대화 시 선형 식별 함수는 모든 훈련 데이터를 올바르게 식별할 수 있어야 함
  - 훈련 데이터와 식별 함수의 값이 0이 되는 식별 초평면과의 최소 거리가 가장 크도록 최적화
  - 라그랑주 승수법으로 해결
  - 식별 함수는 서포트 벡터만 결정할 수 있음

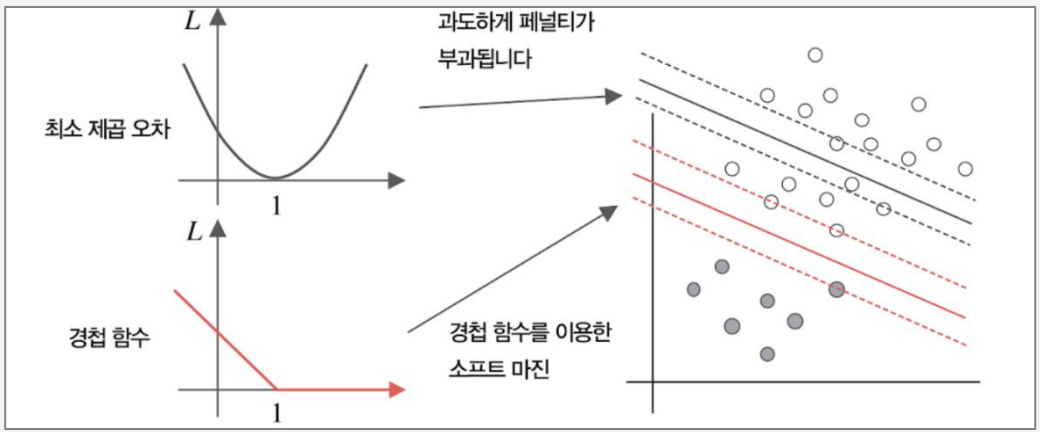
1 서포트 벡터 머신(Support Vector Machine; SVM)

- 식별 함수
  - 입력 데이터의 내적만 이용하는 형태
  - 내적은 비선형 식별 함수를 만드는 방법을 찾는 데도 이용
- 커널 트릭
  - 차원을 높이지 않고도 차원을 올리는 효과
  - 다항식과 가우스 함수 등의 커널 함수를 이용하여 원래 공간의 데이터 분포를 선형 분리할 수 있는 공간으로 변형하는 것
  - 주성분 분석과 클러스터 분석 등에 응용

1 서포트 벡터 머신(Support Vector Machine; SVM)

- 소프트 마진
  - 실제로 데이터를 깔끔하게 분리가 되지 않으므로 잘못 식별된 데이터에 페널티를 설정하여 오차에 대응
  - 마진 최대화와 페널티를 함께 고려해 최적화 하는 방법
- 손실 함수는 경칩(hinge) 모양을 나타내므로 경첩 함수 또는 경첩 손실 함수라고도 함

1 서포트 벡터 머신(Support Vector Machine; SVM)



[경첩함수]

3 의사결정 트리

3 | 의사결정 트리

[결정트리 종류]

알고리즘	분리기준(목표 변수)	비고
ID3	Entropy	다지 분리(범주형)
C4.5	Information gain	다지 분리(범주형), 2진 분리(연속형)
C5.0	Information gain	C4.5와 유사
CHAID	카이제곱, F검정(연속)	통계적 접근 방식
CART	Gini index(범주형), 분산의 차이(연속형)	통계적 접근 방식 항상 2진 분리

3 | 의사결정 트리

1 ID3

- 정답 데이터를 이용해 결정 트리를 만드는 알고리즘의 하나
- 의사 결정 트리를 기반으로 모든 데이터를 제대로 분류할 때까지 노드 추가
- 데이터를 제대로 분류하는 결정 트리는 여러 개가 나올 수 있음
- 분류 효율성과 의사 결정 트리의 일반성을 고려해 최대한 단순한 형태가 되는 것이 목표

1 ID3

- ID3 장점
  - 계산 비용이 적음
  - 학습된 결과를 사람이 이해하기 쉽고, 누락값에 대한 처리 가능
  - 분류와 관련 없는 속성에 대해서도 처리 가능
- ID3 단점
  - 연속형 수치 사용이 불가
  - 속성의 값이 많을 경우 가지의 수도 많아짐(overfitting)

2 ID3에서 결정 트리 만드는 방법

- 모든 데이터를 포함한 하나의 노드로 구성된 트리에서 시작
- 반복적인 노드 분할 과정

[1]	분할 속성(splitting attribute)을 선택
[2]	속성값에 따라 서브 트리(subtree)를 생성
[3]	데이터를 속성값에 따라 분배

2 ID3에서 결정 트리 만드는 방법

분할 속성 결정

- 어떤 속성을 선택하는 것이 효율적인가?
  - 가능하면 분할 결과가 동질적인 것으로 만드는 속성
- Entropy(엔트로피)
  - 동질적인 정도 측정 가능 척도
  - 원 의미는 정보량(Amount of information) 측정 목적의 척도

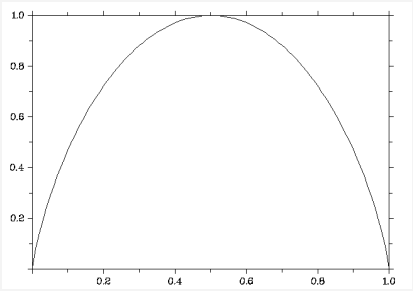
2 ID3에서 결정 트리 만드는 방법

Entropy

- $p(c)$: 부류 c에 속하는 것의 비율

$$I = -\sum_{c} p(c) \log_2 p(c)$$

- 2개의 부류가 존재할 경우



◁ 가장 모호성이 큰 경우

2 ID3에서 결정 트리 만드는 방법

Information Gain

- 원래의 엔트로피와 세부 클래스로 분할된 후의 엔트로피의 차이
- information Gain이 클수록 분류하기 좋다는 것을 의미
- ID3알고리즘은 Information Gain를 사용하여 노드를 결정
- $IG = I - I_{res}$
- I_{res}: 특정 속성으로 분할한 후의 각 부분 집합의 정보량의 가중 평균

$$I_{res} = -\sum_{v} p(v) \sum_{c} p(c|v) \log_2 p(c|v)$$

$$IG = I - I_{res}(A) = -\sum_{c} p(c) \log_2 p(c) + \sum_{c} p(v) \sum_{c} p(c|v) \log_2 p(c|v)$$

2 ID3에서 결정 트리 만드는 방법

Information Gain

- 단점
  - 속성값이 많은 것을 선호하여 선택하게 됨
  - 예) 학번 또는 이름 등
  - 속성값이 많으면 데이터 집합을 많은 부분집합으로 분할
  - 작은 부분 집합은 동질적인 성향으로 분석
- 개선
  - 정보 이득비(Information Gain Ratio)
  - 지니 지수(Gini Index)

2 ID3에서 결정 트리 만드는 방법

Information Gain Ratio

- 속성값이 많은 속성에 대해서 페널티를 부여
- $I(A)$
  - 속성 A의 속성 값을 클래스로 간주하여 계산한 엔트로피
  - 속성 값이 많을 수록 커지는 경향

$$GainRatio(A) = \frac{IG(A)}{I(A)} = \frac{I - I_{res}(A)}{I(A)}$$

2 ID3에서 결정 트리 만드는 방법

지니 지수

- 데이터의 균형과 불균형에 대한 지수

$$Gini = \sum_{i \neq j} p(i)p(j)$$

- 속성 A에 대한 지니 지수의 가중 평균 값

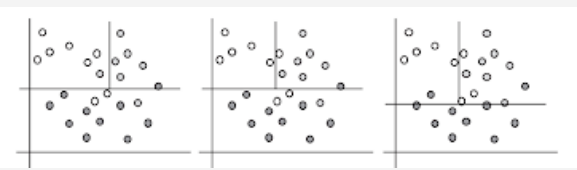
$$Gini(A) = \sum_{v} p(v) \sum_{i \neq j} p(i|v)p(j|v)$$

- 지니 지수 이득(gini index gain)

$$GiniGain(A) = Gini - Gini(A)$$

3 랜덤 포레스트

- 서포트 벡터 머신과 함께 데이터의 분포를 분류하는 방법
- 무작위로 뽑은 데이터를 이용해 학습하면서 많은 결정 트리를 구축하며 의사 결정 트리를 만들 때 마다 결정 트리 구성을 약간씩 변화
- 최종 단계에서 구축한 결정 트리 중 최적의 결정 트리를 선택
- 타당성 검증
  - 모델을 만들 때는 모델이 얼마나 정확한 결과를 계산하는지를 객관적으로 측정



… 결정 트리 다수를 만듭니다 → 적합한 최적의 모델을 대표로 선정합니다

분류 트리 ⇒ 다수결
회귀 트리 ⇒ 평균값

4 식별 모델의 평가와 ROC 곡선

- ROC 곡선(Receiver Operating Characteristic curve)
  - 식별 모델의 성능을 평가하는 방법
  - 제 2차 세계 대전 때 수신된 레이더 신호에서 적 전투기를 찾으려는 미국의 레이더 연구에서 탄생한 개념
- ROC 곡선의 생성
  - 데이터의 정답 결과 세트와 식별 결과 세트를 준비해 혼동 행렬을 생성
  - 식별 결과가 두 종류인 경우 혼동 행렬은 2x2 표 형태가 됨

혼동 행렬은 참 긍정(True Positive; TP), 거짓 부정(False Negative; FN), 거짓 긍정(False Positive; FP), 참 부정(True Negative; TN)의 개수를 나타내는 행렬

4 식별 모델의 평가와 ROC 곡선

[혼동 행렬]

	식별 결과 양성(+)	식별 결과 음성(-)	합계
양성(+)	TP (TP/(TP+FN)⇒진양성률=민감도) (TP/(TP+FP)⇒정밀도) (TP/(TP+TN)⇒재현율)	FN (FN/(TP+FN)⇒거짓 음성률)	TP+FN
음성(-)	FP (FP/(FP+TN)⇒거짓 양성률)	TN (TN/(FP+TN)⇒진음성률=특이도)	FP+TN
합계	TP+FP	FN+TN	TP+FN+FP+TN

4 식별 모델의 평가와 ROC 곡선

혼동 행렬

민감도 (sensitivity)	양성인 수를 분포로 했을 때의 TP 비율	$= \frac{TP}{(TP+FN)}$
정밀도 (precision)	식별 결과가 양성인 수를 분모로 했을 때의 TP 비율	$= \frac{TP}{(TP+FP)}$
재현율 (recall)	양성인 수를 어느 정도 비율로 제대로 식별 했는지를 나타내는 지표	$= \frac{TP}{(TP+TN)}$
특이도 (specificity)	식별 결과가 음성인 수를 분모로 했을 때의 TN 비율	$= \frac{FP}{(FP+TN)}$
F값 (F measure)	정밀도와 재현율의 조화 평균을 계산하는 값	$= 2 \times \frac{(\text{정밀도} \times \text{재현율})}{(\text{정밀도} + \text{재현율})}$

5 ROC 곡선을 이용한 평가

- AUC 값
  - AUC는 ROC 곡선의 아래 부분 면적 값을 의미
  - 0.9 이상이면 정확도가 높음을 의미

- 왼쪽 위에서 곡선까지 거리
  - AUC 값이 높을수록 ROC 곡선은 왼쪽 위에 가까운 형태
  - 왼쪽 위에서 곡선의 거리인 a가 짧으면 짧을 수록 성능이 좋음

5 ROC 곡선을 이용한 평가

- Youden Index
  - AUC 값과 길이 0.5의 대각선 사이 거리 b가 가장 멀 때 '진양성율 + 거짓양성율=c'로 표현되는 값
  - 이 값이 큰 모델은 가장 좋은 평가를 받은 매개변수

5 ROC 곡선을 이용한 평가

[ROC 곡선]