Föreläsning 1 - Introduktion till Data Mining och Maskininlärning

Josef Wilzen

2020-08-17

Outline

Outline

Kursupplägg

- Lärare och examinator: Josef Wilzén
- Upplägg
  - Förläsningar
  - Datorövningar/labbar
  - Projektarbete = 2.5 hp (U,G)
  - Tenta = 5 hp (U,G,VG)

Kursupplägg

- Föreläsningar
  - Ny information och metoder presenteras
- Datorövningar
  - Parvis lösning av övningar som behandlar tidigare material
- Projektarbete:
  - Större uppgift där ni ska analysera data mer självständigt
- Seminarium
  - Projektarbetet presenteras och opponeras
- Tentamen
  - Påminner om datorövningarna, kan ha teoretiska frågor

Kursutvärdering 2019

- Antal respondenter: 23
- Antal svar: 10 (Svarsfrekvens: 43,48 %)
- Vilket helhetsbetyg ger du kursen? 3.40
- Förändingar:
  - Variabelselektion
  - Icke-linjär regression
  - Nya kursböcker
  - Lite mer fokus på innehåll/material vs projekt

Förkunskaper

Vad behövs sen tidigare?

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

Outline

Statistik

Från Wikipedia: "Statistik är en gren inom tillämpad matematik som sysslar med insamling, utvärdering, analys och presentation av data eller information."

- Vad ni har sysslat med i två år...
- Tradionellt så har statistiker sysslat med
  - Relativt små och "enkla" dataset
  - Inferens: hypotestest och konfindesintervall
  - Prediktioner
  - Strävar efter att använda så korrekta antagaden som möjligt
  - Modeller som är lätta att beräkna

Statistik modellering

- Startar med ett problem som relaterar till data
- Konstruktikon av abstrakta modeller som beskriver sambandet/relationen mellan ett antal variabler
- Utgår från att det finns osäkerhet eller slumpelement i problemet och att den ska vara med i modellen
- Beskrivs matematiskt: sannolikhetsfördelningar och ekvationer

Statistik modellering (forts.)

- Metoder för att skatta modellens parameterar härleds och implementeras
- Sen används modellen för:
  - inferens = kvantifiera osäkerhet i modellen och dess parameterar
  - prediktion = utvärdera modellen i nytt sammanhang/ny data/framtiden etc
- Del i större vetenskapligt arbete

Exempel: linjär regression

Repons: y och förklarande variabler X

- Vi antar att det finns ett linjärt samband mellan y och X
- En deterministiskt modell:

$$y = X\beta$$

- En slumpmodell

$$y = X\beta + \varepsilon \qquad \varepsilon \sim N\left(0, \sigma^2\right)$$

- Skattning: $\hat{\beta} = \left(X^T X\right)^{-1} X^T y$
- Inferens: vi kan konstuera olika test för modellen, tex t-test, eller konfidensintervall för β
- Prediktion: Vi kan beräkna anpassade värden $\hat{y}$ för nya värden på X_{test}

Maskininlärning

Maskininlärning (Machine learning):

- System/modell/algoritm som kan lära sig från data utan att att explicita instruktioner/regler behöver ges
- Systemet lär sig av "erfarenhet" från data
- Används för att fatta datadrivna beslut och prediktioner
- Kan beskrivas som matematiska modeller eller algoritmer
- Ofta fokus på skalbarhet och beräkningseffektivitet
  - Stora datormängder, snabba beräkningar
- Ofta del i ett större system/programvara
- Ursprung i datorvetenskapen, del av området artificiell intelligens (AI)

Data Mining

Data Mining (datautvining):

- Metoder för att hitta mönster, samband stora datamängder
- Skärningspunkten mellan:
  - datorvetenskap, databashantering och statistik
- Ofta fokus på skalbarhet och beräkningseffektivitet
- Fokus är ofta på explorativ dataanalys (vi vet inte var vi letar efter...)

Maskininlärning vs Data Mining

Stort överlapp mellen fälten!!!

- Maskininlärning: mer fokus på prediktioner
- Data mining: mer explorativ dataanalys och hitta nya samband

Statistical learning $\approx$ Maskininlärning
Inom "learning" maskininlärning betyder ofta "estimation" i statistik.

Diskussionsfråga

Diskutera i grupper om 3:

- Ge ett exempel på två stora eller komplexa datamängder där man kan finna intressant information.
- Beskriv mängden och ange vilken information som kan finnas däri

Exempel på stora datamaterial

- Transaktionsdatabaser
- Elektroniska hälsoregister
- Register av telefonsamtal
- Sociala nätverk
- Webbsidologgar
- Korpus (stor samling språkliga data)
- Väder och klimatdata
- Astrofysisk data

Maskininlärning

- Metoderna kan vara antingen deterministiska eller probabilistiska
  - Del beskivs med hjälp av algoritmer (=ingen explicit "modell")
- Många klassiska statistika metoder används inom maskininlärning, ex:
  - Linjär regression
  - Logistik regression
  - Principal komponent analys (kommer i Multivariata metoder)
- Maskininlärning $\approx$ (nya) statistiska metoder som passar för komplexa/stora dataset
  - Mer fokus på att lära sig en "funktion" än linjära modeller
  - Mer fokus på prediktion

Dataset

Tabellformat

case	variable 1	variable 1	variable 1	variable 1	·	·	·	·	·
1									
2									
3									
⋮									

- Rader:
  - object, record, observation, transaction
- Kolumner:
  - Variable, attribute, feature, covariate
- Finns många andra format: tidseriser, text, kartor, grafer, videor

Olika dataskalor

1. Nominell: ”Röd”, ”Grön”, ”Blå”
    1. Binär: 0 eller 1, TRUE eller FALSE
2. Ordinal: ”låg”, ”mellan”, ”hög”
3. Intervall/Kvot: $[0, 1]$, $\mathbb{R}$, $\mathbb{R}^+$

Analysprocessen

Övergripande:

- Problemformulering
- Samla in data
- Datahatering och utforskning: Rådata $\rightarrow$ Användbar data
  - NA, felaktigheter, rensa, outliers, skalning
  - Metadata, plottar, beskrivande statistik
- Modellering: skatta modeller och gör prediktioner
- Evaluering: jämför med problemformuleringen!

Datahatering

- Saknade värden
  - Eliminera objekt eller attribut (problem många saknade värde, attribut kan vara viktiga)
  - Skatta saknade värde
  - Ignorera saknade värde
- Förbearbeta data
  - Aggregering
  - Urval
  - Reducera dimensionalitet
  - Diskretisering: göra en numerisk variabel till Nominell/Ordinal
  - Variabelomvandling (feature engingering)

Översikt av metoder

- Supervised learning (övervakand inlärning):
  - Klassificering och regression
  - Varje obs består av en repsonevariabel (y) och förklarande variabler (X).
  - Värdet på repsonevariabel: etiketter/ufall
- Unsupervised learning (oövervakad inlärning)
  - Ett antal variabler (X), men ingen reponsevariabel

Översikt av metoder

- Semisupervised learning
  - Värdet på responsvariabeln finns bara tillgängligt på en delmängd av data
  - Data: $\{X_a, y_a\}$,$\{X_b\}$
  - Använda all data för att förstå relationen $X \rightarrow y$
- Reinforment learning:
  - Handlar om att lära sig att agera optimat i en miljö: robotar, självkörande bilar
  - Data $\rightarrow$ Lära sig beteende

Exempel på olika inlärning

- Övervakad inlärning:
  - Prediktera morgondagens elkonsumtion. Utgå från väderdata, kalenderregister samt tidigare konsumtion.
  - Identifiera vilka siffror som finns i en bild.
  - Skapa beskrivande texter för bilder automatiskt
- Oövervakad inlärning:
  - Identifiera köpmönster vilket kan användas för reklamkampanjer.
  - Hitta liknade grupper av läkemedelsmolekyler
  - Hitta bedragare bland bankkunder

Kursens områden – översikt

- Klassificering:
  - y: Binär, nominell, ordinal
  - Bygg modell för ändligt antal klasser (utfall)
  - Prediktera klass för ny observation
- Regression:
  - y: $\mathbb{R}$,$\mathbb{R}^+$, $a \leq y \leq b$
  - Bygg modell för kontinuerligt utfall
  - Prediktera utfall för ny observation
- Modeller: beslutsträd, neurala nätverk, k-närmaste grannar

Kursens områden – översikt

- Klusteranalys:
  - Dela upp observationer i grupper (kluster)
  - Placera ny observation i ”rätt” kluster
  - Undersöka klustrens egenskaper
- Associationsanalys:
  - Hitta samband som sker ofta i transaktionsdata, ex. $\{\text{blöjor}\} \rightarrow \{\text{öl}\}$

Hur definieras en lämplig modell?

- **Modellen ska fånga den relevanta strukturen i problemet och kunna svara på frågeställningen**
  - En modells lämplighet beror alltid på sin kontext!
- Givet ovan: modellen ska vara så enkel som möjlig
  - Ockhams rakkniv
  - Vad vinner jag på att ha en mer komplicerad modell?
- Modellen ska att gå att beräkna i en rimlig tid

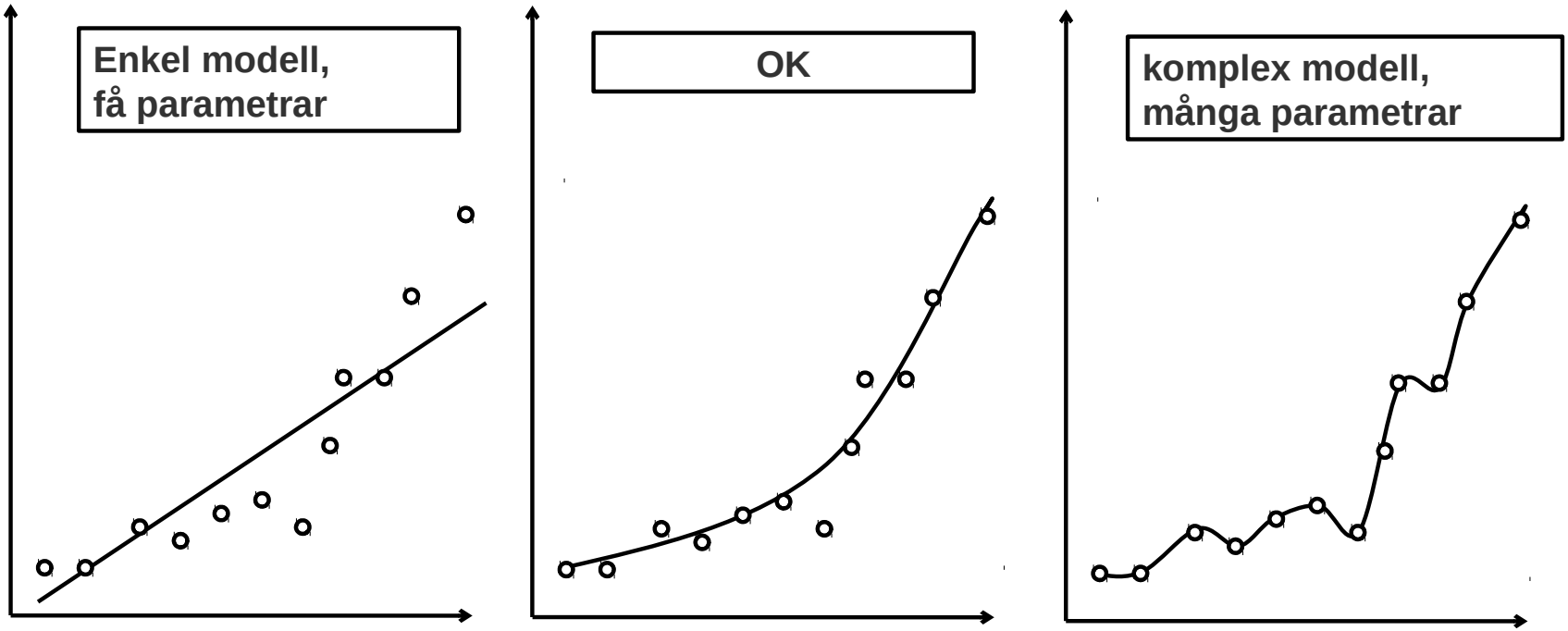
Hur definieras en lämplig modell?

$$y = f(x|w) + \varepsilon \qquad E[\varepsilon] = 0 \qquad V[\varepsilon] = \sigma^2$$

- f är en okänd funktion
- w är ev parameterar till f
- ε är en slumpmässig felterm
- Ex:
  - $f(x|w) = w_1 \cdot sin(2\pi \cdot x \cdot w_2)$
  - $f(x|w) = w_1 \cdot exp(-x \cdot w_2)$
  - $f(x|w) = x_1 w_1 + x_2 w_2 + x_3 w_3$

Hur definieras en lämplig modell?

- Underanpassning (underfitting): modellen fångar inte relevanta strukturer i problemet
- Överanpassning (overfitting): Modellen fångar upp bruset i data



Modelval

- Vi vill att modellen ska fungera bra på ny data
- Definera en felfunktion/kostnadsfunktion (error funciton/cost function):
  - Givet en datamängd så anger felfunktion hur bra modellen anpassar den aktuella datamängden
- Vanliga exempel:
  - Squared error loss: Tänk normalfördelning, ex: linjär regression

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

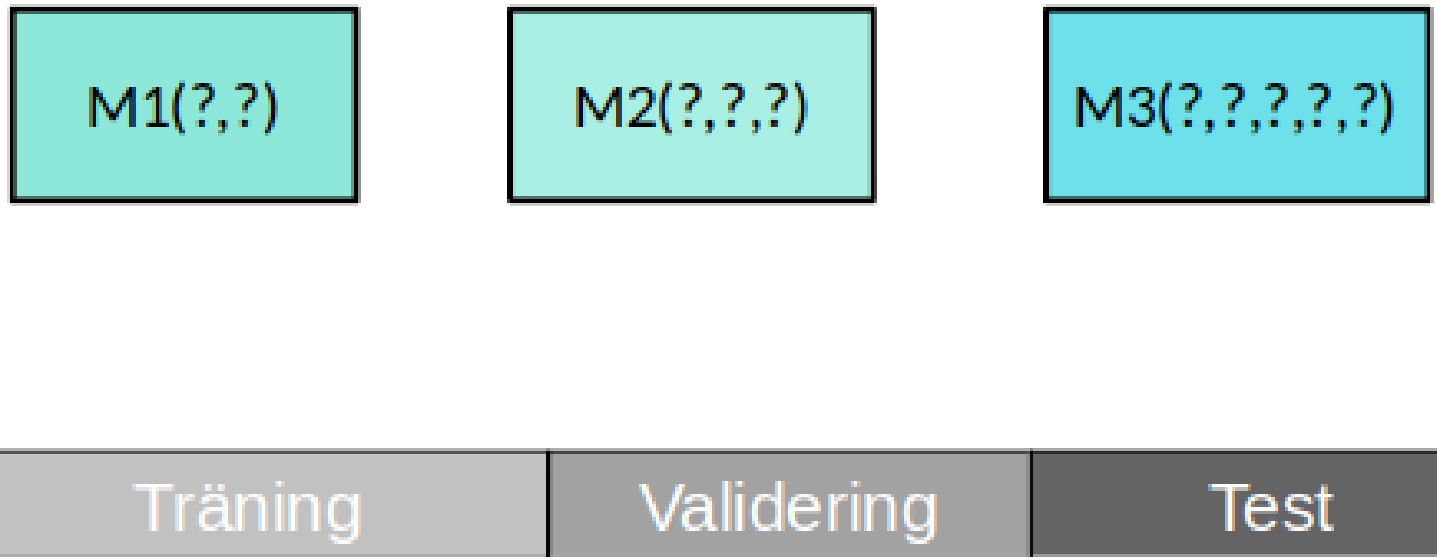
  - Cross entroy loss: Vanligt för klassificering med neurala nätverk
  - Misclassification loss

Modelval

- Dela upp dataset (om ej för liten) i slumpmässiga delar:
  - Träning (train)
  - Validering (validation)
  - Test
- Olika proportioner kan väljas: $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$,$(0.6, 0.2, 0.2)$,$(0.7, 0.15, 0.15)$
- OBS! Träning bör inte ha mindre proportion än de andra.

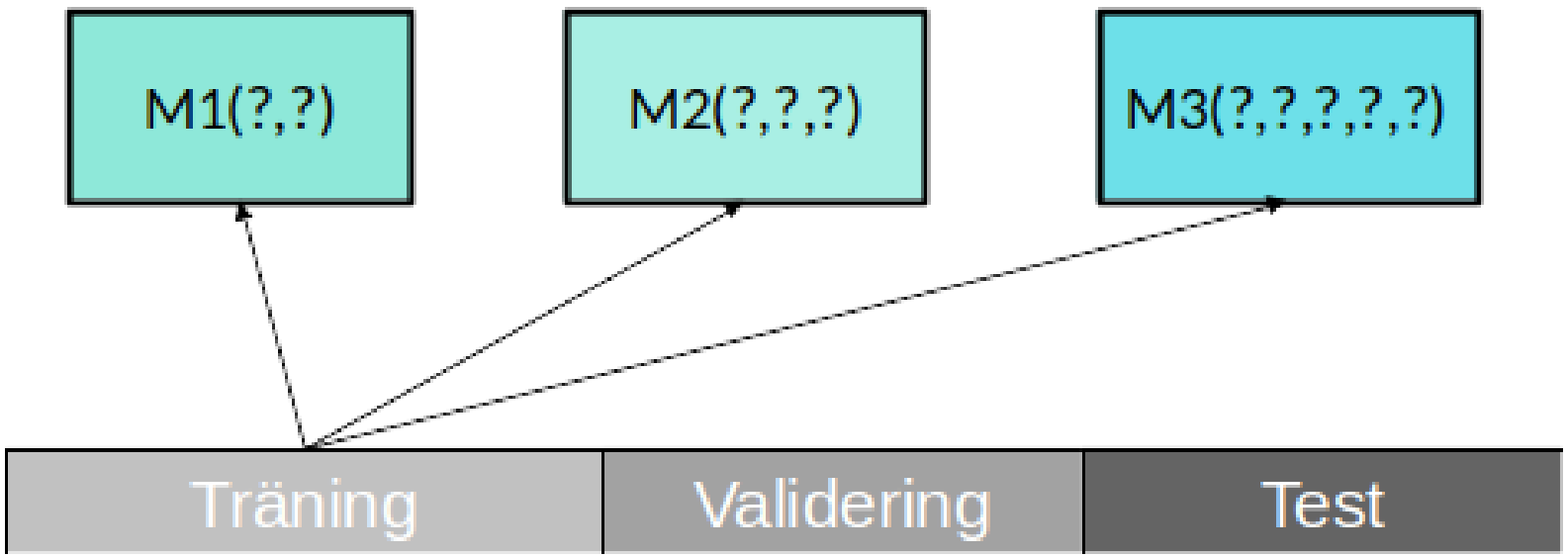
Hur hittas den bästa modellen?

- Ta fram några kandidatmodeller. De kan ha olika komplexitet.



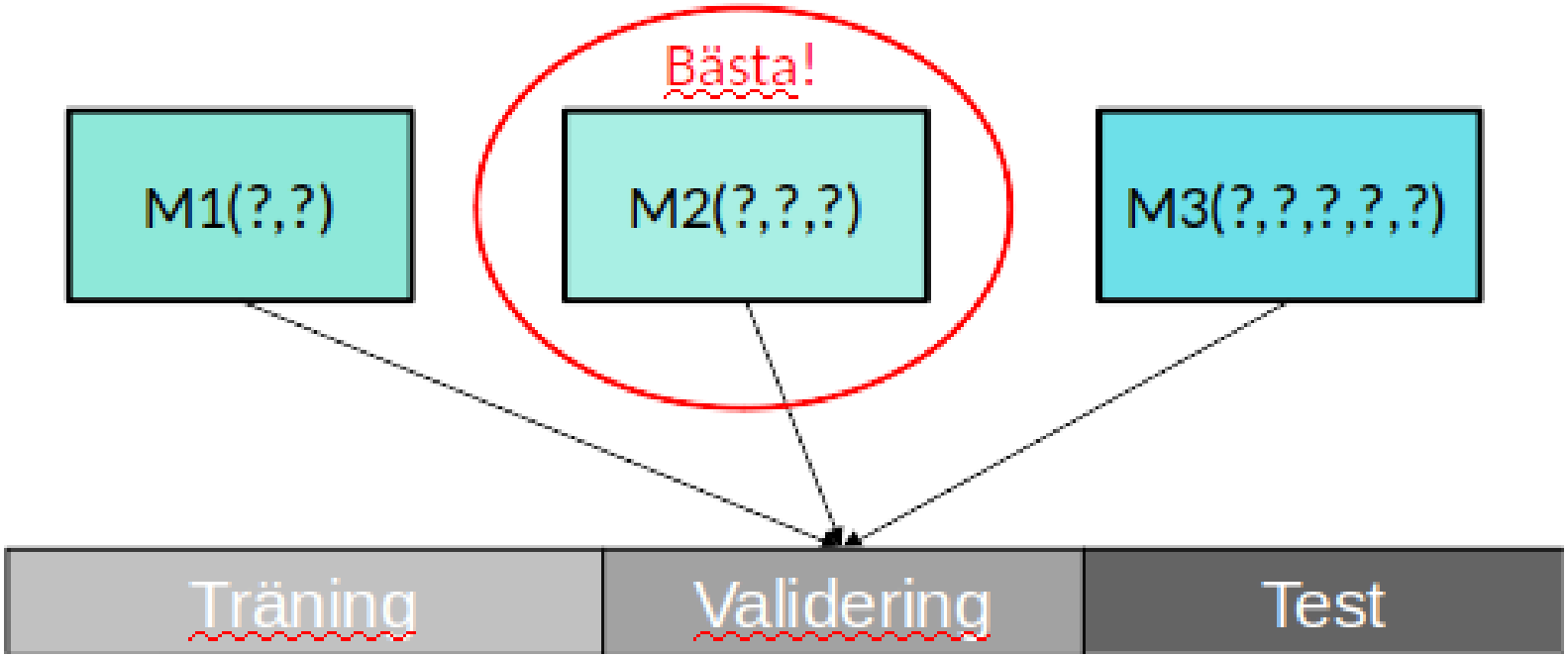
Träningsmängden

- Används för att skatta parametrarna i modellerna

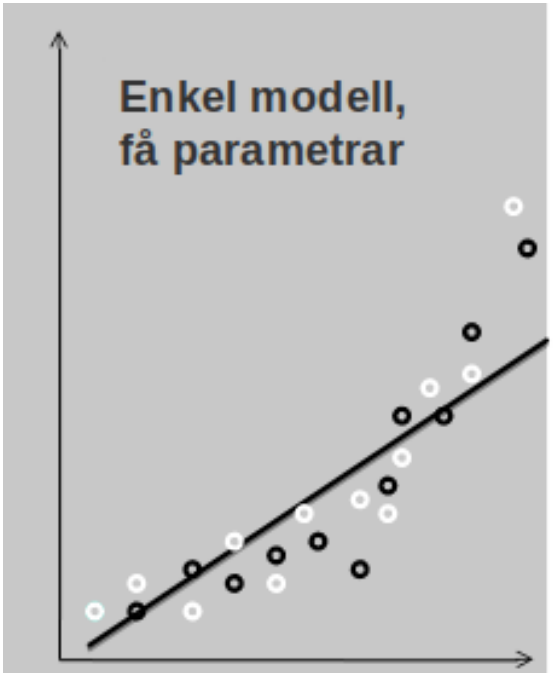


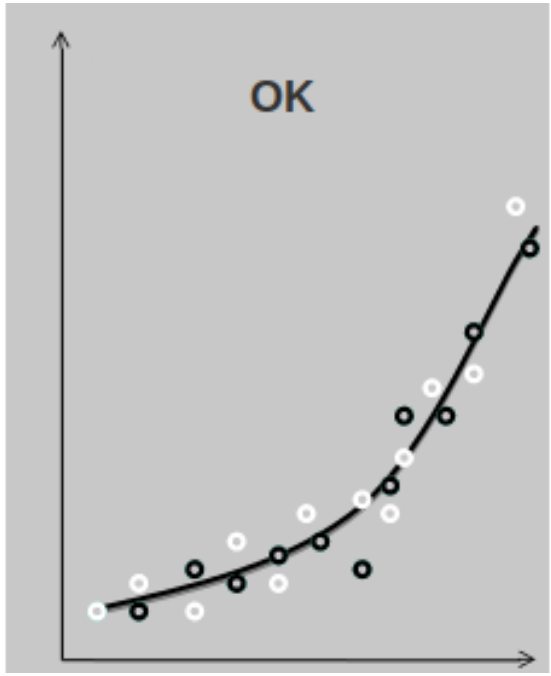
Valideringsmängden

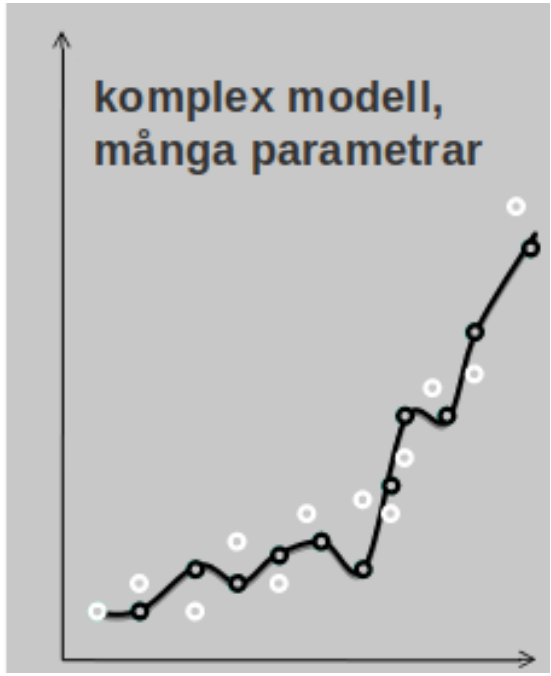
- Används för att välja den bästa skattade modellen utifrån lämplig felfunktion.
- Vi kan iterara mellan att skatta nya modeller på träningsdata och utvärdera dem på valideringsdata.



Valideringsmängden



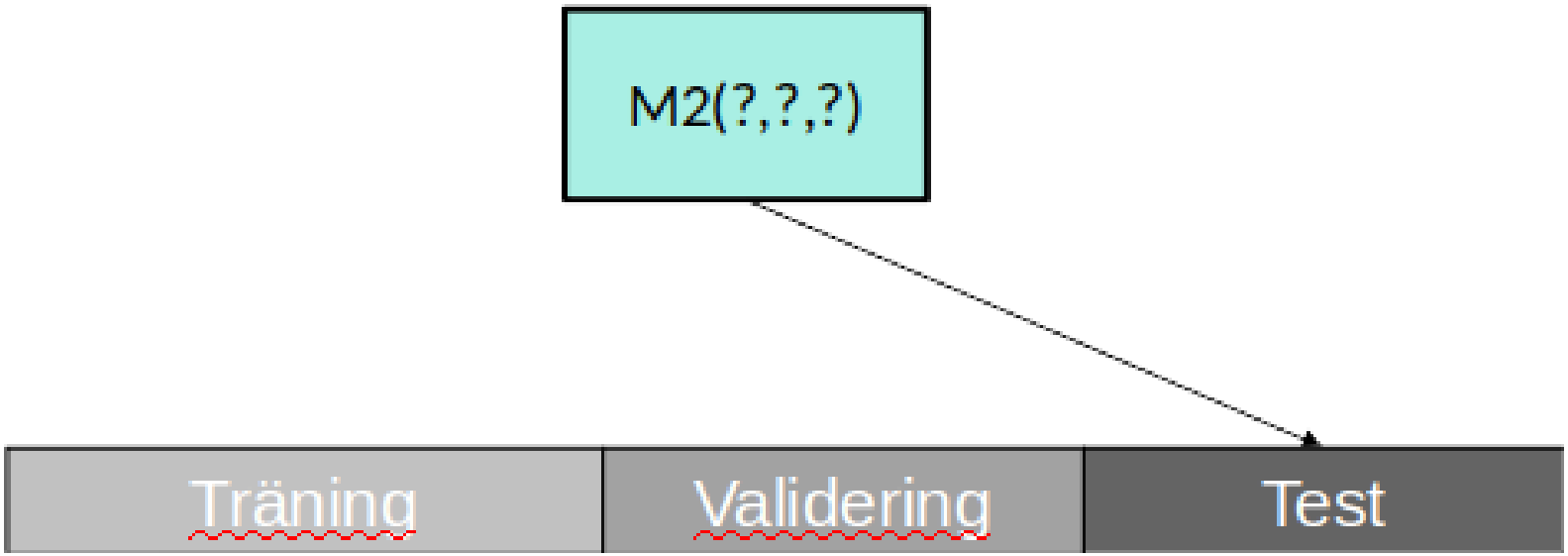




Träningsdata=svart, valideringdata=vit

Testmängden

- Används för att få en väntevärdesriktig skattning av felfunktionen på ny data.
- Vi bör **inte** ändra något på modellen när vi ska använda den på testdata. Varför?



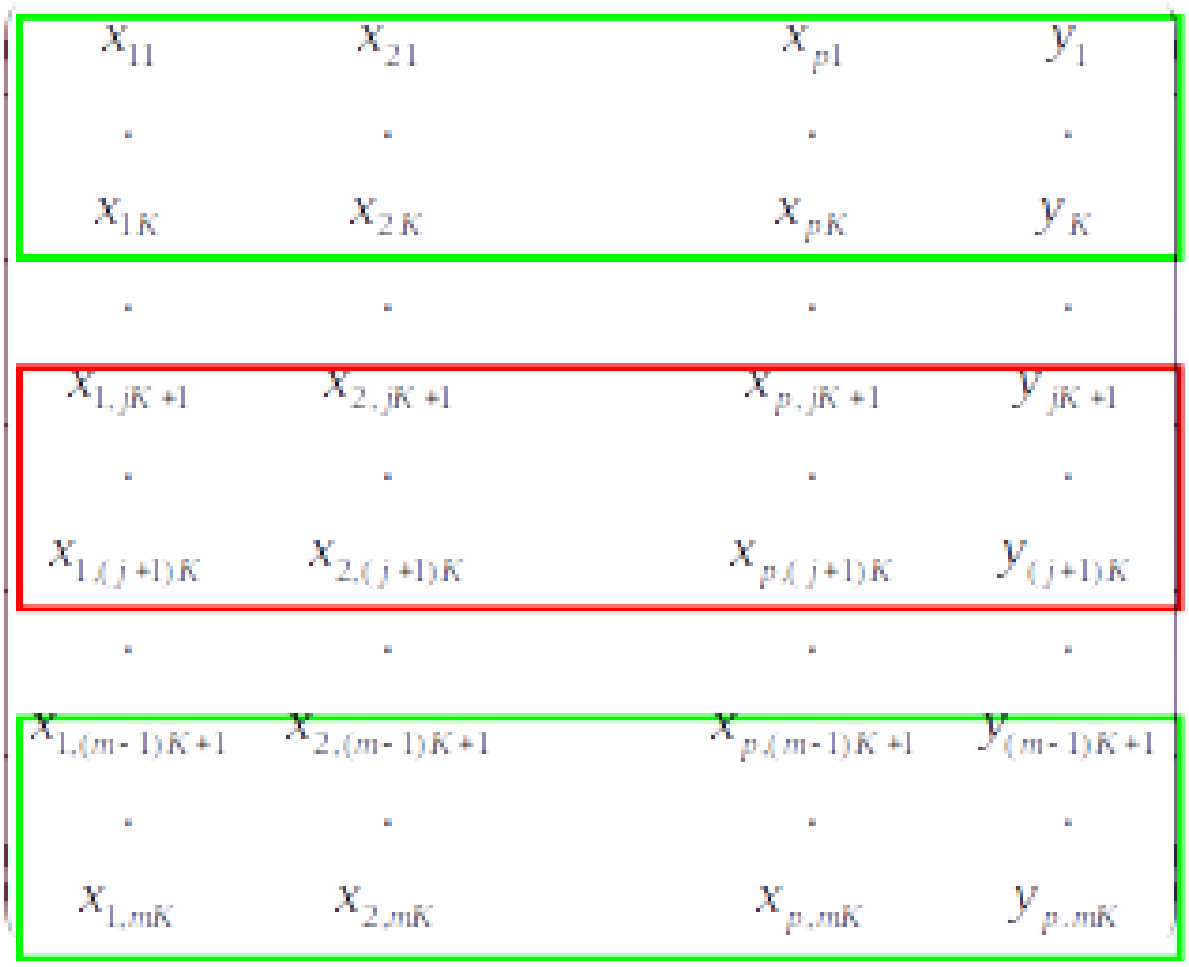
K-fold cross-validation

Utgå från ett antal kandidatmodeller.
Dela upp data i K block och för varje modell:

1. Ta bort ett block och anpassa modellen till återstående data
2. Använd den anpassade modellen för att skatta felfunktionen för de borttagna observationerna (valideringsdata)
3. Upprepa 1. och 2. för alla block och skatta felfunktionen för alla valideringsdata. Beräkna genomsnitt för felfunktionen över alla olika block.

Välj modellen med lägst genomsnittlig felfunktion.
Observera: Om vi har K block så måste vi skatta varje modell K gånger.

K-fold cross-validation



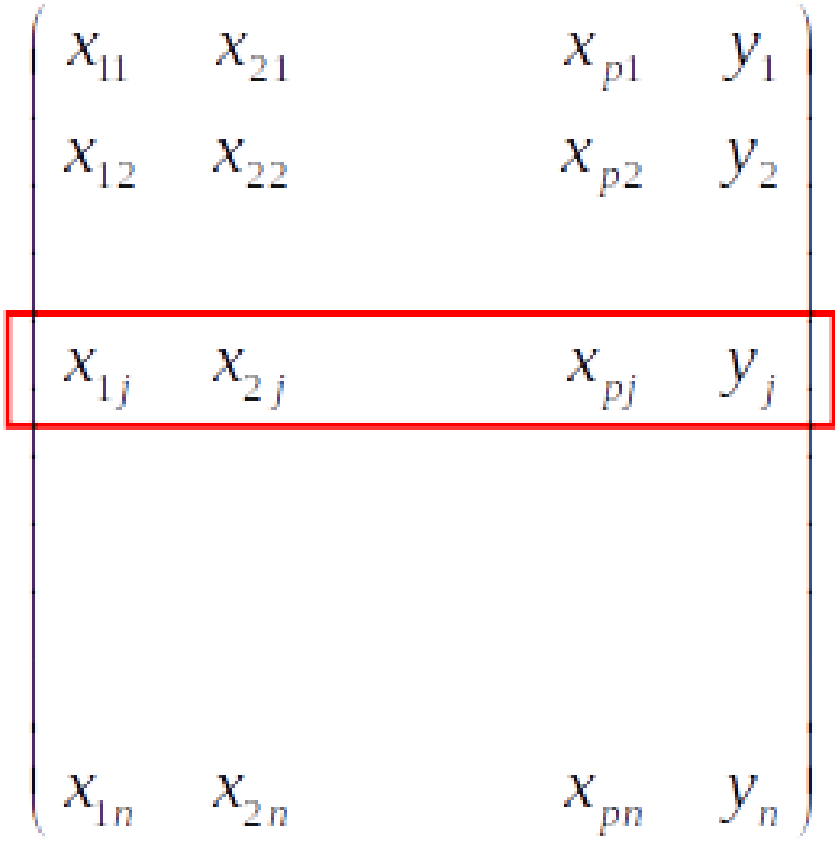
Leave-one-out cross-validation

Utgå från ett antal kandidatmodeller.
För varje modell:

1. Ta bort en observation och anpassa modellen till återstående observationer
2. Använd den anpassade modellen för att skatta felfunktionen för den borttagna observationen (valideringsdata)
3. Upprepa 1. och 2. för alla observationer och skatta felfunktionen för alla valideringsdata. Beräkna genomsnitt för felfunktionen över alla olika block.

Välj modellen med lägst genomsnittlig felfunktion.
Observera: Om vi har n obs så måste vi skatta varje modell n gånger.

Leave-one-out cross-validation



Bias, varians, brus

$$y = f(x) + \varepsilon \qquad E[\varepsilon] = 0 \qquad V[\varepsilon] = \sigma^2$$

$$\hat{y} = \hat{f}(x_{test})$$

Förväntad test MSE:

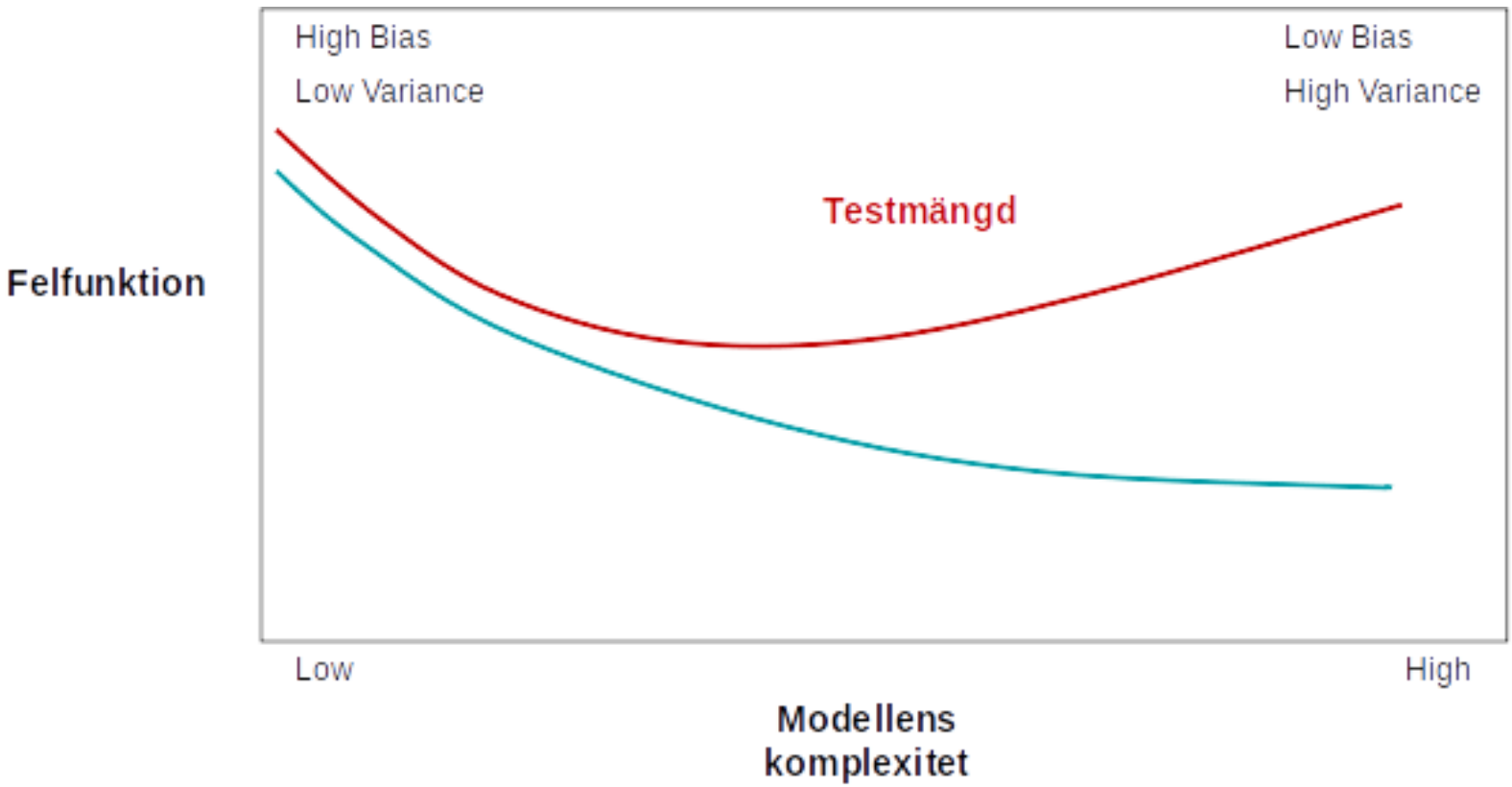
$$E\left[y_{test} - \hat{f}(x_{test})\right]^2 = V[\varepsilon] + V\left[\hat{f}(x_{test})\right] + Bias\left[\hat{f}(x_{test})\right]^2$$

- Brusvarians $V[\varepsilon]$: irreducibel brus
- Modellens varians $V\left[\hat{f}(x_{test})\right]$: Hur mycket kommer $\hat{f}$ att ändras när vi byter dataset
- Modellens skewhet $Bias\left[\hat{f}(x_{test})\right]$: Sytematisk skewhet eller modelleringsfel i modellen

Bias, varians, brus

Bias-variance-trade-off

- Vi fill ha
  - Lågt bias
  - Låg varians



Avslut

- Frågor? Kommentarer?
- Mail:
  - Teams
  - Distanskurs
- Kurshemsidan
- Labben