Prediktiv analys

FÖRELÄSNING 2

Dagens fråga

- Om inte blå, vilken färg tycker du himlen skulle ha?

Dagens agenda

- Forking GitHub Desktop
- Vad är prediktiv analys och hur gör man det?
- Supervised VS unsupervised learning
- Supervised learning regression och klassificering
- Modeller och algoritmer

Förra föreläsning

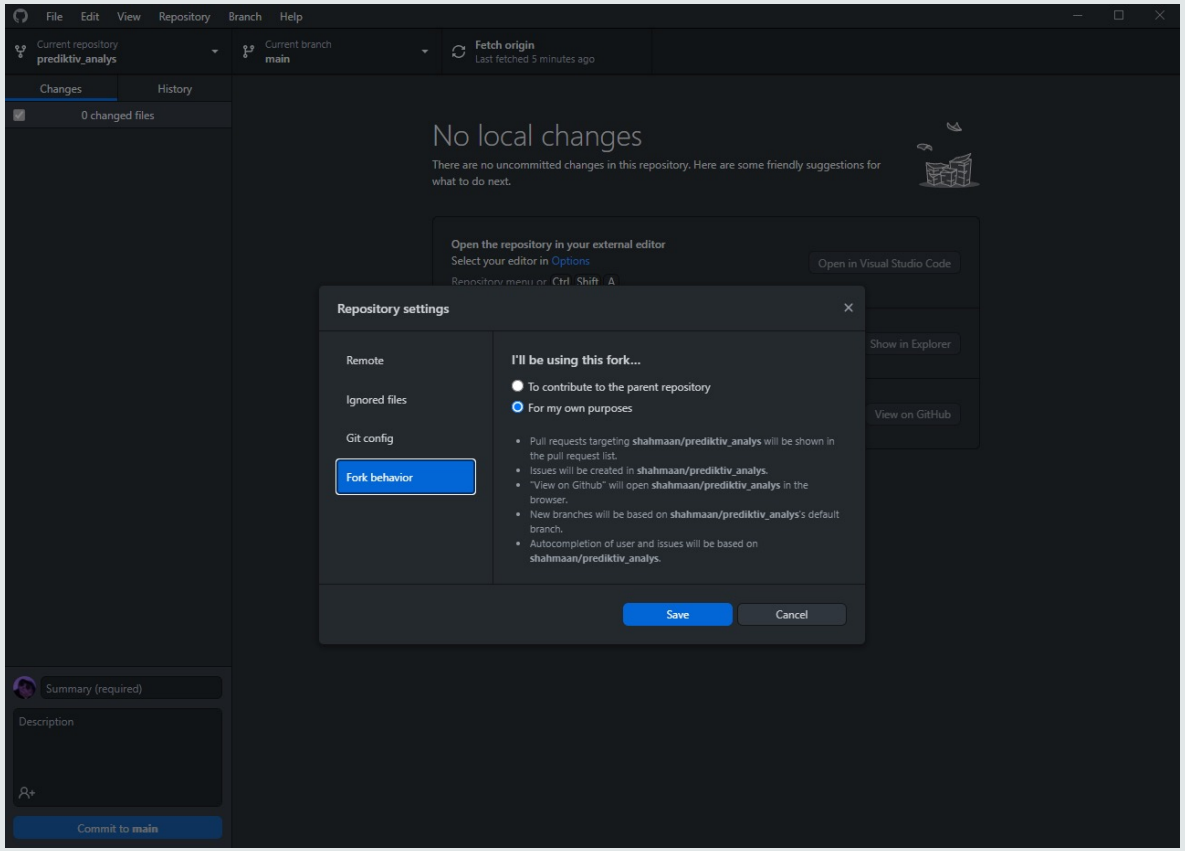
- Intro till kursen och kursplanering
- Etablering (repetition?) av koncept: prediktiv analys, algoritm, statistik, machine learning, neuralt närverk, data mining, business intelligence, regression, klassificering, datarensning
- Repetition installera Conda, VSC, Jupyter Notebook, Virtual Environment, GitHub Desktop och kursens repository

Forking Git

- Fork är en kopia av ett Git repo, men tillåter att du kan göra ändringar utan att påverka originalrepot
- Om ni önskar att jobba i kursens repository, göra ändringar, men också pusha ändringarna till er GitHub måste ni skapa en fork
- Man kan skapa en fork genom terminalen. Följ instrukser https://docs.github.com/en/get-started/quickstart/fork-a-repo

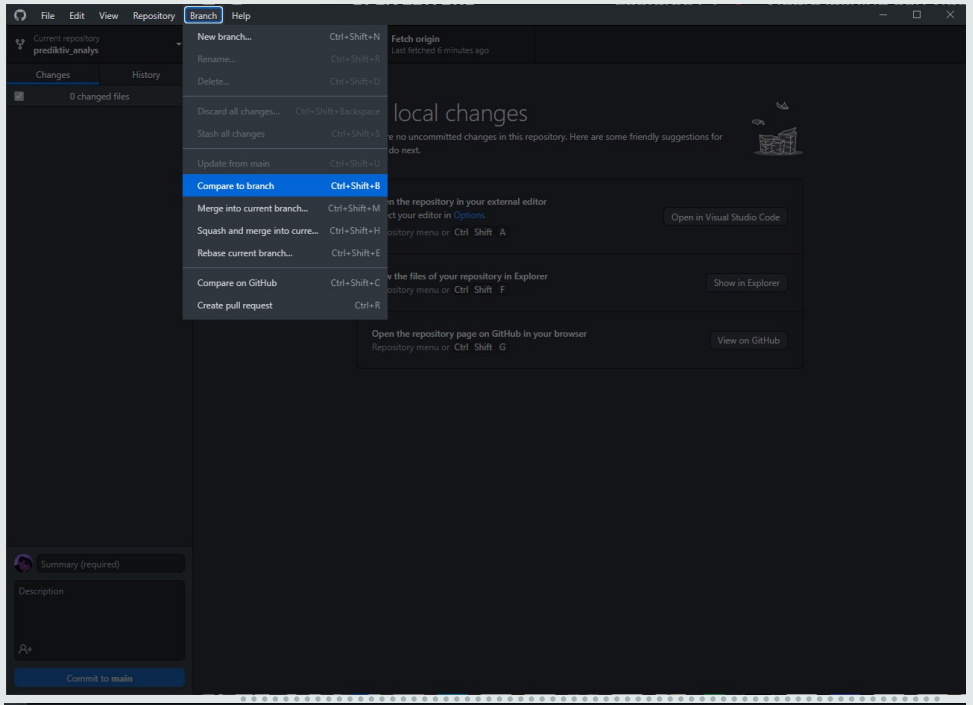
Forking GitHub Desktop

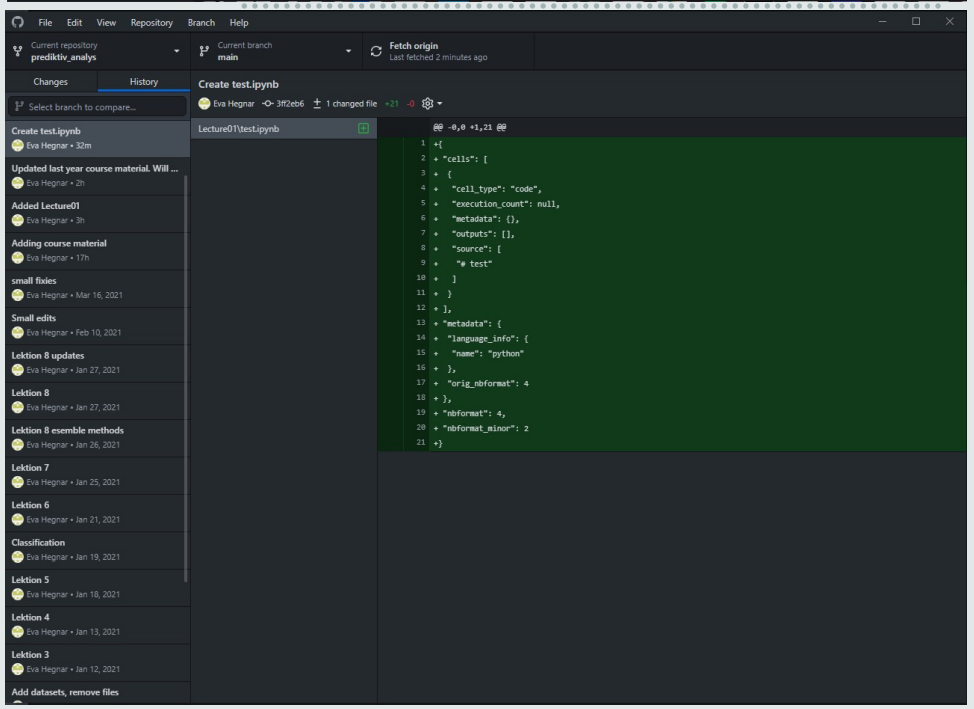
- Man kan göra det genom GitHub Desktop
- Gå till repot ni redan har klonat och vill forka i GitHub Desktop
- Välj Repository – Respository settings – Fork behavior – For my own purposes



Forking GitHub Desktop

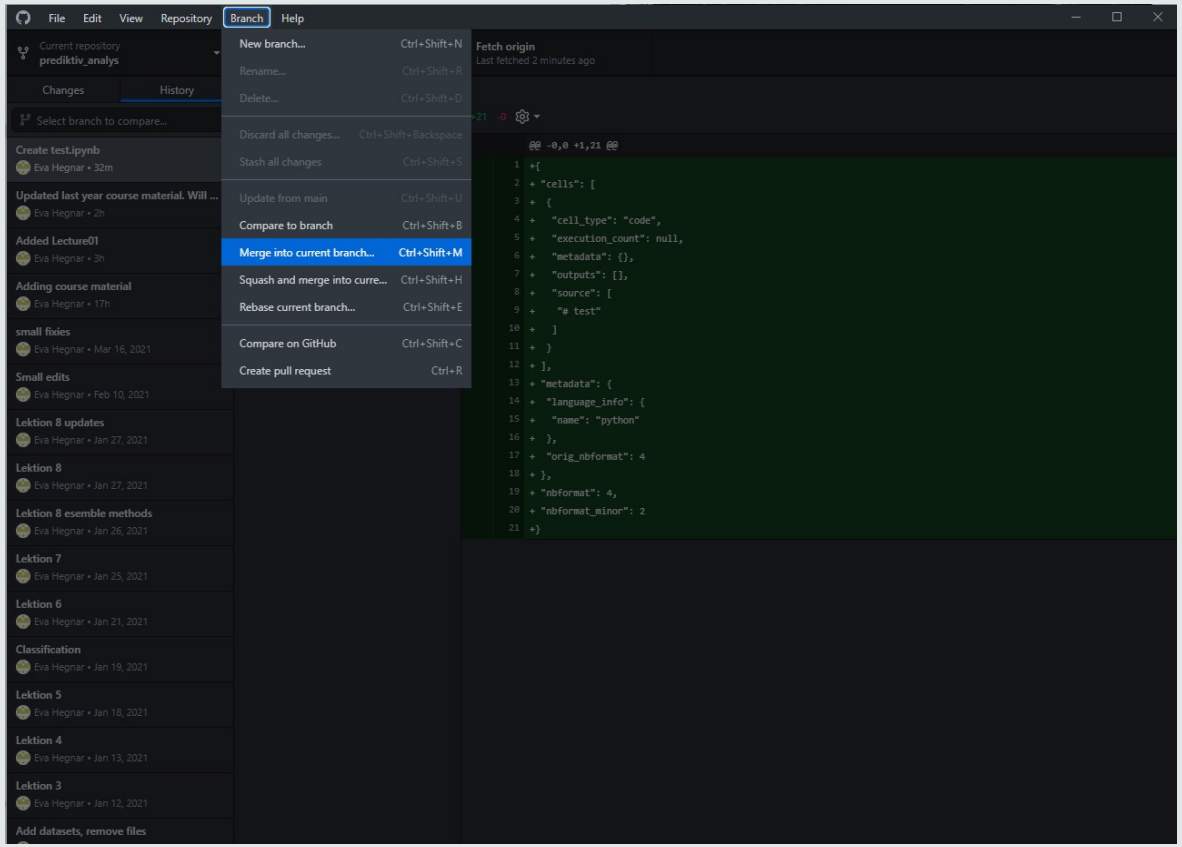
- Om du har gjort ändringar i repot och vill se de
- Välj Branch – Compare to branch





Forking GitHub Desktop

- För att merge ändringar man har gjort till forken, dvs spara sina ändringar till GitHub och få senaste ändringar i repot pulled
- Branch – Branch into current branch



Vad är prediktiv analys

- Med ”prediktiv” menar vi ”en gissning för något som är okänt
  - Medicinsk diagnos
  - Kommer denna användaren klicka på en annons
  - Elektrisk förbrukning per timme
  - Kommer denna studenten hoppa av utbildningen inom ett år
  - Konkurs

Finns många metoder att prediktera

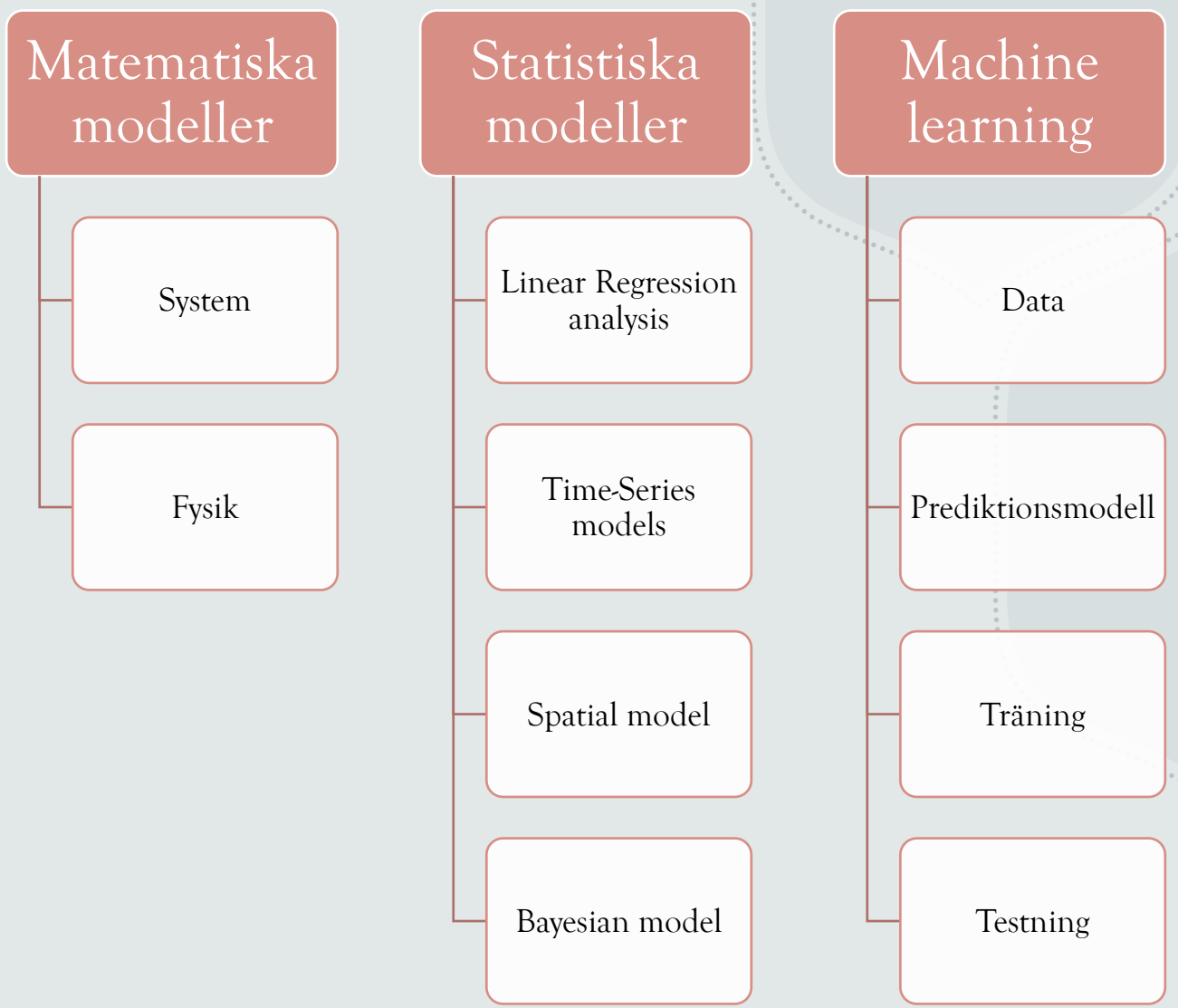
- Ej vetenskapliga metoder
  - Tur
  - Astrologi
  - Tarot kort
  - Magiska kristall kulor
  - Kommunikation med gud
  - Andra övernaturliga fenomen

- Vetenskapliga metoder
  - Matematiska lagar
  - Statistik

Prediktiv analys

- Prediktiv analys är användningen av **data**, kombinerat med tekniker från matematik, statistik och datavetenskap för att göra prediktioner.
- Målet med prediktiv analys är att producera en god approximering av vad som kan hända med okända händelser.

Hur gör man prediktiv analys

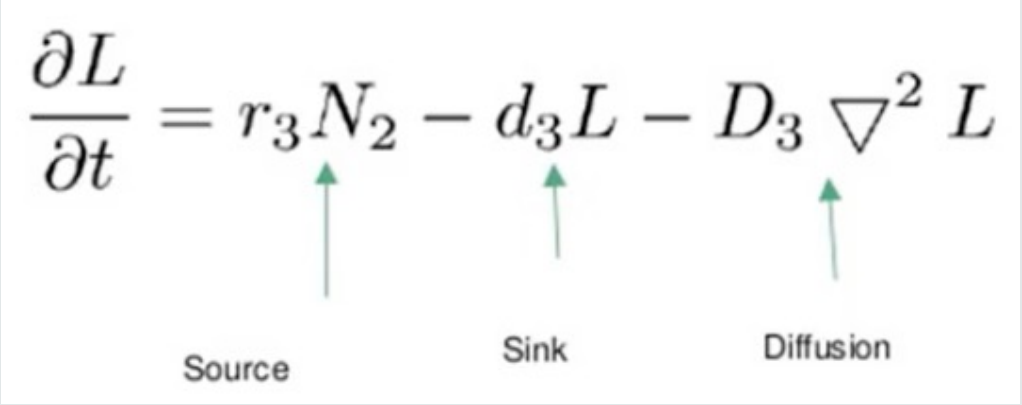


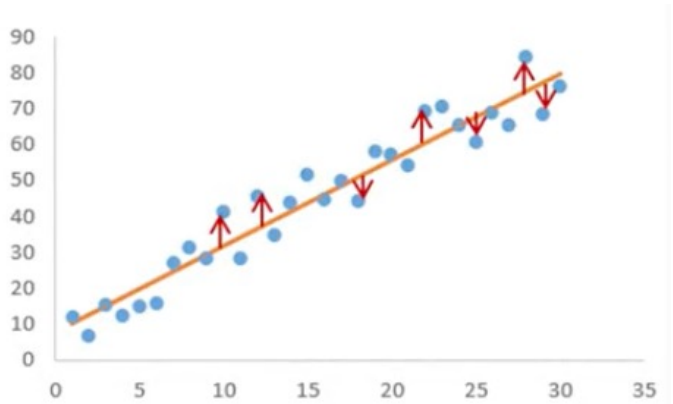
Matematiska modeller

En matematisk modell är en beskrivning av ett system där man använder matematiska koncept och språk.

Används mycket inom fysik

Oftast logiska derivat från existerande teorier



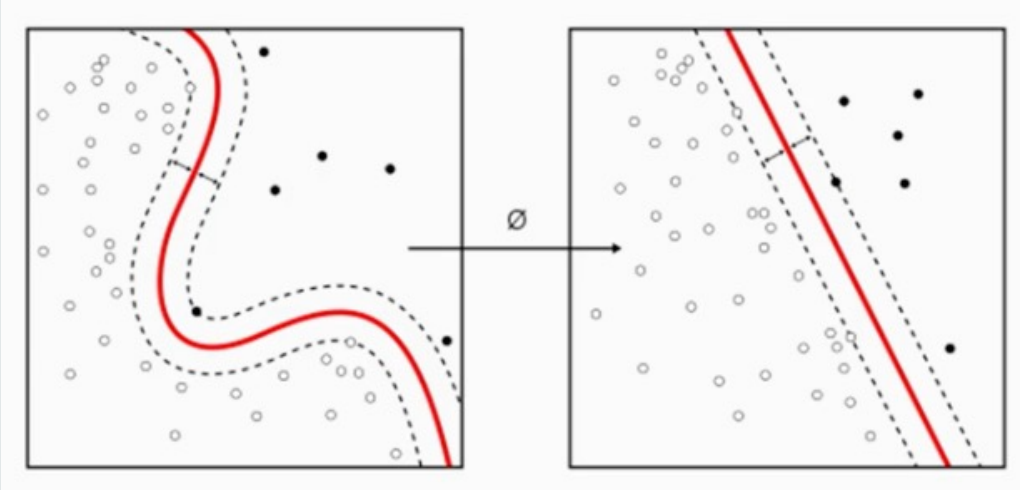


Statistiska modeller

- Statistiska modeller är en klass av matematiska modeller som försöker modellera system som har inslag av slumpmässighet
- Några exempel
  - Linjär regressionsanalys (Linear Regression analysis)
  - Tidsserie modeller (Time-Series models)
  - Rums modell (Spatial model)
  - Bayersk modell (Bayesian model)

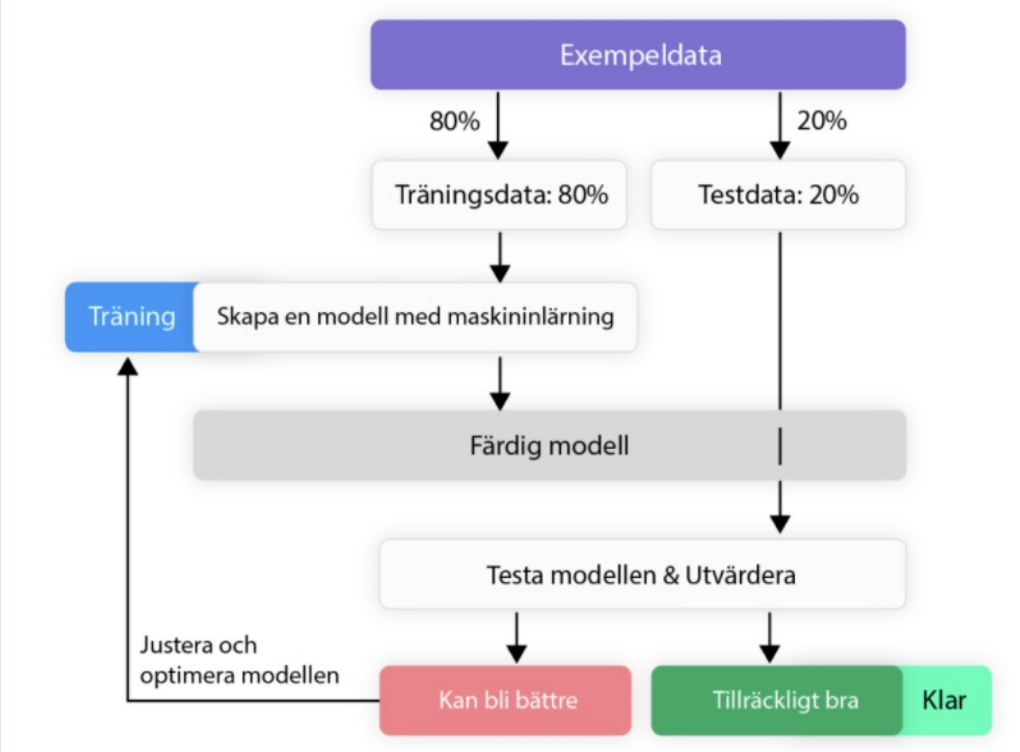
Machine Learning

- Machine Learning är en underkategori till datavetenskap.
- Generellt kan man enkelt beskriva fältet som:
  ”Man ger datorn möjlighet att lära sig utan att den blivit explicit programmerad”



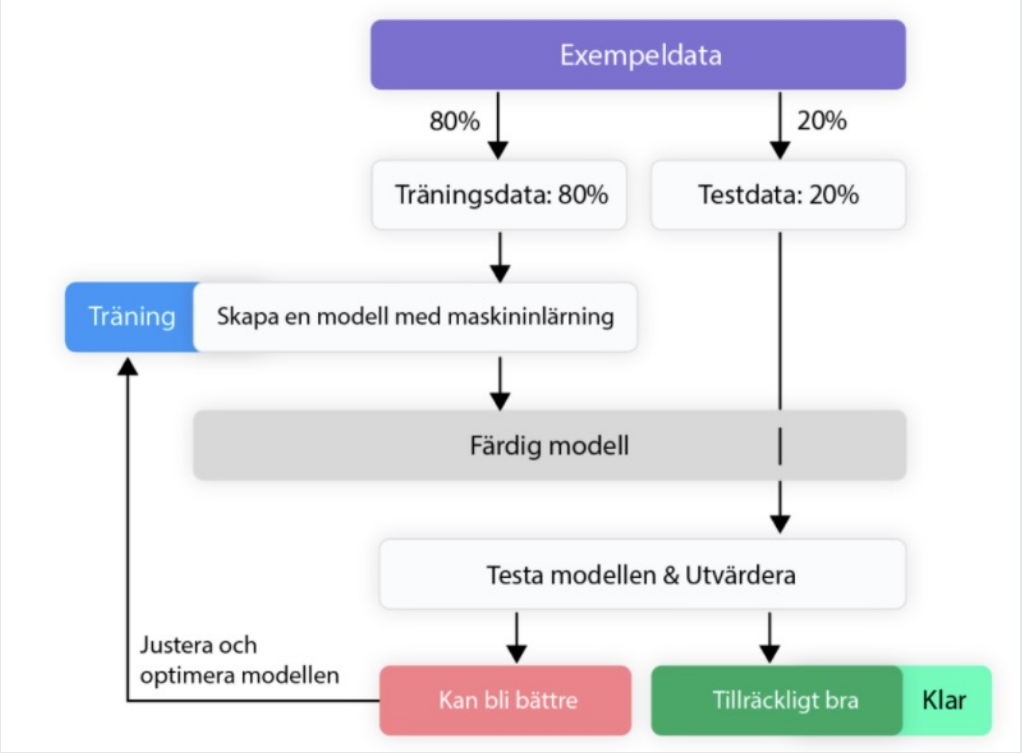
Data

- Data som man har till sitt förfogande.
- Data behöver inte vara i tabellform, kan vara bilder, videofilmer, diagram m.fl. är också data.



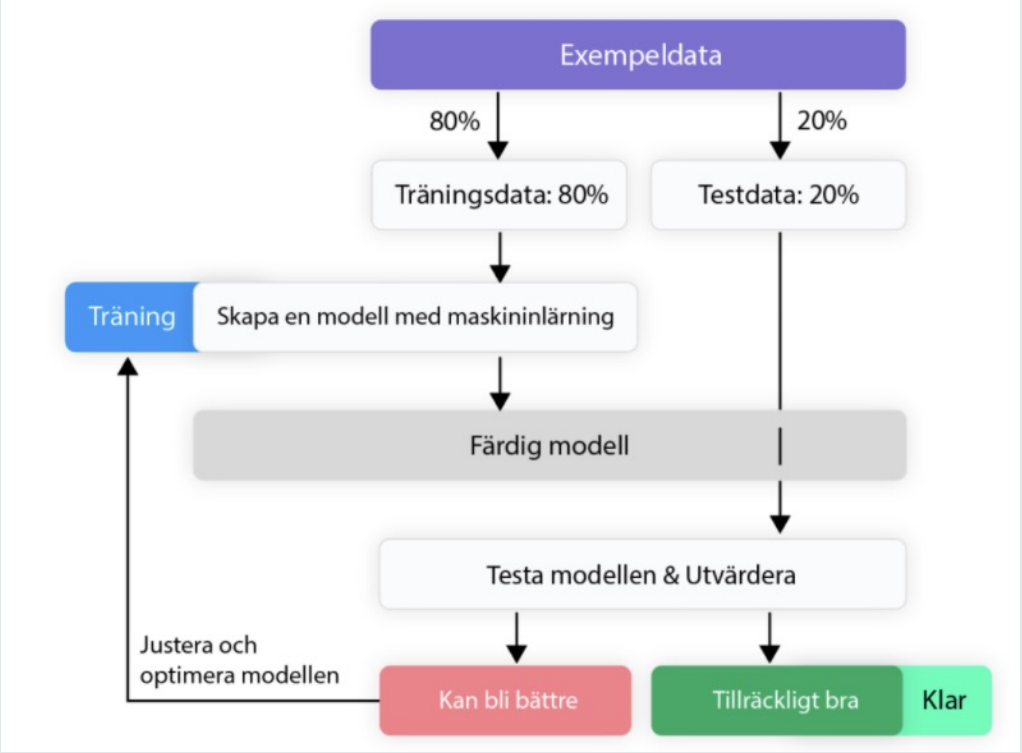
Modell

- När man studerar data så gör man det med matematiska funktioner och sammanfattningen av dessa funktioner kallas ***modell***.
- Man säger att man använder data för att bygga en modell.
- Denna modellen kan sedan användas för att exempelvis prediktera om nya patienter kommer utveckla cancer.
- Modellen sammanfattar vad datorn lärt sig med machine learning.



Träning (training)

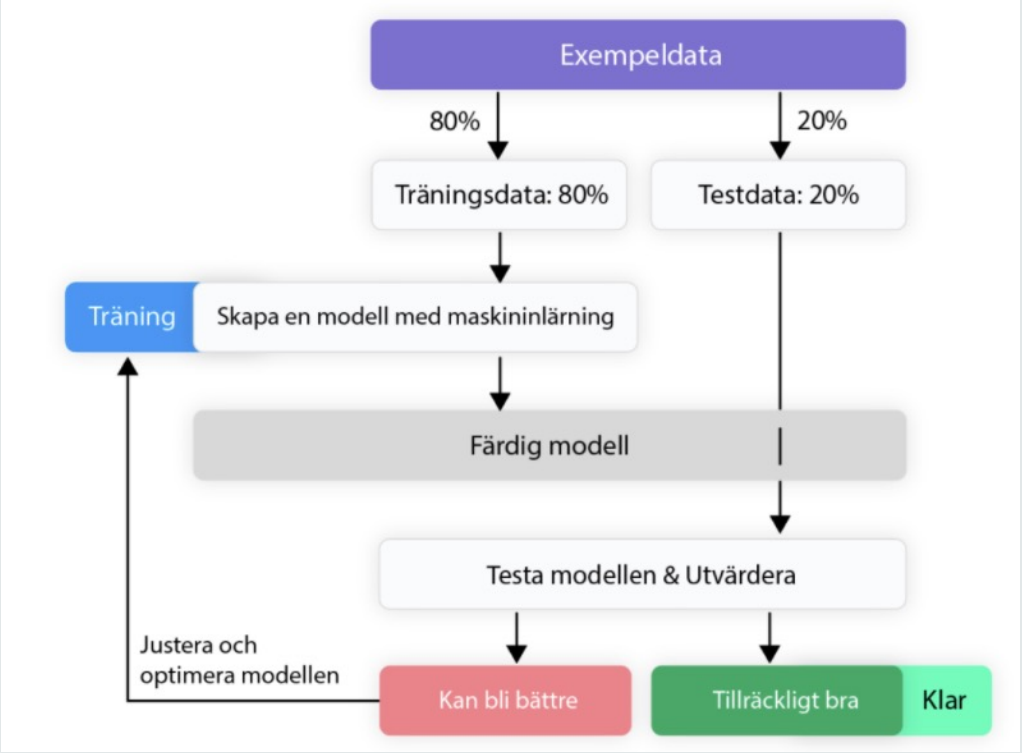
- För att skapa en modell behövs data.
- Data (träningsdata) används för att maskininlärningen skall hitta mönstren som finns i data.
- Denna fasen kallas träning (eng. training), eftersom maskinen tränas med hjälp av exempeldata.
- Man brukar som regel använda 80% av all data till träning.



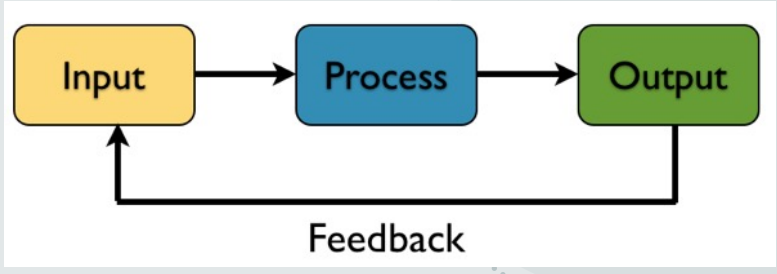
Testning

De 20% av data som inte används till träning kan användas till testning. Under testningen utvärderar man hur bra modellen är och detta måste göras på data som modellen inte tränats på. Testningen går alltså ut på att utvärdera modellens precision/förmåga.

Det är viktigt att man alltid testar modellen på data som modellen inte studerat under träningsfasen. Modellen kommer nämligen alltid ha hög precision på data som den ”sett under träningen”. Detta är vad man kallar ”overfitting”. Precisionen på testdata kommer alltid vara lägre och det är den precisionen som vi är intresserade av, eftersom det ger en bättre indikation av hur modellen presterar på framtida data. Fördelningen 80/20 kan justeras beroende på situation och data.



Input, output, features



- Variabel, input

  Med variabel menar man vanligtvis kolumnerna i en tabell eller attribut. Varje kolumn beskriver en egenskap hos enheterna man studerar (patienter, besökare, osv).

- Utfall, Output, Label, target

  Utfall är det man försöker förutsäga med hjälp av modellen. Utfallet är alltså det man är intresserad av att prediktera.

- Prediktorer, features, attributes

  Alla variabler som används för att förutsäga utfallet är features. Man kan ha ett dataset med många varaiabler, men de som väljs till modellen är features. Med andra ord är en feature en variabel som används för att prediktera (förutsäga) utfallet.

Supervised och unsupervised learning

Supervised Learning

Unsupervised Learning

Reinforced Learning

Supervised Learning

- All machine learning börjar med någon form av data
- Vi har ett urval eller observation av något
- För varje observation har vi ett set av egenskaper (features) (attribut, variabler) och en mål (target) variabel som vi vill förutspå.

LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	default payment next month
20000	2	2	1	24	2	2	-1	-1	1
120000	2	2	2	26	-1	2	0	0	1
90000	2	2	2	34	0	0	0	0	0
50000	2	2	1	37	0	0	0	0	0
50000	1	2	1	57	-1	0	-1	0	0
50000	1	1	2	37	0	0	0	0	0
500000	1	1	2	29	0	0	0	0	0
100000	2	2	2	23	0	-1	-1	0	0

Mer exempel

Vi har data om (features)...	... och vill prediktera (target):
E-mails	Hur mycket är spam resp. inte spam
Finansiell data	Aktiepriset
Akademiska och socioekonomiska data	Vilka kommer hoppa av
Nyhetsartiklar	Hur många kommer se dem
Samt mycket mera	

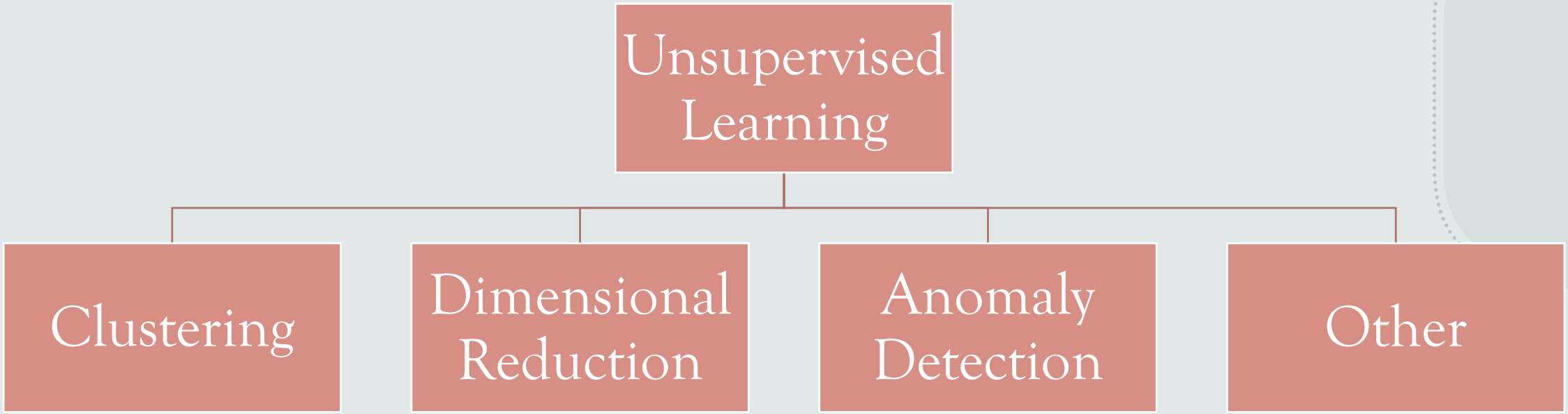
Unsupervised Learning

- Träningsdatan består av ett set av egenskaper (features) (attribut, variabler) utan några relaterade mål (target) variabler
- Vi har alltså bara features!

LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4
20000	2	2	1	24	2	2	-1	-1
120000	2	2	2	26	-1	2	0	0
90000	2	2	2	34	0	0	0	0
50000	2	2	1	37	0	0	0	0
50000	1	2	1	57	-1	0	-1	0
50000	1	1	2	37	0	0	0	0
500000	1	1	2	29	0	0	0	0
100000	2	2	2	23	0	-1	-1	0

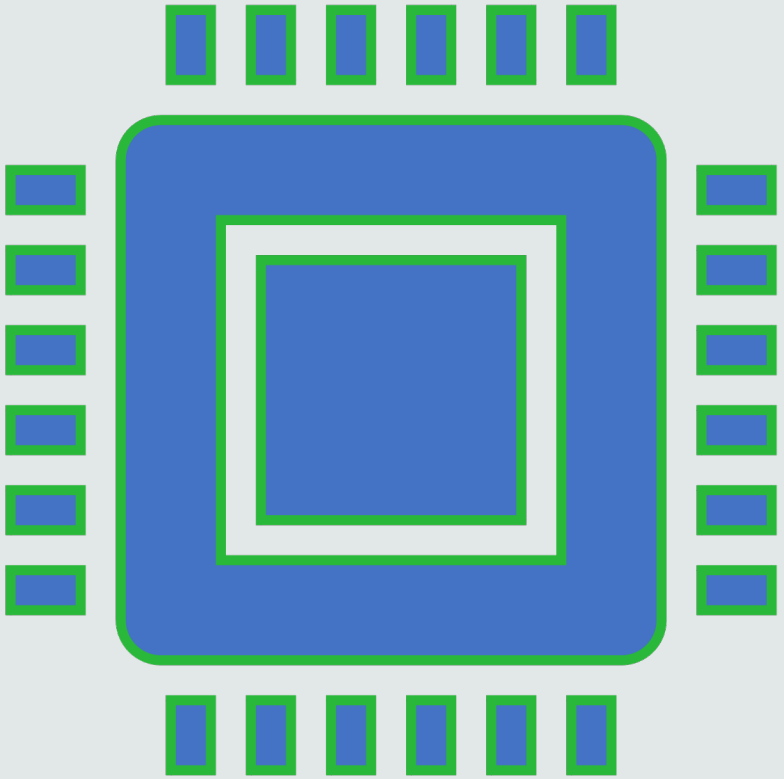
Brett urval

- Det finns många områden inom unsupervised learning
- De flesta försöker hitta någon form av struktur i datan



Mer exempel

Vi har data om (features)…	… och vill finna:
Kunder	Kundsegment
Kreditkorts transaktioner	Konsumtionsmönster
Betyg och socioekonomiska data om studenter	Gruppera egenskaper baserat på hur studenter kommer prestera baserat på data om föräldrarna
Genetiska data	Grupper av gener relaterade till någon biologisk funktion
Samt mycket mera	



Reinforcement Learning

- En form av maskininlärning där algoritmen interagerar med en dynamisk miljö där den måste prestera en viss funktion eller mål
- Programmet får feedback i form av ”morot eller piska” när det lär sig om problemområdet
- Exempel:
  - Tillverkningsrobotar
  - Självkörande bilar
  - Automatisk ”tradingmjukvara”

Supervised Learning

Regressionsproblem

Klassificeringsproblem

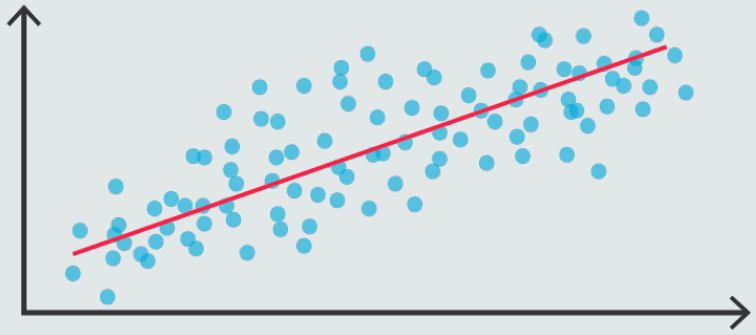
Krav på vår data

- Egenskaper (features) (attribut, variabler) från observationer
- Mål (target) variabel som vi vill förutspå.

LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	default payment next month
20000	2	2	1	24	2	2	-1	-1	1
120000	2	2	2	26	-1	2	0	0	1
90000	2	2	2	34	0	0	0	0	0
50000	2	2	1	37	0	0	0	0	0
50000	1	2	1	57	-1	0	-1	0	0
50000	1	1	2	37	0	0	0	0	0
500000	1	1	2	29	0	0	0	0	0
100000	2	2	2	23	0	-1	-1	0	0

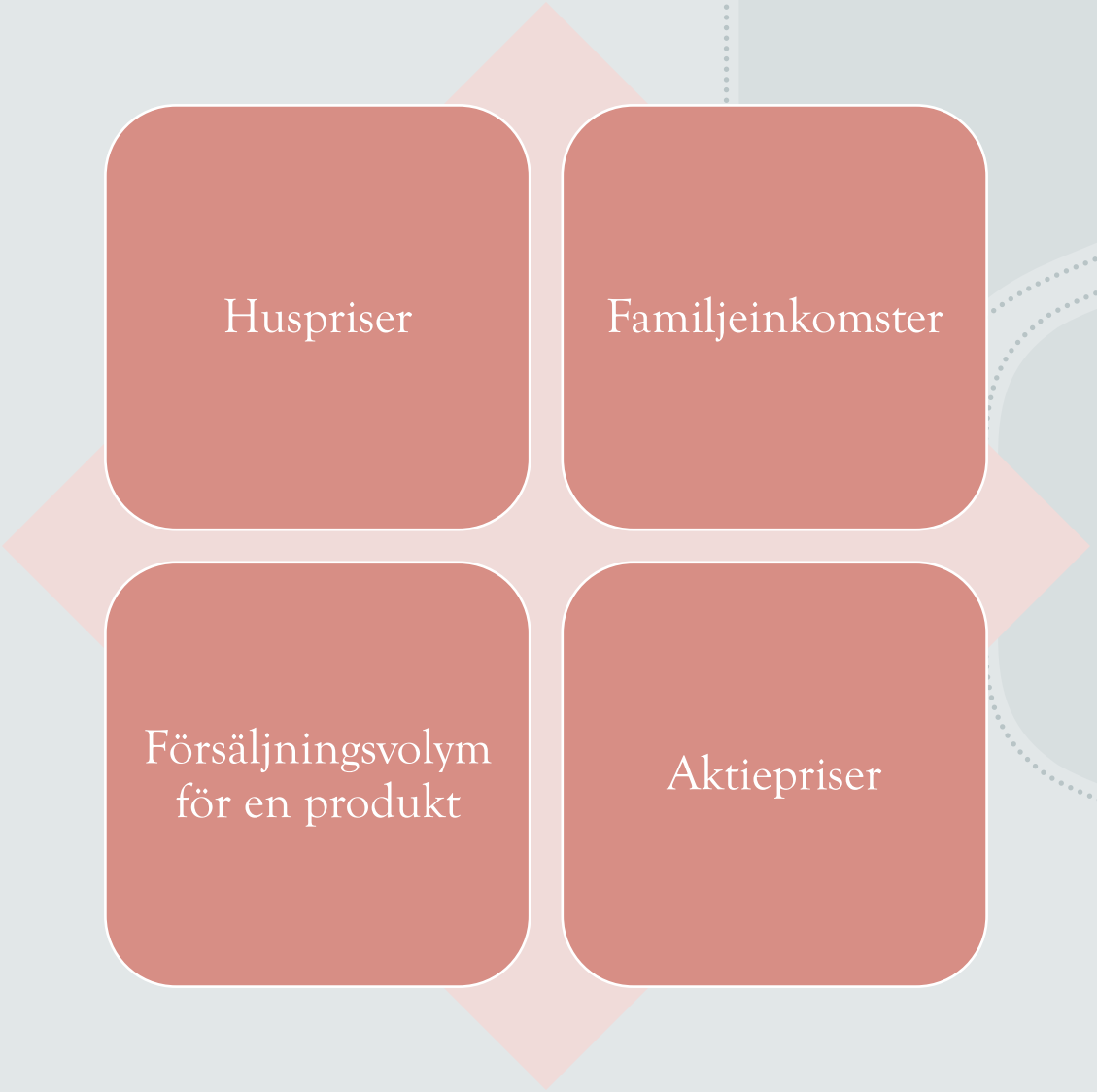
Regression

- När målet (target) (produkten, beroende värdet, kvantiteten som ska predikteras) är en numeriskt, **kontinuerlig**, variabel.



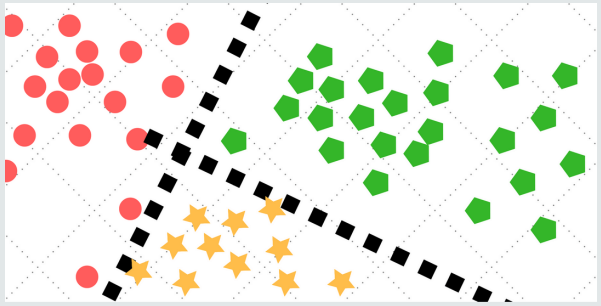
Dvs - Prediktera ett numeriskt värde

Exempel på regressionsproblem



Klassifikation

- När målet (taget) (produkten, beroende värdet) är en kategorisk variabel.



- Prediktera kategorier.
- Dvs – Icke numeriska variabler (fast kan representeras med tal)

Exempel på klassifikationsproblem

Hur gör vi regression och klassificering?

För båda typer av uppgifter har vi många modeller

Regression	Klassifikation
Linear Regression	Logistic Regression
K-Nearest Neighbour Regression	K-Nearest Neighbour Classifier
Lasso and Ridge Regression	Classification Trees
Random Forest	Random Forest Classifier
Artificial Neural Networks	Artificial Neural Networks
M.fl	M.fl

Vad behöver vi för att göra Prediktiv Analys med Supervised Learning

LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	default payment next month
20000	2	2	1	24	2	2	-1	-1	1
120000	2	2	2	26	-1	2	0	0	1
90000	2	2	2	34	0	0	0	0	0
50000	2	2	1	37	0	0	0	0	0
50000	1	2	1	57	-1	0	-1	0	0
50000	1	1	2	37	0	0	0	0	0
500000	1	1	2	29	0	0	0	0	0
100000	2	2	2	23	0	-1	-1	0	0

Ingående funktioner (features)

- Attribut
- Variabler

Råmaterialet som används för att prediktera något

Målet (target)

Det vi är intresserade av att prediktera

Målet är att använda en **inlärningsalgoritm** som använder **ingående data** och **målet(target)** för att producera en **modell.** Den ger oss ett förhållande mellan ingående data och target.

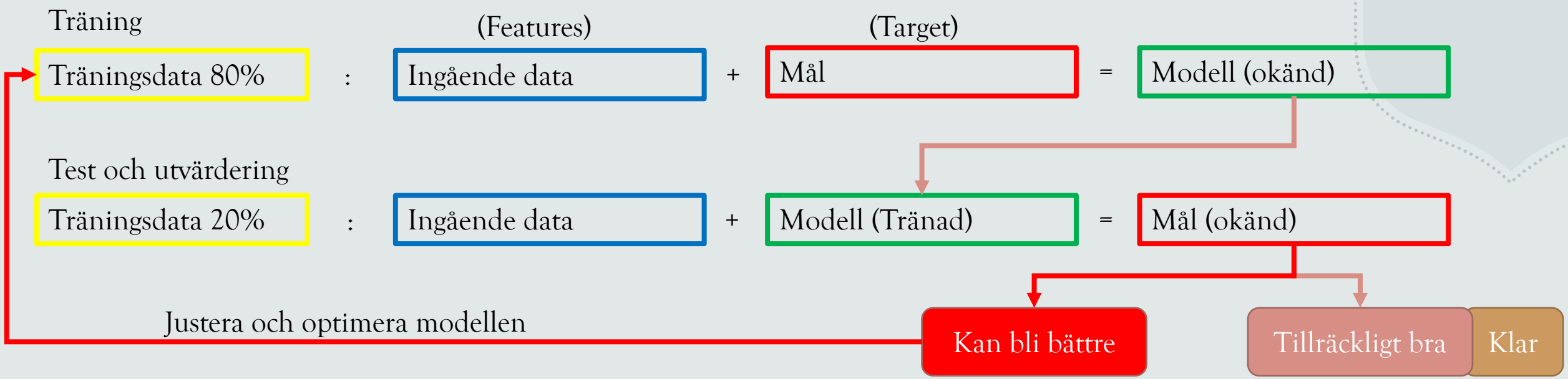
Inlärningsalgoritmen ska därefter kunna användas för att prediktera ett okänt taget från givna ingående värden.

Förenklat

Målet är att använda en **inlärningsalgoritm** som:

- använder **ingående data** och **target** för att producera en **modell**
- ger oss ett förhållande mellan ingående data och target

Inlärningsalgoritmen ska därefter kunna användas för att prediktera ett okänt mål från givna ingående värden.



Så vad är en inlärningsmodell?

Learning Model = *Model* + *Algorithm*

Modell: Den generella formuleringen av förhållandet mellan ”features” och ”target”

Learning Algorithm: Tillvägagångssättet för att finna den **specifika formen** för *Modellen*, vanligtvis genom inlärning av parametrar från data

Model	Learning Algorithm
Linear Regression	Ordinary Least Squares Method
Logistic Regression	Gradient descent
Artificial Neural Networks	Back propagation
Support Vector Machines	Quadratic programming
Perceptron	Perceptron learning algorithm

Exempel 1 – ”The Perceptron Model”

Model 1: Standard för kredit (ja/nej) (Credit Default)

if CD = $((w_1 * Pay\_0) + (w_2 * Pay\_1)) > w_0$:

Predict Yes

else:

Predict No

w = weight

features

”The Perceptron Learning Algorithm”: kommer lära sig de lämpliga w’s (weight’s) för modellen.

Learning Model 1 = Perceptron + Perceptron Learning Algorithm

(Model) (Learning Algorithm)

Exempel 2 – ”Logistic Regression Model”

Model 2: Standard för kredit (ja/nej) (Credit Default)

$$Z = ((w_1 * Pay\_0) + (w_2 * Pay\_1) + (w_3 * Age))$$

if $CD = \frac{1}{1 + \exp^{-Z}} > 0.5$:

Predict Yes

else:

Predict No

w = weight

features

”The Gradient Decent Algorithm”: kommer lära sig de lämpliga w’s (weight’s) för modellen.

Learning Model 2 = Logistic Regression + Gradient Descent

(Model) (Learning Algorithm)

När ska man använda ”Supervised Machine Learning” för att göra Prediktiv Analys?

Man måste ha följande 3 element:

1. Det får inte finnas något färdigt känt förhållande mellan ”features” och ”target”
2. Det finns ett mönster (eller förhållande) mellan ”features” och ”target”
3. Tillräcklig kvantitet med ”data” och av tillräcklig kvalitet (skräp in = skräp ut)

Övning

- Gör dagens tillhörande frågor (separat dokument)
- Vidare läsning

Vad har vi gått igenom idag?

- **Vad är prediktiv analys**

  Med ”prediktiv” menar vi ”en gissning för något som är okänt

- **Hur gör man prediktiv analys**

  Matematiska modeller

  Statistiska modeller

  Machine Learning

- **Supervised vs Unsupervised learning**

  Supervised - För varje observation har vi ett set av egenskaper (features) (attribut, variabler) och en mål (target) variabel som vi vill förutspå.

  Unsupervised - Träningsdatan består av ett set av egenskaper (features) (attribut, variabler) utan några relaterade mål (target) variabler

  Reinforcement - En form av maskininlärning där algoritmen interagerar med en dynamisk miljö där den måste prestera en vis funktion eller mål

- **Supervised Learning – Regression och klassificering**

  Regression - När målet (kvantiteten som ska predikteras) är en numeriskt (kontinuerlig) variabel. Dvs - Prediktera ett numeriskt värde.

  Klassificering - När målet (beroende värdet) är en kategori variabel. Dvs – Icke numeriska variabler.

- **Modeller och algorithmer**

  Learning Model = *Model + Algorithm*

  ”The Perceptron Model”

  ”Logistic Regression Model”

Nästa lektion

- Regression
- Stegen för att bygga en Machine learning algoritm i Python.
- Exempel på hur man genomför regression i Python.