Dimensionsreducering

- "curse of dimensionality" - tveksam proposition
- dock beräkningsmässigt dyrt
- många parametrar kan dölja kolinjäritet
- små modeller med hög R^2 / liten MSE att föredra, de är bättre på att generalisera (okänd data).

- Forward selection

Kör p enkla integrationer. Välj den med högst förklaringsgrad. Testa resterande p-1 variabler; lägg till om R^2/MSE ökar/minskar osv..

- Backward Elimination

  - Börja med alla p variabler.
  Testa alla modeller med p-1 variabler.
  Fortsätt tills R^2/MSE blir sämre.

- PCA Principal Component Analysis

  Grundidé: hitta en ortogonalbas till designmatrisen

  detta översätter alltså X till en icke-korrelerad matris.

$$X = \Phi B$$

(Φ: Basbytesmatris)

Pga stokastiskt, inga exakta lösningar. Ist väljer vi basvektorer statistiskt:

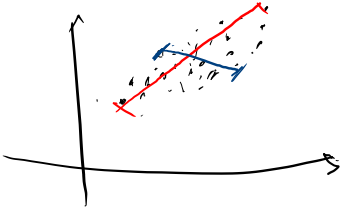
Varje variabel X_i skrivs som en linjärkombination z_i:

$$z_{i1} = \phi_{11} x_1 + \phi_{21} x_2 + \ldots + \phi_{p1} x_p$$

$$\vdots$$

$$z_{ip} = \phi_{1p} x_1 + \phi_{2p} x_2 + \ldots + \phi_{pp} x_p$$

Vi väljer ϕ_{ij} sådana att variationen är maximal!



Nya koordinaterna är rätlinjiga.

$$\max\left[\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\phi_{ij}x_j\right)^2\right]$$

bivillkor: $\sum_{j=1}^{p}\phi_{ji}^2 = 1$ (annars kan vi välja godtyckligt stora ϕ)

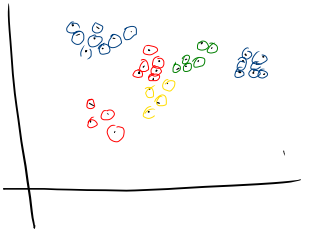
PCA formuleras som ett egenvärdesproblem och löses med Single Value Decomposition (lin. algebra). $O(n^2 m) \rightarrow O(n^3)$ (kvadratisk)

Notera att vi inte har något Y. Detta är inlärning direkt i featurespace, utan några etiketter som säger vad den rätta lösningen är. Dvs $Y = \beta X \rightarrow \beta X$

Oövervakad inlärning

- Data-analys (automatiskt)
- Feature generering
- Label generering
- Typiskt förbehandling i autonoma system

k-means clustering



Grundidé: minimera intraklustervariation
→ hitta k mängder där punkterna är så nära varandra som möjligt.

$$\min_{C_1,\dots,C_k} \left[\sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right]$$

$|C_k|$ – antal punkter i C_k

Kan som vanligt använda en annan norm. Här är det l_2, men gör att använda andra simil. än avstånd; tex inre produkt eller log-odds över frekvenser i datan etc.