LINKÖPING UNIVERSITY

732G12 Data Mining

Föreläsning 1

Josef Wilzén
IDA, Linköping University, Sweden

Dagens föreläsning

- Introduktion
  - Kursupplägg
  - Introduktion till Data Minig och Maskininlärning
- Översikt av metoder
- Modellval och generalisering

Introduktion - Kursupplägg

- Lärare och examinator: Josef Wilzén, mail: josef.wilzen@liu.se
- Struktur: Kurshemsidan, Lisam, Teams
- Kurslitteratur
- Upplägg:
  - Föreläsningar
  - Datorövningar/labbar
  - Projektarbete
  - Datortentamen

Introduktion - Kursupplägg

- Föreläsningar
  - Presenterar metoder, teori
- Datorövningar
  - Använd metoderna för att lösa övningar/problem
  - Förberedelse för projekt samt tentamen
- Projektarbete
  - Större uppgift där ni ska analysera eget datamaterial
  - Seminarie där man ska presentera och opponera
- Datortentamen
  - Påminner om datorövningar, kan ha något teoretiskt inslag
- Generativ AI/chatGPT

Introduktion - Kursutvärdering 2023

- Antal respondenter: 14
- Antal svar: 7 (50%)
- Helhetsbetyg: $4,43$
- Förändringar till detta år:
  - Ny kursplan: (2023)
    - Associations- och sekvensanalys har tagits bort.
    - Mer om icke-linjära metoder har tillkommit $\rightarrow$ lite annan ordning i år
  - Ev mindre förändingar på labbar

Introduktion - Bonuspoäng

- Möjligt att ha 2 bonuspoäng på tentamen.
- Krav: Delta aktivt på 9 av 14 datorlaborationern
- Delta aktivt innebär:
  - Vara på plats och jobba med kursen.
  - Ställa en fråga och visa upp en lösning för läraren.
- Använda bonuspoängen:
  - Kan bara användas under de tre tillfällen kopplade till årets omgång.
  - Kan bara användas för att bli godkänd.

Introduktion - Projektarbete

- Grupper om två.
- Gruppanmälan på Lisam.
- Eget datamaterial.
- Två steg:
  - Välja data, separat inlämning
  - Själva projektet: skriftlig och muntlig redovisning

Introduktion - Förkunskaper

Vad behöhver ni ha med er?

- Linjär algebra
- Matematisk analys
- Programmering
- Regression och Variansanalys
- Statistik teori

Introduktion - Vad är statistik

Från SCB: Med statistik menas vetenskapen om metoder för insamling, bearbetning, redovisning och analys av data. Statistik är också den siffermässiga beskrivningen av en viss företeelse, till exempel i tabellform. Statistik har alltså två betydelser: dels metodiken eller processen, dels informationen eller själva produkten.

Introduktion - Statistisk modellering

- Börjar med data som vi vill analysera.
- Konstruerar en modell som beskriver något samband.
- Osäkerhet/slumptal kan finnas i problemet och ska vara med i modellen.
- Beskriv vår modell matematiskt med fördelningar och ekvationer.
- Härled och implementera parameterskattningar.
- Använd vår modell för inferens och predikation.

Intoduktion - Exempel: Linjär regression

Vi har respons $\mathbf{y}$ och förklarande variabler $\mathbf{X}$.

Vi antar ett linjärt samband mellan $\mathbf{y}$ och $\mathbf{X}$.

Deterministik modell:

$$\mathbf{y} = \mathbf{X}\beta.$$

Stokastisk modell:

$$\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_{\varepsilon}^{\mathbf{2}}).$$

Vi kan skatta parametrar med $\hat{\beta} = (\mathbf{X}^{\top}\mathbf{X})^{-\mathbf{1}}\mathbf{X}^{\top}\mathbf{y}$

Inferens genom att konstruera test eller konfidensintervall

Predikation genom att beräkna respons $\hat{y}$ för nya värden $\mathbf{X}_{\text{test}}$.

Introduktion - Maskininlärning

Från Wikipedia: Det handlar om metoder för att med data ”träna” datorer att upptäcka och ”lära” sig regler för att lösa en uppgift, utan att datorerna har programmerats med regler för just den uppgiften.

Introduktion - Data Mining

Från Wikipedia: ... betecknar verktyg för att söka efter mönster, samband och trender i stora datamängder. Verktygen använder beräkningsmetoder för multivariat statistisk analys kombinerat med beräkningseffektiva algoritmer för maskininlärning och mönsterigenkänning hämtade från artificiell intelligens.

Introduktion - Maskininlärning vs Data Mining

Sorta överlapp mellan maskininlärning och data mining.

- Maskininlärning, mest fokus på predikationer.
- Data mining, mer fokus på att utforska dataset och hitta samband.

"Statistical learning" $\approx$ Maskininlärning.

Introduktion - Diskussion

Bilda grupper om 3 och diskutera er fram till två stora eller komplexa datamängder, beskriv vad det är för mängder och vilken information som kan finnas.

Introduktion - Exempel på stora datamaterial

- Transaktionsdatabaser
- Hälsoregister
- Sociala nätverk
- Väder / klimatdata
- Börsdata
- Korpus

Introduktion - Dataset

Tabellformat:

id	Variabel 1	Variabel 2	Variabel 3	$\cdots$
1	.	.	.	$\cdots$
2	.	.	.	$\cdots$
3	.	.	.	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

- Rader är objekt, record, observation, transaktion etc.
- Kolumner är variabler, attribut, kovariat etc.
- Finns andra format, t.ex. tidsserier, texter, bilder etc.

Introduktion - Dataskalor

Finns många olik dataskalor:

- Nominell
  - Ex. "RÖD", "GRÖN", "BLÅ"
- Binär
  - 0 eller 1
- Ordinal
  - Ex. "LÅG", "MELLAN", "HÖG"
- Intervall/kvot
  - Ex. $[0, 1], \mathbb{R}, \mathbb{R}^+$

Analysprocessen

Exempel på processen:

1. Problemformulering
2. Samla in data
3. Datahantering
   - Gå från rådata till användbar data
   - Ta bort sakanade värden, outliers, skalning
   - Kolla på metadata, plottar
4. Modellering
   - Skatta modeller och gör predikationer
5. Evaluering
   - Jämför med problemformuleringen!

Datahantering

- Saknade värden
  - Eliminera objekt eller attribut
  - Skatta saknade värden (imputering)
  - Ignorera saknade värden
- Förbearbeta data
  - Aggregering
  - Urval
  - Reducera dimensionalitet
  - Diskretisering
  - Variabelomvandling
  - Skalning

Metoder

- Supervised learning
  - Klassificering och regression
  - Varje observation består av en responvariabel och förklarande variabler
- Unsupervised learning
  - Ett antal variabler, men ingen responsvariabel
- Semisupervised learning
  - Värdet på responsvariabeln finns bara på en delmängd av observationerna
- Reinforcement learning
  - Handlar om att lära sig agera optimalt i en miljö.
  - Använda data för att lära sig ett beteende.

Exempel på metoder

- Supervised learning:
  - Prediktera morgondagens elpris
  - Identifiera objekt i en bild
  - Skapa en beskrivande text från en bild
- Unsupervised learning:
  - Identifiera köpmönster
  - Hitta liknande grupper av läkemedelsmolekyler
  - Hitta bedragare bland bankkunder

Kursens områden

- Modellering
- Klassificering
- Regression
- Klusteranalys

Att välja en lämplig modell

För en lämplig modell gäller:

- Modellen ska fånga upp den relevanta strukturen i problemet och kunna svara på frågeställningen!
- Modellen ska vara så enkel som möjligt.
- Modellen ska gå att beräkna i rimlig tid.
- Vi vill att modellen ska **generalisera** till ny liknande data.

Att välja en lämplig modell - Exempel

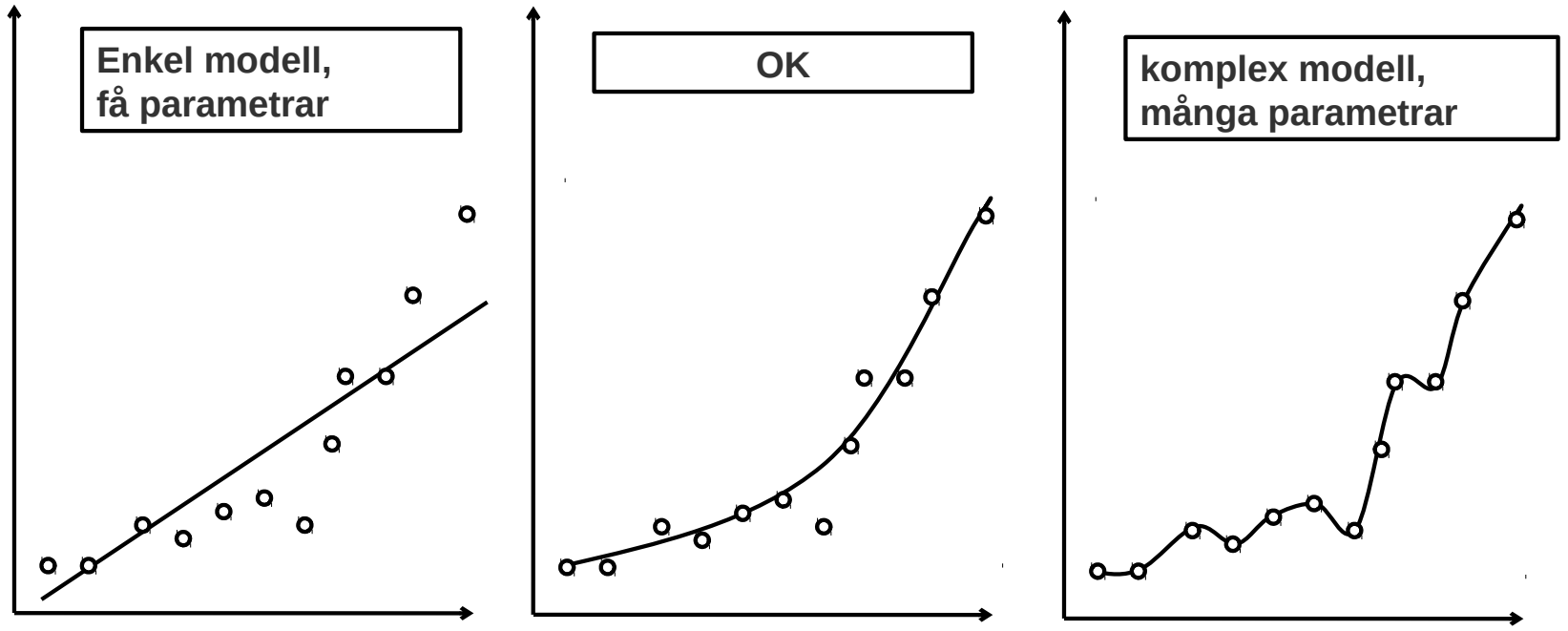
Standardmodell med additivt brus:

$$y = f(x|\omega) + \varepsilon, \qquad \mathbb{E}[\varepsilon] = 0, \qquad \mathbb{V}[\varepsilon] = \sigma_\varepsilon^2.$$

- f är en okänd funktion
- ω är parametrar till f
- ε är slumpmässig felterm
- Exempel på funktioner:

$$f(x \mid \omega) = \omega_1 \cdot x + \omega_2$$
$$f(x \mid \omega) = \omega_3 \cdot \sin(2\pi\omega_1 x + \omega_2) + \omega_4$$
$$f(x \mid \omega) = \exp(-\omega_1(x - \omega_2)^2)$$
$$f(x \mid \omega) = \omega_1 x_1 + \omega_2 x_2 + \omega_3 x_3$$

Vilken modell är lämplig?



- Underanpassning: Modellen fångar inte upp relevanta strukturer i problemet.
- Överanpassning: Modellen fångar upp bruset i datan.

Modellval

- För att hjälpa oss i modellval definerar vi en felfunktion/kostnadsfunktion
  - Givet data ger kostnadsfunktionen ett värde på hur bra modellen anpassar datan.
- Vanliga exempel:
  - Mean squared error (MSE)

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i \mid \hat{\omega}))^2$$

  - Mean absolute error (MAE)

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - f(x_i \mid \hat{\omega})|$$

  - Cross entropy loss
  - Missclassification loss

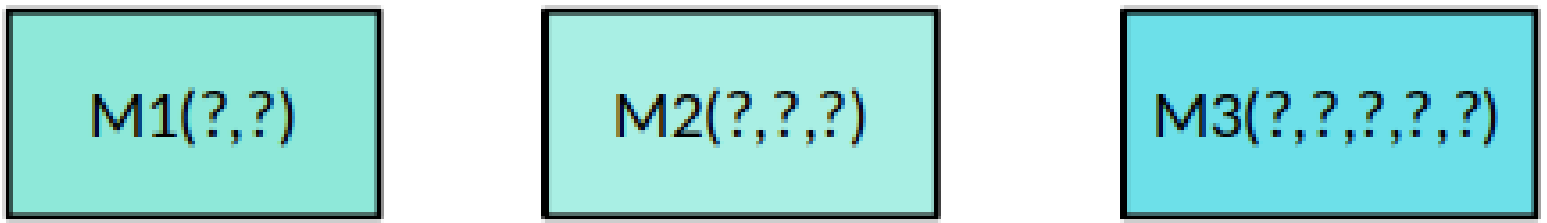
Modellval

Standardprocessen:

- Dela upp din data i slumpmässiga delar:
  - Träningsdata
  - Valideringsdata
  - Testdata
- Olika proportioner kan användas
  - $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$
  - $(0.5, 0.25, 0.25)$
  - $(0.8, 0.1, 0.1)$
- Träningsdata bör inte vara minst.
- Notera: i många källor/på internet så används ofta ”test data” för att beskriva valideringsdata. Dvs de använder ”test data” för att välja den bästa modellen

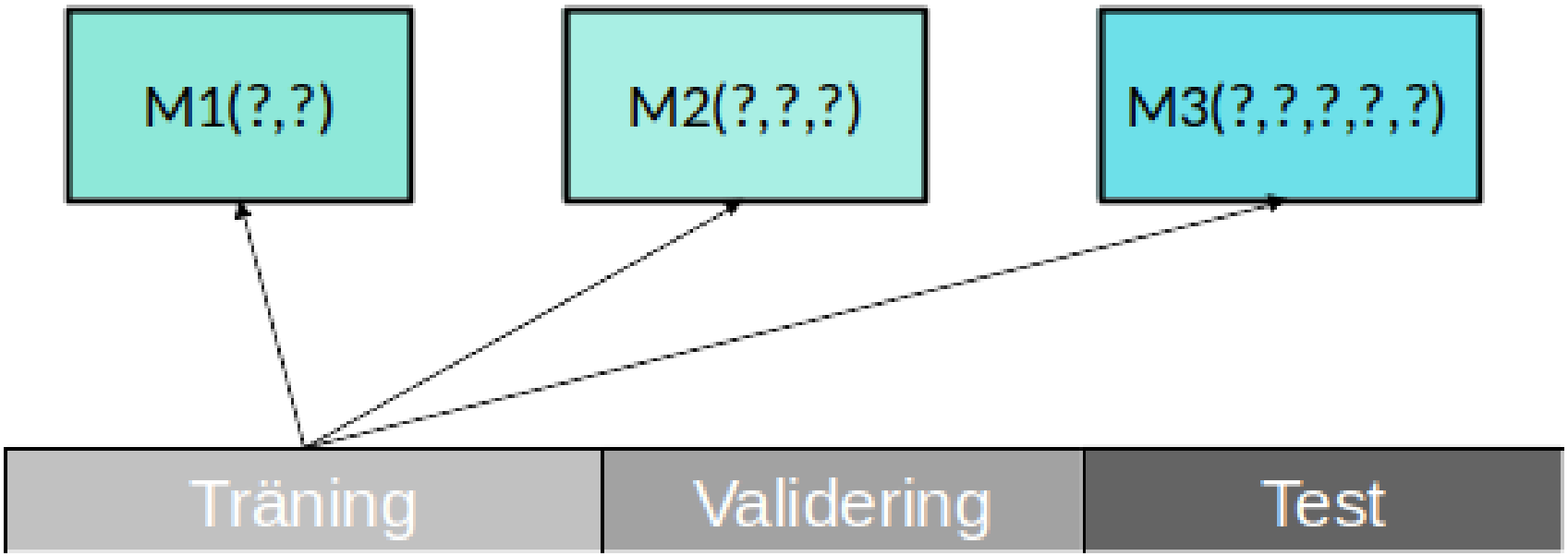
Hur hittar vi bästa modellen?

Börja med att ta fram några kandidatfunktioner



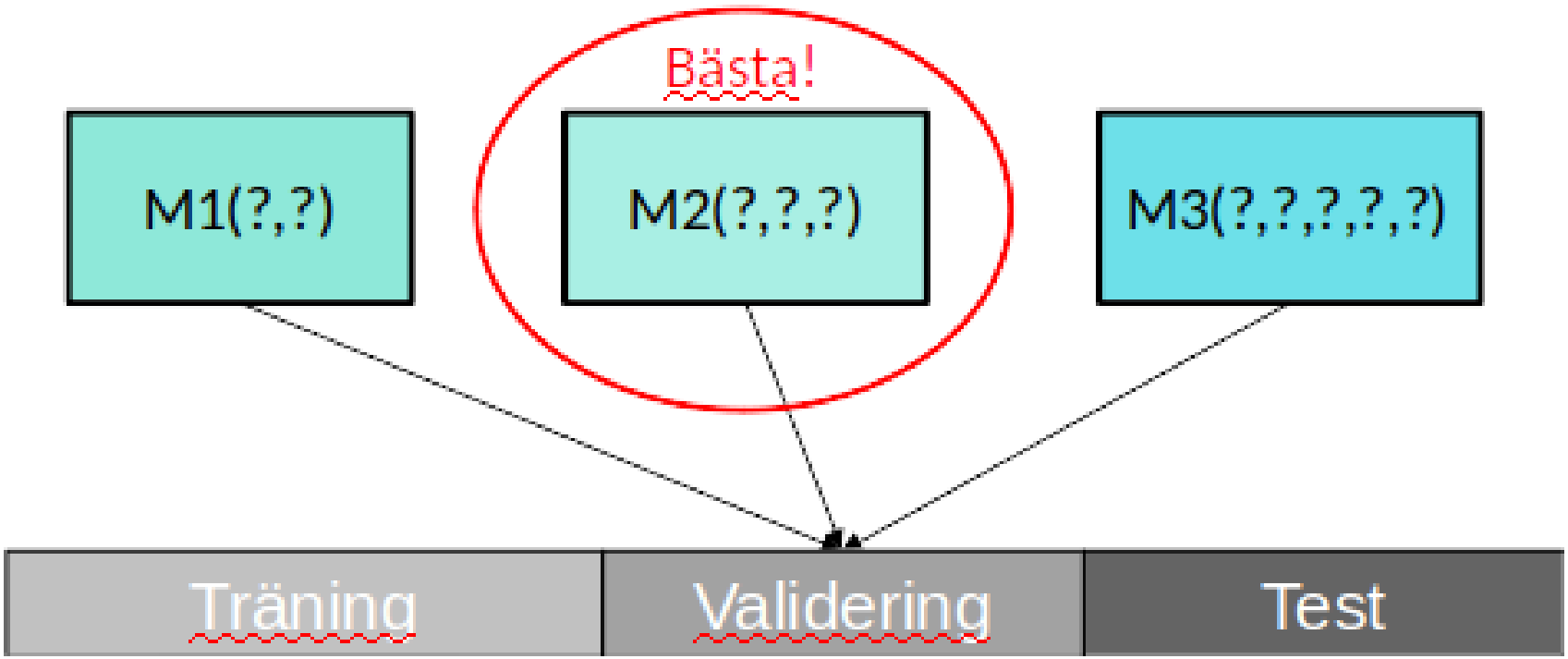
Träningsdata

Använd träningsdata för att skatta parametrarna i modellerna.



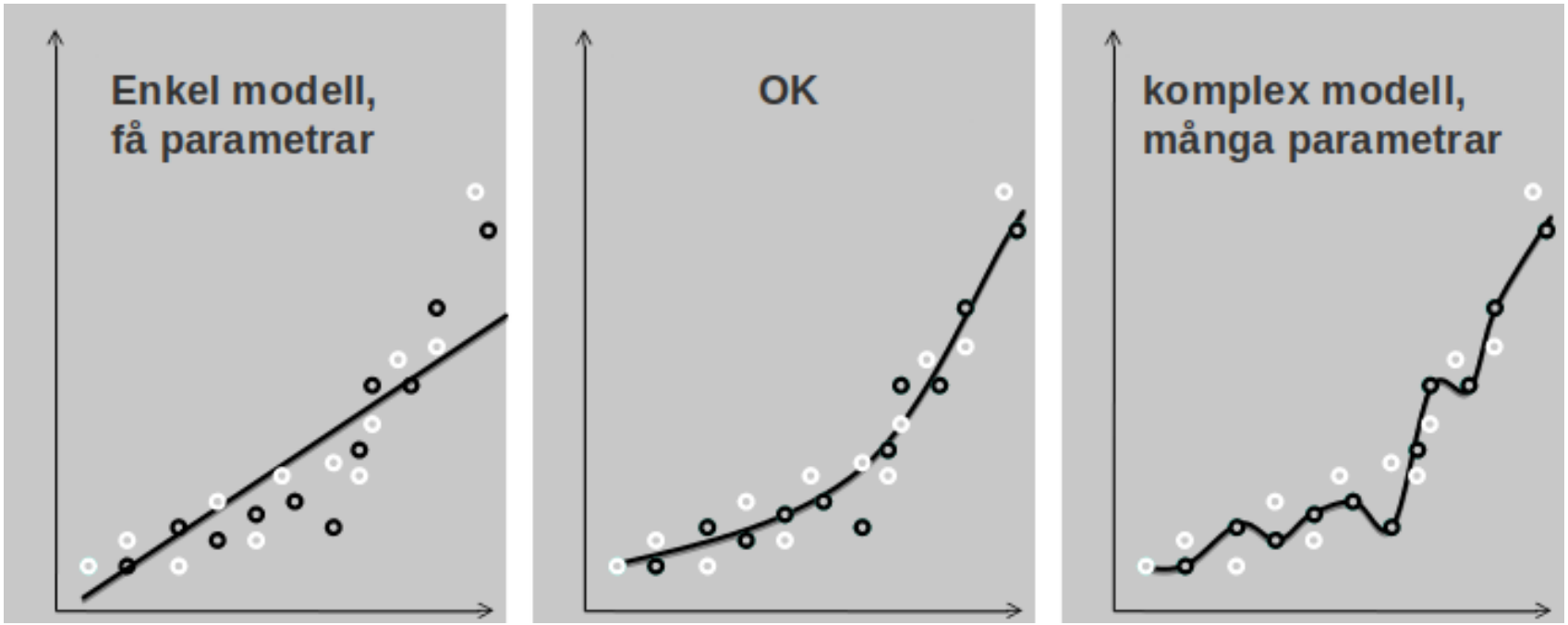
Valideringsdata

Använd valideringsdatan för att välja den bästa skattade funktionen utifrån lämplig felfunktion.



Vi kan iterera mellan att skatta med träningsdata och utvärdera på valideringsdatan

Validringsdata

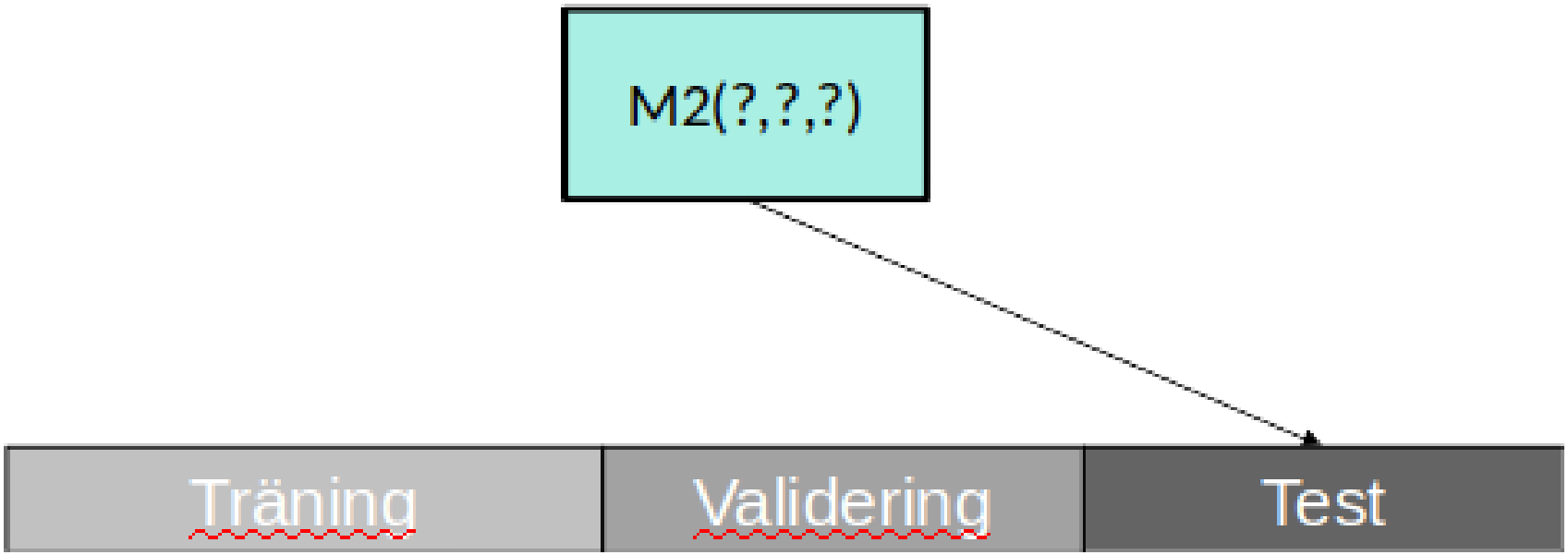


Här är träningsdata svart och valideringsdata vit.

Testdata

Vi använder testdatan för att få en väntevärdesriktig skattning av felfunktionen på ny data $\rightarrow$ vi skattar testfelet

Vi bör *inte* ändra något på modellen när vi ska använda den på testdata. Varför?



Leave-one-out cross-validation

Givet ett antal modeller görs följande för varje modell:

1. Ta bort en observation och anpassa modellen till återstående data.
2. Använd den anpassade modellen för att beräkna felet på den borttagna observationen.
3. Upprepa 1 och 2 för alla datapunkter. Beräkna genomsnittet för felfunktionen över alla observationer.

Välj modellen med lägs genomsnittlig felfunktion.

OBS! Om vi har n observationer måste vi anpassa varje modell n gånger.

Leave-one-out cross-validation

$$\begin{pmatrix} x_{11} & x_{21} & & x_{p1} & y_1 \\ x_{12} & x_{22} & & x_{p2} & y_2 \\ & & & & \\ x_{1j} & x_{2j} & & x_{pj} & y_j \\ & & & & \\ & & & & \\ x_{1n} & x_{2n} & & x_{pn} & y_n \end{pmatrix}$$

K-fold cross-validation

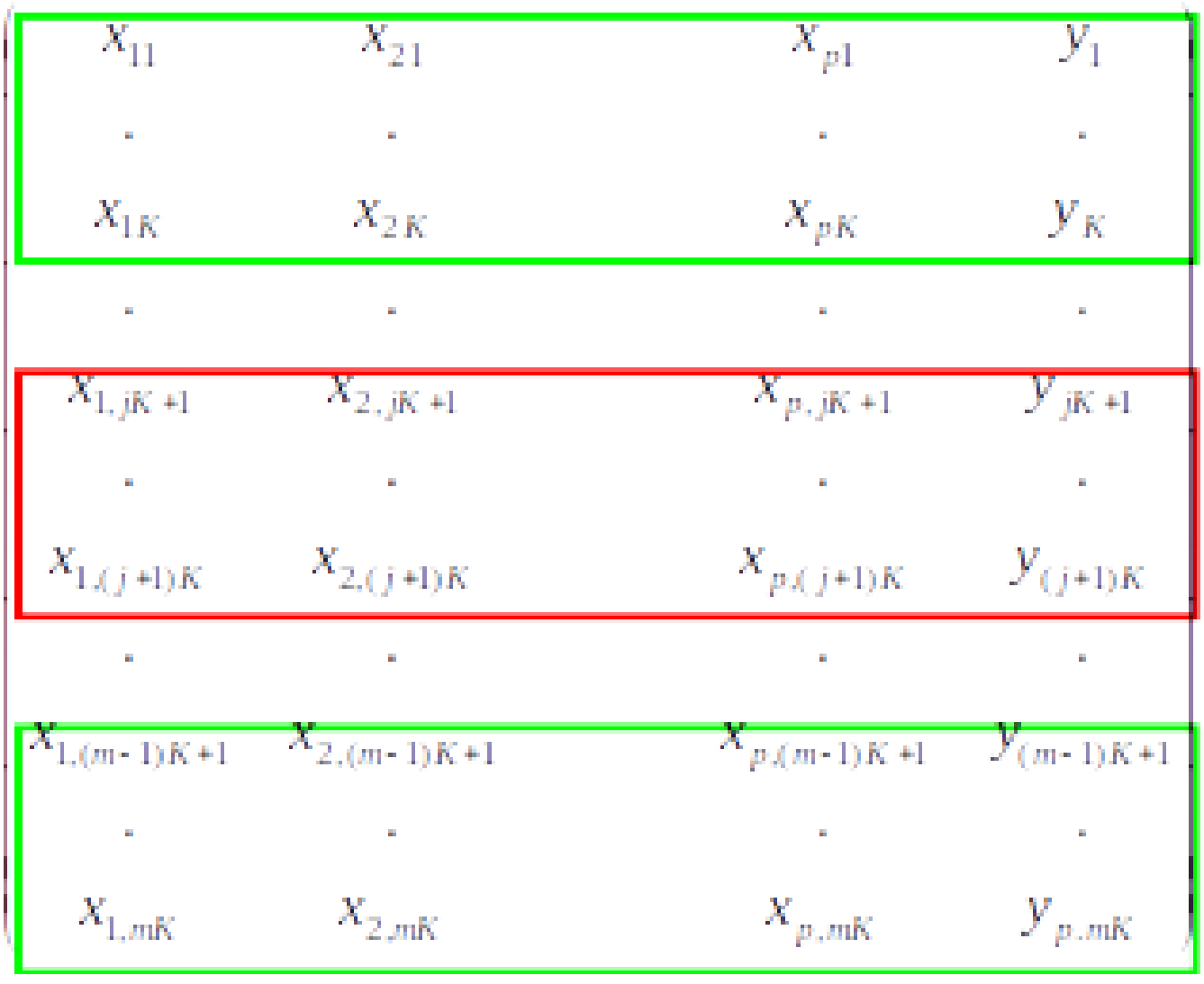
Dela upp datan i k olika block. Givet ett antal modeller görs följande för varje modell:

1. Ta bort ett block och anpassa modellen till återstående data.
2. Använd den anpassade modellen för att beräkna felet på de borttagna observationerna.
3. Upprepa 1 och 2 för alla block. Beräkna genomsnittet för felfunktionen över alla olika block.

Välj modellen med lägs genomsnittlig felfunktion.

OBS! Om vi har K block måste vi anpassa varje modell K gånger.

K-fold cross-validation



Bias, varians och brus

Givet en modell

$$y = f(x) + \varepsilon, \qquad \mathbb{E}[\varepsilon] = 0, \qquad \mathbb{V}[\varepsilon] = \sigma_\varepsilon^2$$

med skattad funktion $\hat{y} = \hat{f}(x_{\text{test}})$.

Om vi beräknar MSE får vi:

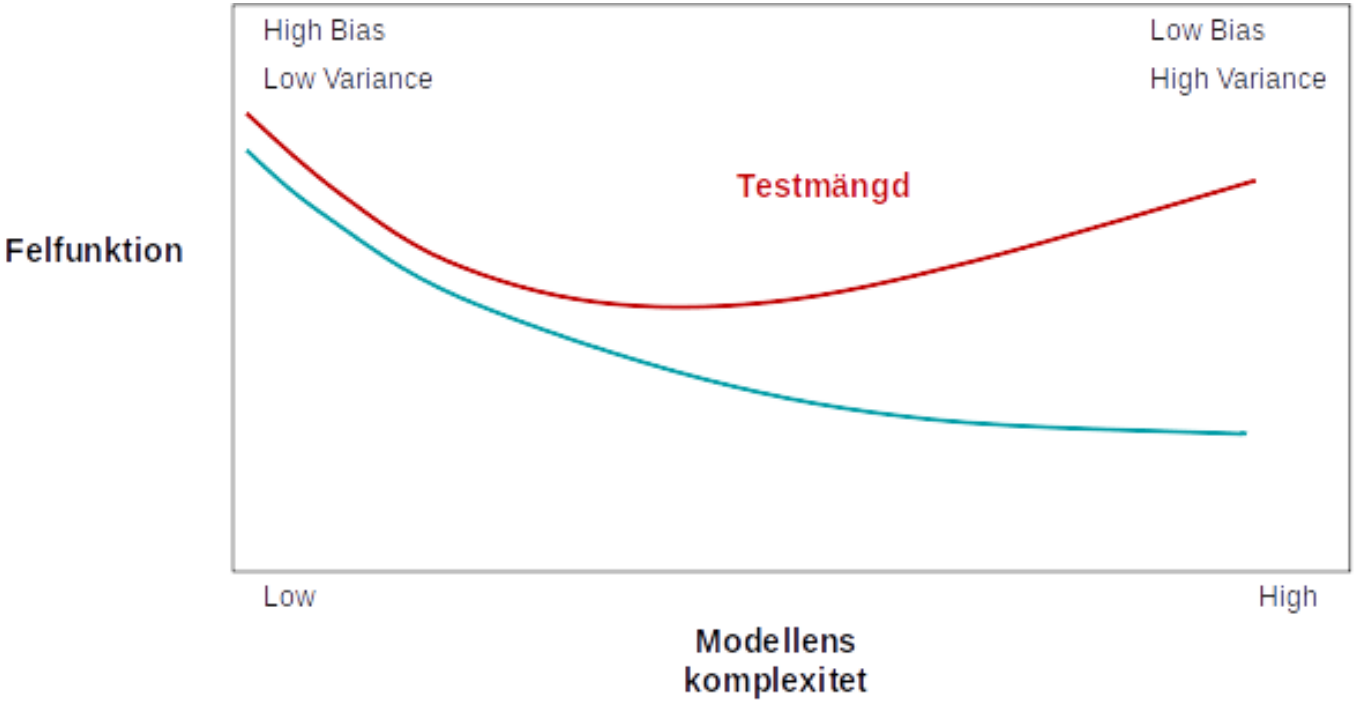
$$\mathbb{E}\left[\left(y_{\text{test}} - \hat{f}(x_{\text{test}})\right)^2\right] = \mathbb{V}[\varepsilon] + \mathbb{V}\left[\hat{f}(x_{\text{test}})\right] + \text{Bias}\left[\hat{f}(x_{\text{test}})\right]^2$$

- Brusvariansen $\mathbb{V}[\varepsilon]$: irreducibelt brus
- Modellens varians $\mathbb{V}[\hat{f}(x_{\text{test}})]$: hur mycket ändras $\hat{f}$ när vi byter dataset
- Modellens systematiska fel $\text{Bias}[\hat{f}(x_{\text{test}})]$: modelleringsfel i modellen

Bias-variance-trade-off

Vi vill ha:

- Låg bias och låg varians $\rightarrow$ en modell som generaliserar bra på ny liknande data!



- För vissa komplexa modeller: Double descent $\rightarrow$ testfelet minskar för att sedan öka, men för att sedan minska igen

Frågor?