Prediktiv analys

FÖRELÄSNING 12

Dagens fråga

- Vad är otroligt billigt, och som du skulle betala betydligt mer för?

Dagens agenda

- Repetition
- Jobba med inlämning

Förra föreläsning

- Optimering av prediktionsmodeller
- Evaluering av modell: K-fold Cross Validation
- Hyperparameter tuning – hur välja bäst parameter till modellen?
- Exhaustive Grid Search
- Feature Selection

Mentimeter

Repetition

Koncept i prediktiv analys

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

Matematisk modell - beskrivning av ett system där man använder matematiska funktioner.

Statistiska modeller - klass av matematiska modeller som försöker modellera system som har inslag av slumpmässighet

Machine Learning - datorn lär sig reglar utan att den blivit explicit programmerad

Träning – Modellen tränas med data för att lära strukturen den behöver för att prediktera korrekt

Testning – Avsatt data testar hur bra modellen faktiskt fungerar

Koncept i prediktiv analys

Attribut – variabler, kolumner i ett dataset, beskriver en egenskap hos enheten man studerar

Features – input, alla attribut som används i modellen för att prediktera

Target – output, label, det man försöker prediktera med modellen

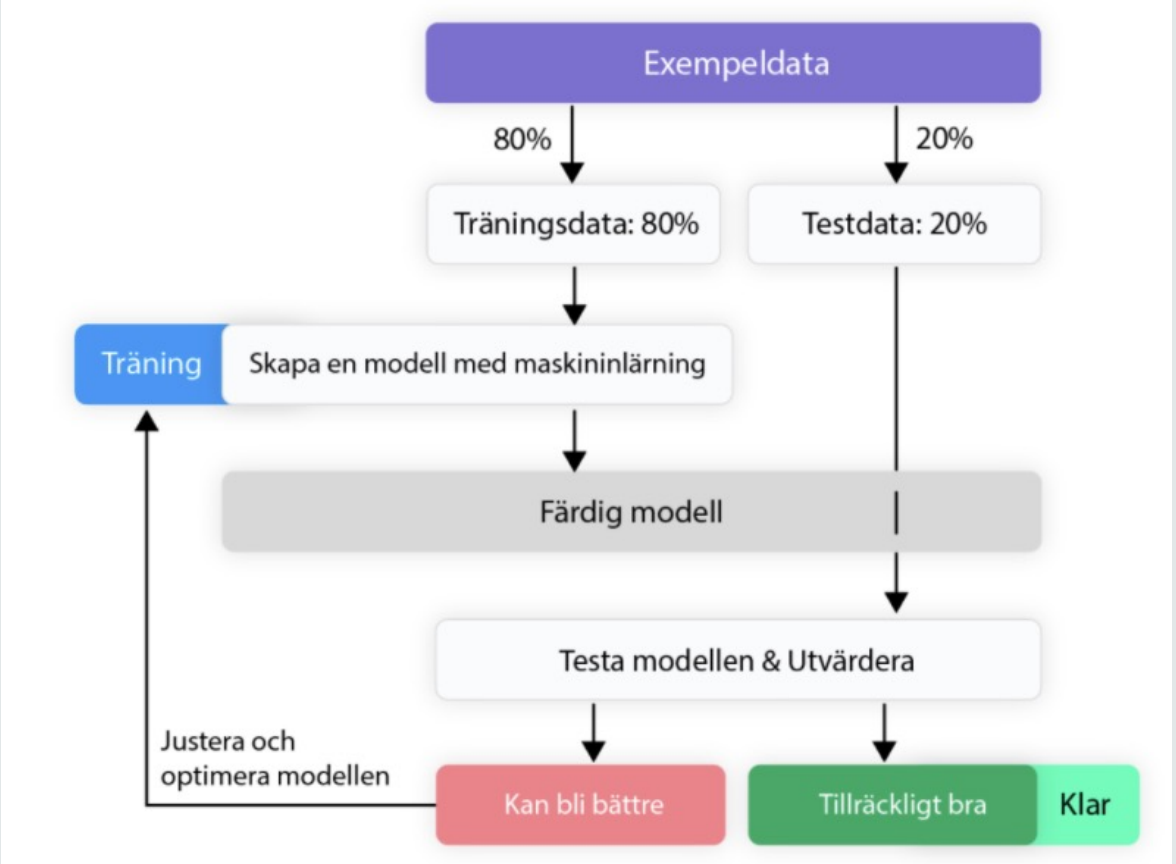
Supervised Learning - För varje observation finns ett sätt features och ett target man vill prediktera

Unsupervised Learning - För varje observation finns ett sätt features, men ingen target

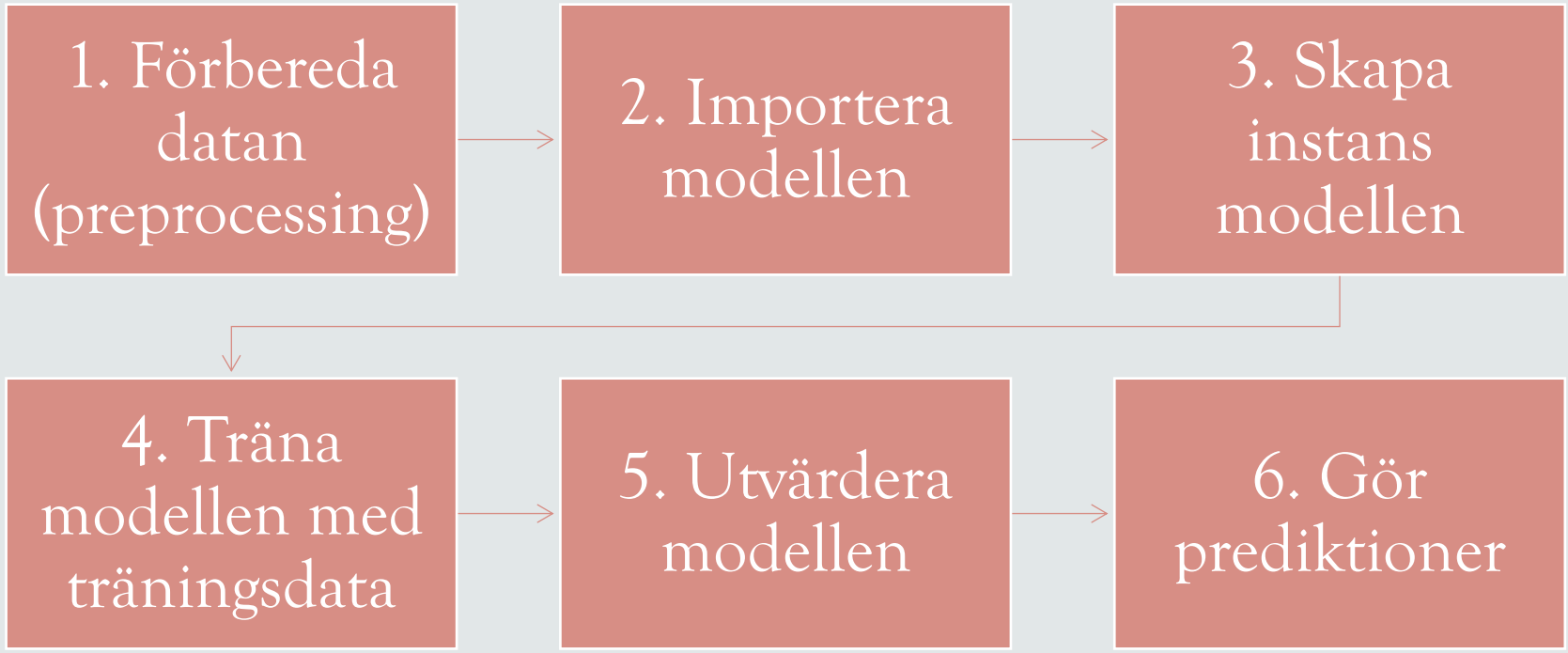
Modell - Den generella formuleringen av förhållandet mellan features och target

Learning Algorithm - Tillvägagångssättet för att finna den specifika formen för modellen genom inlärning av parametrar från data

Koncept i prediktiv analys



Steg i prediktiv analys



Regression

Regression – När target är en numeriskt (kontinuerlig) variabel

Multiple Linear Regression - Beskriver relation mellan variabler mellan att anpassa en linje till den observerade datan

K-Nearest Neighbor - Observationer med features som ligger nära varandra kommer att ha target värden som också ligger nära varandra

Lasso Regression – Modifireat Multiple Linear Regression som exkluderar egenskaper som är irrelevanta för modellen

Model evaluation – Regression

Error metrics – Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R-squared

Önskar dessa error metrics för regressionproblem så små som möjligt. R2 nära 0

Holdout cross validation – random uppdelning av datan i tränings- och testdataset

Overfitting - Modellen lär sig alla aspekter av träningsdatan inklusive slumpmässiga avvikelser Som konsekvens av detta så kommer modellen ge dåliga predikteringar för osedd test data. Kan hända i för komplexa modeller

Regularization – teknik för att förhindra overfitting som straffar för komplexa modeller

Klassificering

Klassifikation - När target är en kategorisk (diskret) variabel

Binär klassifikation – när det är två kategorier 1 (positiva) klassen och 0 (negativa) klassen

Logistic Regression – Beskriver relation mellan variabler mellan att anpassa en linje till den observerade datan när det är kategorisk target

Classification Trees – Lär sig simpla beslutsregler som delar datan baserad på features (kan också ha Regression Tree)

Naive Bayes – Baserad sannolikhetsteori och på Bayes theorem. Gaussian NB och kontinuerliga features och Multinomial NB om diskreta features

Model evaluation – Klassificering

- **Confusion matrix** – beskriver prestanda för binär klassifikation
- **Accuracy** = (TP +TN) / Total
- **Precision** = TP / (TP + FP)
- **Recall** = TP / (TP + FN)
- Ger värde mellan 0 och 1 och vill ha error metrics så nära 1 som möjligt

- **Threshold** – Det är minsta sannolikhet som klassificerar en observation i den positiva klassen (1). Default threshold är 0.5 (50%)

		Predicted	
		0	1
Observed	0	True Negatives ☺ (TN)	False Positives ☹ (FP)
	1	False Negatives ☹ (FN)	True Positives ☺ (TP)

Data preprocessing

01

Handling Missing values (NaN)

02

Feature scaling

03

Handling categorical features (Onehot endcoding/ Dummy variables)

04

Outliers

05

Multicollinearity

Optimering av prediktionsmodeller

Ensamble methods

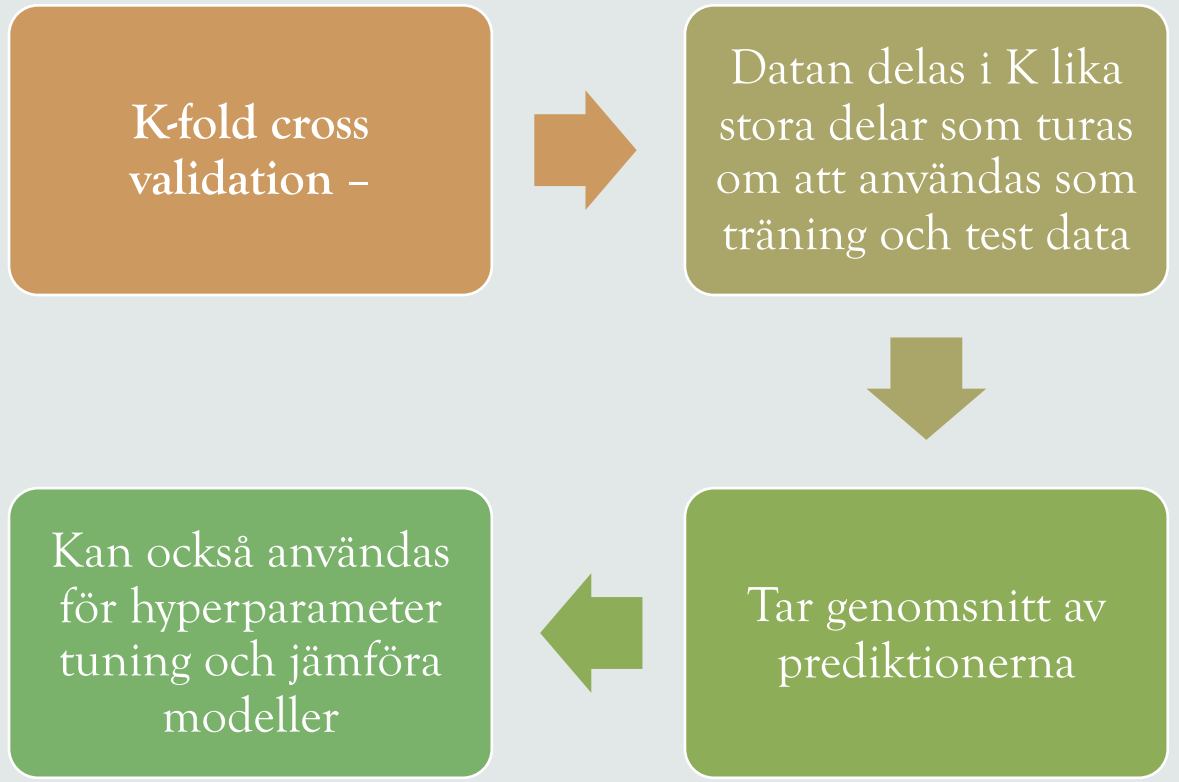
Bagging – bootstrap aggregation. Resamplar ett antal dataset från träningsdatan med replacement. Samma typ modell tränas på varje samplade dataset. Aggregerad prediktion: genomsnitt när regression och majority vote när klassificering

Random Forest – flera simpla träd tränas på boostrappad sampel från träningsdata. Aggregerad prediktion: genomsnitt när regression och majority vote när klassificering

AdaBoost – svaga modeller tränas i sekvens. Efter varje tränad modell re-weights träningsdatan till nästa modell. Fel prediktion viktas högre.

Optimering av prediktionsmodeller

Utvärdera modell



Optimering av prediktionsmodeller

Hyperparameter tuning

Hyperparameter – parameter som anges till modellen och inte lärs

Exhaustive Grid Search – testa att givet antal hyperparametrar från ett grid av parametervärden. Bästa kombination vinner

Använd K-fold cross validation

Optimering av prediktionsmodeller – Feature selection

- Ta bort dummy features med låg varians
- Ta bort features som är högt korrelerade med andra features (multicollinearity)
- Vilka features är högt korrelerade med target?
- Använda statistiska tester för att identifiera relevanta features, ANOVA och Chi-squared
- Recursive Feature Selection – modeller som beräknar koefficienter eller feature importance (MLR, logistic regression, random forest etc)