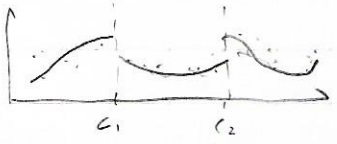
Spline : 다항식 기반 회귀 함수

단계함수를 기초함수로 사용하며, 다항식을 기반으로 함. (단계함수 + 다항 회귀)

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

단, 여기서 C_1, 함수로 1과 0을 return하는 함수가 아니고, 내부적으로 다항 회귀를 수행 함. (일반적으로 3차항)



→ 불연속적인 면을 방지하기 위해 제약을 줌.

$C_1(c_1) = C_2(c_2)$

위와같이 하더라도 해당포인트에서 미분이 불가능 함.

→ 추가적인 제약을 줌

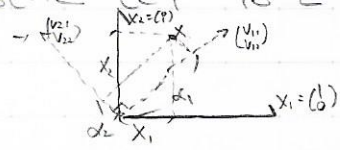
$C_1'(c_1) = C_2'(c_1)$, $C_1''(c_1) = C_2''(c_1)$

이를 통해 모든 점에서 미분이 가능해짐.

Spline은 구간을 사용하기 때문에 loess보다 운용성이 적음. (loess는 모든 점이 데이터)

PCR (Principal Component Regression) : 주성분 회귀

주성분이란 분산이 가장 큰 좌표로 변환한 것.



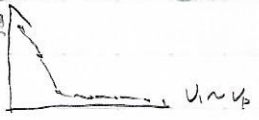
$X = X_1\begin{pmatrix}1\\0\end{pmatrix} + X_2\begin{pmatrix}0\\1\end{pmatrix} = \alpha_1\begin{pmatrix}v_{11}\\v_{12}\end{pmatrix} + \alpha_2\begin{pmatrix}v_{21}\\v_{22}\end{pmatrix}$

→

X_1과 X_2로는 데이터의 최대 분산을 설명할 수 없음. 새로운 좌표계 V_1을 통해서만 최대 분산을 설명할 수 있음

V_1은 첫번째 주성분으로 가장 큰 데이터의 분산을 설명하고, V_2는 두번째 주성분으로 V_1과 직교함. p개의 변수에 대해선 p개의 주성분이 있음. 데이터의 정보는 곧 데이터의 분산임. 주성분은 순서대로 데이터의 분산 정보를 많이 포함하고 있게 됨.

→ 분산설명



이를 통해 적은 변수의 갯수로 회귀식을 찾을 수 있음. PC를 찾는 과정은 비지도 학습의 방법으로 수행되며, (Y를 고려하지 않음) 이로 인해 회귀분석을 수행하는 과정을

→ 많은 경우에 개별적 방향과 출력이 관계가 있다. 모든 경우에 적절한 아님.

PCR에서 사용하는 PC의 갯수에 따라 유연성이 결정됨.

PLS (Partial Least Square) : 최소 부분 자승

좌표변환 : $Z_m = \sum_j \phi_{jm} X_j$

새로운 좌표 군을 원래 좌표 X의 가중합임.

PCR은 X의 최대 분산으로 Z를 찾는 비지도 방법.

PLS는 X와 Y의 최대 분산을 통해 Z를 찾고자 하는 지도학습의 방법으로 이루어짐.

→ $X_1 \cdots X_p$의 변수가 있을때 각 변수를 기준으로 선형회귀를 수행하고 각 회귀계수인 $\beta_1 \cdots \beta_p$ ($= \phi_1 \cdots \phi_p$)를 얻음.

→ 단 이때 모든 변수는 분산이 1이고 평균이 0인 상태로 정규화 되어 있어야 함.

표준편차를 나눔으로서 각 X와 Y의 상관계수를 얻고, 새 좌표계도 상관계수에 따라 변화함.

$Z_1 = \sum_j \phi_{j1} X_j$, Z_1을 기준으로 회귀식을 구하면, $Y \sim \beta_0 + \beta_1 z_1$이 됨.

회귀식이 설명하지 못하는 잔차 $r = Y - \beta_0 - \beta_1 z_1$이 됨. 이후 이 잔차 r을 설명하기 위한 z_2까지 구할 수 있음.