

# Classificatie – Voorspelling

Tot nu toe: vooral 'classificatie'

- Naive Bayes
- k-Nearest Neighbours
- . . .

Gebaseerd op voorspellingsvariabelen  $X_1, X_2, \dots, X_p$

Proberen om klasse  $Y$  te bepalen (= discreet)

- Training data: om model te bouwen
  - Validatie data: om de juistheid van het model te testen
- confusion matrix

Nu: ook '**voorspelling**'

- k-Nearest Neighbours
- Multiple Linear Regression
- . . .

Gebaseerd op voorspellingsvariabelen  $X_1, X_2, \dots, X_p$  probeert de **doorlopende waarde van variabele Y** te voorspellen.

- Training data: om model te bouwen
- Validatie data: om de juistheid van het model te testen

▮ **numerieke maatregelen!**

voor elke observatie  $i$ , voorspellings error (residu):

$$e_i = y_i - \hat{y}_i$$

met

- $y_i$ : de 'echte' waarde
- $\hat{y}_i$ : de voorspelde waarde (door het model)

## **Numerieke maatregelen voor de correctheid van de voorspellingsmodel**

- MAE/MAED (Mean Absolute Error/Deviation)

...

- Average Error

...

- MAPE (Mean Absolute Percentage Error)

...

- RMSE (Root Mean Squared Error)

...

- TSSE Total Sum of Squared Errors

...

## Simpele Lineaire Regressie

gebaseerd op 1 voorspellings waarde X, probeert de waarde van 1 doorlopende output waarde Y te voorspellen.

Theoretisch model (populatie):

$$Y = \beta_0 + \beta_1 \cdot X + \varepsilon$$

met  $\varepsilon$  "ruis", spreiding in Y.

Veronderstellingen:

- spreading in Y hetzelfde voor elke waarde van X ( $\varepsilon \sim N(0, \sigma^2)$ ) = 'homoscedasticiteit'
- voorspellings errors (residu's) onafhankelijk

Training data (vb. 60% van hele dataset)

- schat coëfficiënten  $\beta_0$  en  $\beta_1$   
□  $Y = b_0 + b_1 \cdot X$
- parameter  $\varepsilon$ , schat spreiding in Y  
□ 'Std. Dev. Schatting' in output

Hoe goed is het gevonden model?

- Hangt af van het doel!

- Doel analyse: beschrijving (typische statistiek)

- 'goedheid van pasvorm'

- berekeningen op training data!

- bereken  $R^2$  (or  $R$ )

- hoe dicht bij 1 (of -1), hoe beter

- vierkantswortel  $R$  in de output

- doel analyse: voorspelling (typische data mining)

- numerieke metingen om juistheid te voorspellen

(zie vorige slide)

- berekeningen op validatie data!

- validatie data scoren in output

- balans tussen te 2 vinden

- voor data mining: vooral voorspellende kracht belangrijk

Wat in het geval van een andere partitie?

- bijvoorbeeld een ratio van 60%-40%, maar andere seed

- bijvoorbeeld een ander ratio, maar dezelfde seed

- verschillende schattingen  $b_0$  en  $b_1$  voor  $\beta_0$  en  $\beta_1$

- in welke mate kan dit verschillen van de ene partitie naar de andere?

- schattingen voor de spreading van de parameters van de ene partitie naar de andere?

- 'Std. Error' in output met 'Coëfficient'

## Meerdere Lineaire Regressie

Gebaseerd op meerdere voorspel variabelen  $X_1, X_2, \dots, X_p$  die 1 doorlopende output waarde  $Y$  proberen te voorspellen.

Theoretisch model (populatie):

$$Y = \beta_0 + \beta_1 \cdot X_1 + \dots + \beta_p \cdot X_p + \varepsilon$$

met  $\varepsilon$  "ruis", spreiding in  $Y$ .

Veronderstellingen:

- spreiding in  $Y$  hetzelfde voor elke waarde van  $X$  ( $\varepsilon \sim N(0, \sigma^2)$ ) = 'homoscedasticiteit'
- voorspellings errors (residus) onafhankelijk

Training data (vb. 60% van hele dataset)

- schattings parameters  $\beta_0, \beta_1, \dots, \beta_p$

$$\hat{Y} = b_0 + b_1 \cdot X_1 + \dots + b_p \cdot X_p$$

- parameter  $\varepsilon$ , schat spreiding in  $Y$
- 'Std. Dev. Schatting' in output

Welke voorspellende variabele werkt 'effectief' mee aan de voorspelling?

□ Welke coëfficiënten  $b_0, b_1, \dots, b_p$  zijn aanzienlijk verschillend van 0?

Coëfficiënten in het model die verschillend zijn van 0 kunnen dat per ongeluk zijn.

(bijvoorbeeld door observaties (bij toeval) in the training data)

□ calculeer p-waarde

- p-waarde: kans om deze waarde bij toeval te vinden voor een coëfficient met het model in de training data als die coëfficient in populatie gelijk is aan nul.

□ lage p-waarde: coefficient is aanzienlijk verschillend van nul, draagt 'effectief' bij aan de voorspelling

- 'p-waarde in output met 'Coëfficient'



## **Optimaal nummer + keuze van voorspellingsvariabelen?**

- Te veel variabelen: kans op overfitting!
  - Misschien lage voorspellingskracht
- Neem bij voorkeur geen variabelen in die niet bijdragen aan de voorspelling.
  - Leidt tot een grotere spreading in de voorspellingen
- Voorkeur voor geen variabelen te verwijderen die een bijdrage leveren aan de voorspelling.
  - Leidt tot een hoger gemiddelde error in de voorspellingen
- Pas op voor de voorspellingswaarden die sterk gecorreleerd zijn!
  - Kan de coëfficiënten incorrect representeren
  - correlaties opsporen ('matrixplot' of 'correlatiematrix')
- Pas op voor uitschieters!
- Vuistregel: aantal observaties  $n$  in training data is gelijk aan minstens  $5 \cdot (p+2)$

## **Methodes om de beste subset of voorspellende variabelen te kiezen**

- Eerst: vermindering van het aantal voorspellende variabelen door middel van domeinkennis
- Dan: gebruik maken van algoritmes
  - ‘Uitgebreid zoeken’: probeer alle voorspellende variabelen hun subsets
  - ‘Voorwaartse selectie’: start met 1 voorspellende variabele, voeg elke keer de meest relevante toe.
  - ‘Achterwaartse selectie’: start met alle voorspellende variabelen, verwijder telkens de minst relevante
  - ...

