





# Sammenligning av klassifiseringsmetoder

Silje Helgesen 

Hanne-Kristin Schrøder Leiros 

Søketeknologi 

Science fair   18. november 2024

1 Hva er klassifisering? 

2 Statistiske metoder 

3 Vektormodeller 

4 Sammenligning 

# Hva er klassifisering?

- 🔍 Gitt en mengde klasser  $\mathbb{C}$ , avgjør hvilke klasse(r) et dokument tilhører.

# Hva er klassifisering? 🗂️

- Gitt en mengde klasser  $\mathbb{C}$ , avgjør hvilke klasse(r) et dokument tilhører.
- Klassifiserere har en funksjon,  $\gamma : \mathbb{X} \rightarrow \mathbb{C}$ .

# Hva er klassifisering? 🗂️

- Gitt en mengde klasser  $\mathbb{C}$ , avgjør hvilke klasse(r) et dokument tilhører.
- Klassifiserere har en funksjon,  $\gamma : \mathbb{X} \rightarrow \mathbb{C}$ .
- Alle eksemplene her er veiledet læring.

# Hva er klassifisering? 🗂️

- 🔍 Gitt en mengde klasser  $\mathbb{C}$ , avgjør hvilke klasse(r) et dokument tilhører.
- 🔍 Klassifiserere har en funksjon,  $\gamma : \mathbb{X} \rightarrow \mathbb{C}$ .
- 🔍 Alle eksemplene her er veiledet læring.
  - Altså konstruerer de  $\gamma$  basert på anotert treningsdata,  $\mathbb{D}$ . 📖

# Hva er klassifisering? 📁

- 🔍 Gitt en mengde klasser  $\mathbb{C}$ , avgjør hvilke klasse(r) et dokument tilhører.
- 🔍 Klassifiserere har en funksjon,  $\gamma : \mathbb{X} \rightarrow \mathbb{C}$ .
- 🔍 Alle eksemplene her er veiledet læring.
  - Altså konstruerer de  $\gamma$  basert på anotert treningsdata,  $\mathbb{D}$ . 📖
  - $\Gamma(\mathbb{D}) = \gamma$ , der  $\Gamma$  er treningen.

# (Multinomial) Naive Bayes

- Lager statistikk om treningsdataen og bruker dette til å avgjøre klassen.



# (Multinomial) Naive Bayes

- Lager statistikk om treningsdataen og bruker dette til å avgjøre klassen.
- Finner klassen med høyest sannsynlighet,  $c_{map}$ .

# (Multinomial) Naive Bayes

- Lager statistikk om treningsdataen og bruker dette til å avgjøre klassen.
- Finner klassen med høyest sannsynlighet,  $c_{map}$ .
  - Ganger andelen dokumenter som tilhører klassen med sannsynligheten for at hver term i dokumentet tilhører klassen.

# (Multinomial) Naive Bayes

- Lager statistikk om treningsdataen og bruker dette til å avgjøre klassen.
- Finner klassen med høyest sannsynlighet,  $c_{map}$ .
  - Ganger andelen dokumenter som tilhører klassen med sannsynligheten for at hver term i dokumentet tilhører klassen.
  - $\hat{P}(c) = \frac{N_c}{N}$

# (Multinomial) Naive Bayes

- Lager statistikk om treningsdataen og bruker dette til å avgjøre klassen.
- Finner klassen med høyest sannsynlighet,  $c_{map}$ .
  - Ganger andelen dokumenter som tilhører klassen med sannsynligheten for at hver term i dokumentet tilhører klassen.
  - $\hat{P}(c) = \frac{N_c}{N}$
  - $\hat{P}(t_i|c) = \frac{\text{count}(t_i, c)}{\sum_{t \in T} \text{count}(t, c)}$

# (Multinomial) Naive Bayes

- Lager statistikk om treningsdataen og bruker dette til å avgjøre klassen.
- Finner klassen med høyest sannsynlighet,  $c_{map}$ .
  - Ganger andelen dokumenter som tilhører klassen med sannsynligheten for at hver term i dokumentet tilhører klassen.
  - $\hat{P}(c) = \frac{N_c}{N}$
  - $\hat{P}(t_i|c) = \frac{\text{count}(t_i, c)}{\sum_{t \in T} \text{count}(t, c)}$
  - $c_{map} = \arg \max_{c \in C} \hat{P}(c) \prod_{t \in T} \hat{P}(t|c)$

# (Multinomial) Naive Bayes

- Lager statistikk om treningsdataen og bruker dette til å avgjøre klassen.
- Finner klassen med høyest sannsynlighet,  $c_{map}$ .
  - Ganger andelen dokumenter som tilhører klassen med sannsynligheten for at hver term i dokumentet tilhører klassen.
  - $\hat{P}(c) = \frac{N_c}{N}$
  - $\hat{P}(t_i|c) = \frac{\text{count}(t_i, c)}{\sum_{t \in T} \text{count}(t, c)}$
  - $c_{map} = \arg \max_{c \in C} \hat{P}(c) \prod_{t \in T} \hat{P}(t|c)$
- Lineær klassegrense.



# KNN - k nærmeste naboer



- Klassen til de fleste av de  $k$  nærmeste vektorene i vektorrommet avgjør klassen til et dokument.



# KNN - k nærmeste naboer



- Klassen til de fleste av de  $k$  nærmeste vektorene i vektorrommet avgjør klassen til et dokument.
- Avgjøres lokalt





# KNN - k nærmeste naboer



- Klassen til de fleste av de  $k$  nærmeste vektorene i vektorrommet avgjør klassen til et dokument.
- Avgjøres lokalt
- Håndterer utliggere godt.



# KNN - k nærmeste naboer



- Klassen til de fleste av de  $k$  nærmeste vektorene i vektorrommet avgjør klassen til et dokument.
- Avgjøres lokalt
- Håndterer utliggere godt.
- Ikke-lineære klassegrenser.

# Rocchio - Nearest Centroid classifier TOP

- Lager en *sentroide* for hver klasse.

# Rocchio - Nearest Centroid classifier TOP

- Lager en *sentroide* for hver klasse.
- Et nytt dokument som skal klassifiseres blir tildelt klassen til nærmeste sentroide.

# Rocchio - Nearest Centroid classifier TOP

- Lager en *sentroide* for hver klasse.
- Et nytt dokument som skal klassifiseres blir tildelt klassen til nærmeste sentroide.
- Mer påvirket av utliggere enn KNN.

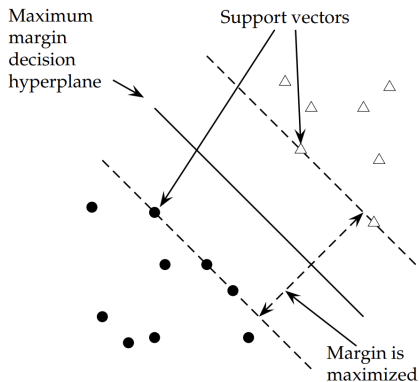
# Rocchio - Nearest Centroid classifier TOP

- Lager en *sentroide* for hver klasse.
- Et nytt dokument som skal klassifiseres blir tildelt klassen til nærmeste sentroide.
- Mer påvirket av utliggere enn KNN.
- Lineær klassegrense.

# Support Vector Machines



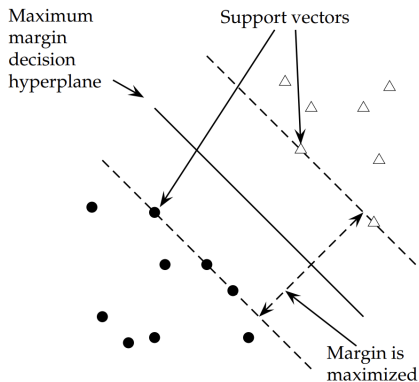
- Finne skilletet som har størst plass mellom to klasser i vektorrommet.



Figur 15.1 fra *Introduction to Information retrieval*. [1]

# Support Vector Machines

- Finne skillet som har størst plass mellom to klasser i vektorrommet.
- Lineære klassegrenser.



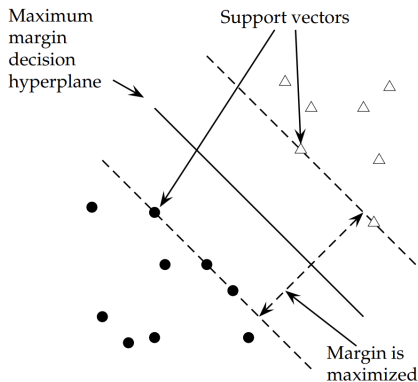
Figur 15.1 fra *Introduction to Information retrieval*. [1]





# Support Vector Machines

- Finne skillet som har størst plass mellom to klasser i vektorrommet.
- Lineære klassegrenser.
- Kan finne ikke-lineære klassegrenser med kjerne triks.



Figur 15.1 fra *Introduction to Information retrieval*. [1]

# Sammenligning

## Klassegrense

### lineær —

- Rocchio
- Naive Bayes
- SVM

### ikke-lineær ~

- KNN
- SVM + kernel

# Sammenligning

## Klassegrense

### lineær —

- Rocchio
- Naive Bayes
- SVM

### ikke-lineær ~

- KNN
- SVM + kernel

Hvor mye data har du? 

# Sammenligning **VS**

## Klassegrense

### 🔍 lineær —

- Rocchio
- Naive Bayes
- SVM

### 🔍 ikke-lineær ~

- KNN
- SVM + kernel

## Hvor mye data har du? 📄

### 🔍 Ingen?

# Sammenligning

## Klassegrense

- 🔍 lineær —
  - Rocchio
  - Naive Bayes
  - SVM
- 🔍 ikke-lineær ~
  - KNN
  - SVM + kernel

## Hvor mye data har du? 📄

- 🔍 Ingen? Håndskrevne regler

# Sammenligning

## Klassegrense

- 🔍 lineær —
  - Rocchio
  - Naive Bayes
  - SVM
- 🔍 ikke-lineær ~
  - KNN
  - SVM + kernel

## Hvor mye data har du? 📄

- 🔍 Ingen? Håndskrevne regler
- 🔍 Lite data?

# Sammenligning **VS**

## Klassegrense

### 🔍 lineær —

- Rocchio
- Naive Bayes
- SVM

### 🔍 ikke-lineær ~

- KNN
- SVM + kernel

## Hvor mye data har du? 📄

- 🔍 Ingen? Håndskrevne regler
- 🔍 Lite data? NB

# Sammenligning **VS**

## Klassegrense

- lineær —
  - Rocchio
  - Naive Bayes
  - SVM
- ikke-lineær ~
  - KNN
  - SVM + kernel

## Hvor mye data har du? 📄

- Ingen? Håndskrevne regler
- Lite data? NB
- Fornuftig mengde data?



# Sammenligning **VS**

## Klassegrense

- 🔍 lineær —
  - Rocchio
  - Naive Bayes
  - SVM
- 🔍 ikke-lineær ~
  - KNN
  - SVM + kernel

## Hvor mye data har du? 📄

- 🔍 Ingen? Håndskrevne regler
- 🔍 Lite data? NB
- 🔍 Fornuftig mengde data? alle de kule

# Sammenligning **VS**

## Klassegrense

- lineær —
  - Rocchio
  - Naive Bayes
  - SVM
- ikke-lineær ~
  - KNN
  - SVM + kernel

## Hvor mye data har du? 📄

- Ingen? Håndskrevne regler
- Lite data? NB
- Fornuftig mengde data? alle de kule
- Mye data?

# Sammenligning **VS**

## Klassegrense

- lineær —
  - Rocchio
  - Naive Bayes
  - SVM
- ikke-lineær ~
  - KNN
  - SVM + kernel

## Hvor mye data har du? 📄

- Ingen? Håndskrevne regler
- Lite data? NB
- Fornuftig mengde data? alle de kule
- Mye data? NB?

- [1] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*. eng. Cambridge: Cambridge University Press, 2008.

Takk for oss 🎉