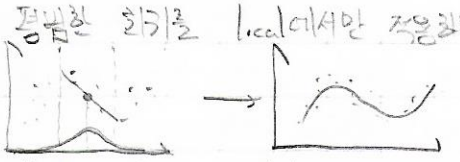
Supervised learning I

다양한 회귀 분석법. (로컬회귀, 다항식 회귀, 단계함수 회귀, spline) 과 주성분 회귀, 부분최소제곱에 대한 내용.

선형회귀의 단점 (단순하고 설득력)
- 비선형관계를 포착하기 어려움 → loess, spline
- 고차원 데이터에 적합하지 않음 → pcr, pls

Loess (Local Regression) : 로컬 회귀

평범한 회귀를 local에서만 적용하는 개념.



한 점을 기준으로 주위의 샘플로만 선형회귀를 수행하면, 한 점과 그 주변의 선형회귀선을 찾을 수 있음. 이를 모든 점에 대해 수행하면 곡선의 형태 (비선형)의 회귀선을 찾을 수 있음.

→ 이웃점만 고려하는 것은 아님 가까울수록 높은 가중치를 주는 개념임.

Loess의 RSS는 가중치를 계산해야 함.

→ $RSS_{loess} = \sum_{i=1}^{n} w_i (y_i - \beta_0 - \beta_i x_i)^2$

흔히 가중치는 한점으로부터의 정규분포의 확률으로

→ $w_i = \exp(-(x_i - x)^2 / 2\tau^2)$

τ (tau)는 정규분포의 분산을 의미함.

다가 크면 높고 좁은 종모양이 되어, 점으로부터 멀리 떨어진 데이터포인트도 고려하게 됨.

다가 작으면 좁고 높은 종모양이 되어 " 는 고려하지 않게 됨.

→ loess의 유연성은 τ에 의해 제어됨.

- 가중치 함수는 사각형의 형태로 쓸 수 있음. 이렇게 되면 밖을 벗어난 데이터 포인트를 아예 고려하지 않게 됨.

Loess vs 일반 선형회귀

Loess는 " 에 비해 비모수적이다.
" 에 비해 유연성이 높다.
" 에 비해 많은 계산 능력이 필요하다.
" 에 비해 많은 데이터가 필요하다.
" 에 비해 클래스하기 않다.

Polynomial Regression : 다항 회귀

고차항 다항회귀는 선형회귀에 비해 유연성이 높고, 이는 최고차항에 의해 좌우된다. 이상치에 매우 민감하며, 복잡한 곡선이 있다.

단계적 회귀는 X의 전체범위를 구간으로 나누고, 각 구간의 값을 상수로 맞추는 아이디어임.

X를 $c_0, c_1 \cdots c_k$의 구간으로 분리함.

각 X에 대해서 구간의 값을 구함.

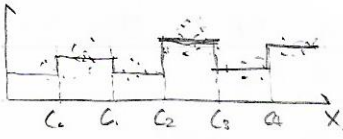
$C_0(X) = I(X < c_0)$, $C_1(X) = I(c_0 < X < c_1) \cdots$

I는 참이면 값으로 1이나 0을 return 함.

위와 같은 C_k를 가설함수로 정함.

→ $y_i = \beta_0 + \beta_1 C_1(x_i) + \beta_2 C_2(x_i) \cdots \beta_k C_k(x_i) + \epsilon_i$

각 범위 내에서 계수인 β에 따라 y_i가 결정됨.



절단점의 갯수에 따라 유연성이 결정되며, 통상 3~5개를 사용함.

→ 3개를 사용할 경우 사분위수와 같이 cut point를 설정함.