IN1140: Introduksjon til språkteknologi

Forelesning #8

Sondre Wold

Universitetet i Oslo

21. oktober 2024



I dag skal vi se på:

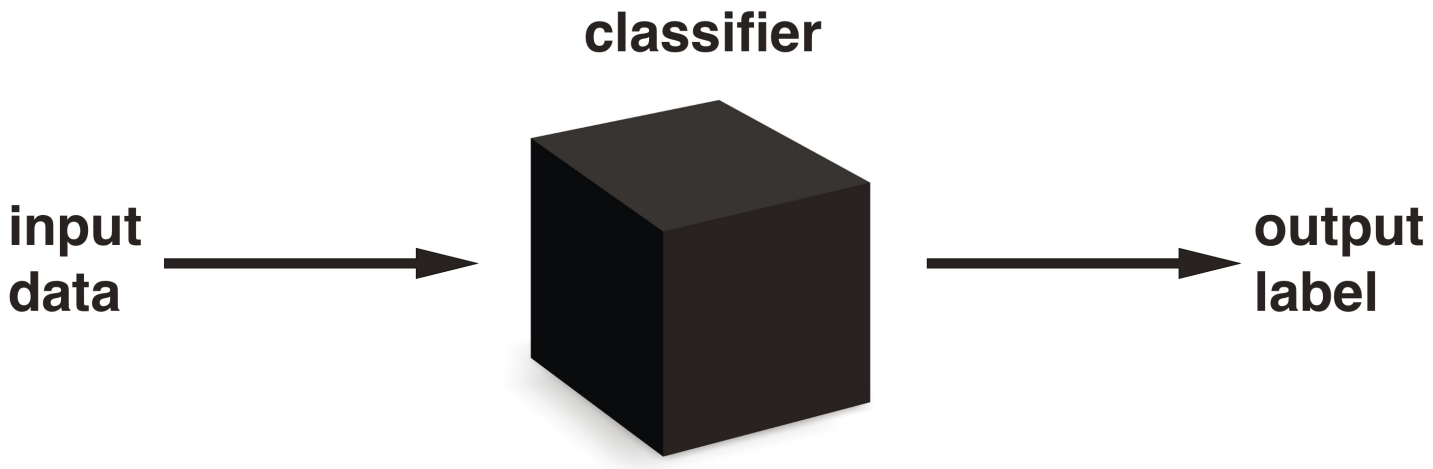
- Klassifisering - hva er egentlig det?
- Bayes regel
- Naive Bayes
- Avisidentifikasjon med Naive Bayes, kodegjennomgang.

Men først...

Rekk opp hånda om det er noe som ikke er klart! Her skal alle med!

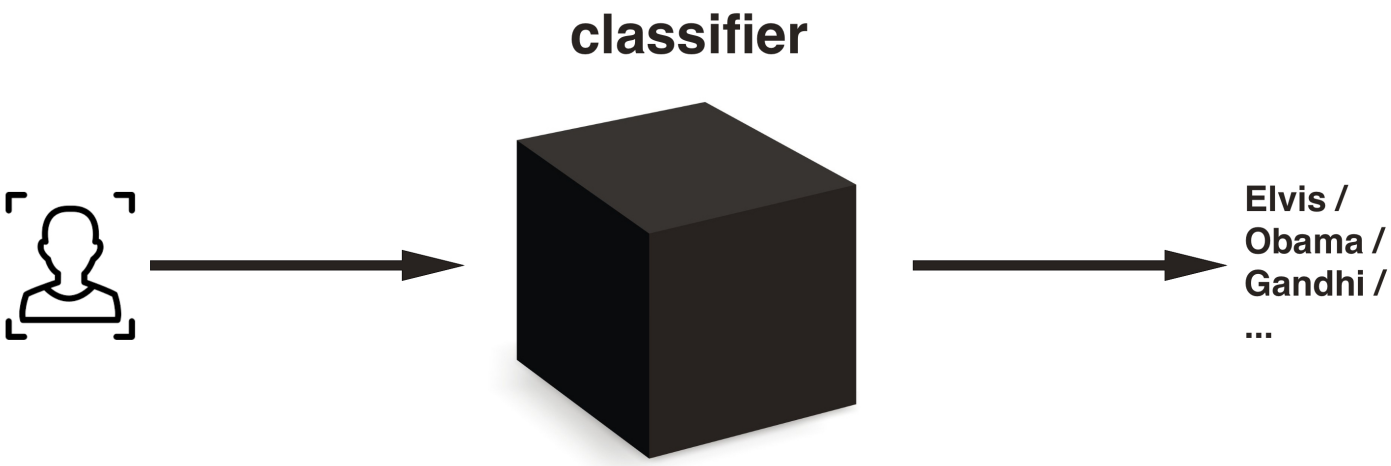
Klassifikasjon

- Gitt et sett av sett av mulige kategorier:
- tildel en kategori (*label*) til et ukjent objekt.



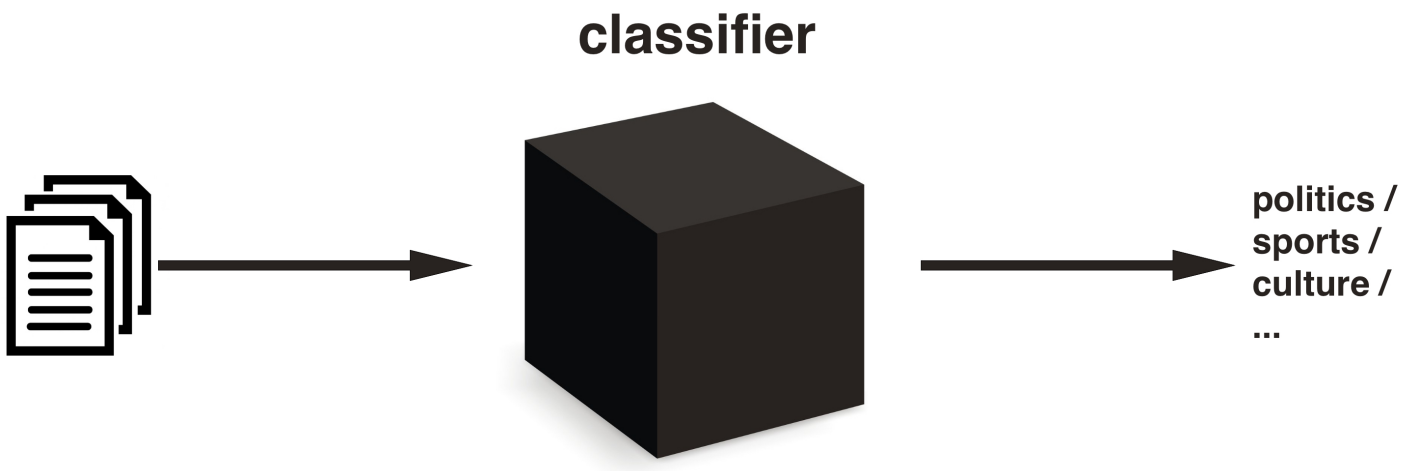
Klassifikasjon

- Gitt et sett av sett av mulige kategorier:
- tildel en kategori (*label*) til et ukjent objekt.



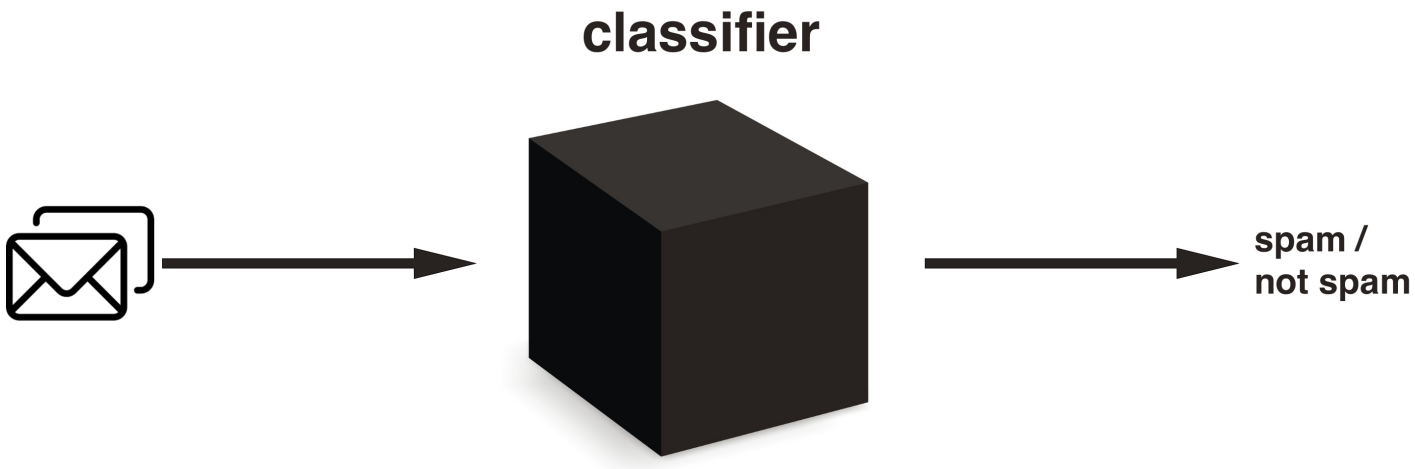
Klassifikasjon

- Gitt et sett av sett av mulige kategorier:
- tildel en kategori (*label*) til et ukjent objekt.

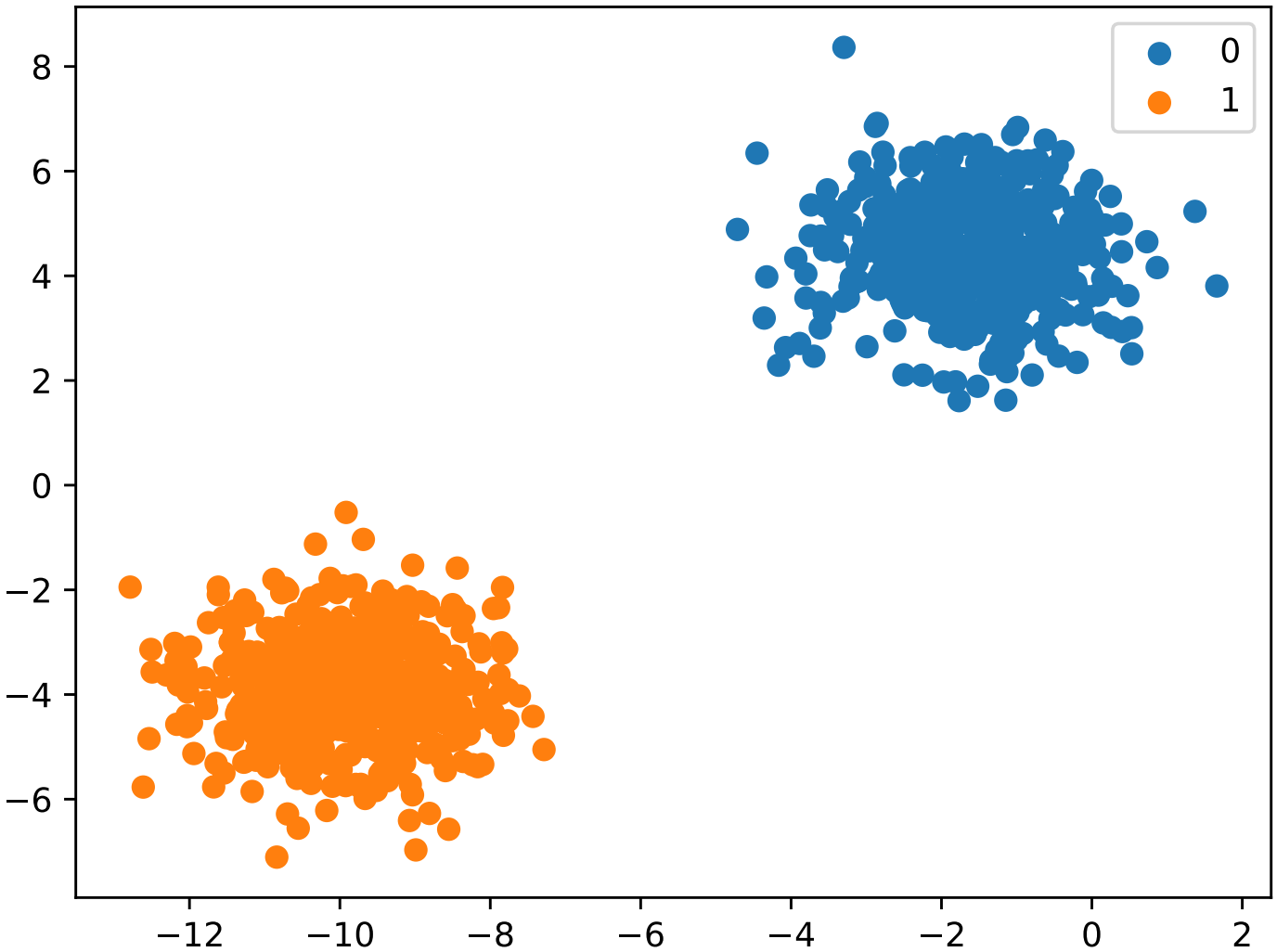


Klassifikasjon

- Gitt et sett av sett av mulige kategorier:
- tildel en kategori (*label*) til et ukjent objekt.



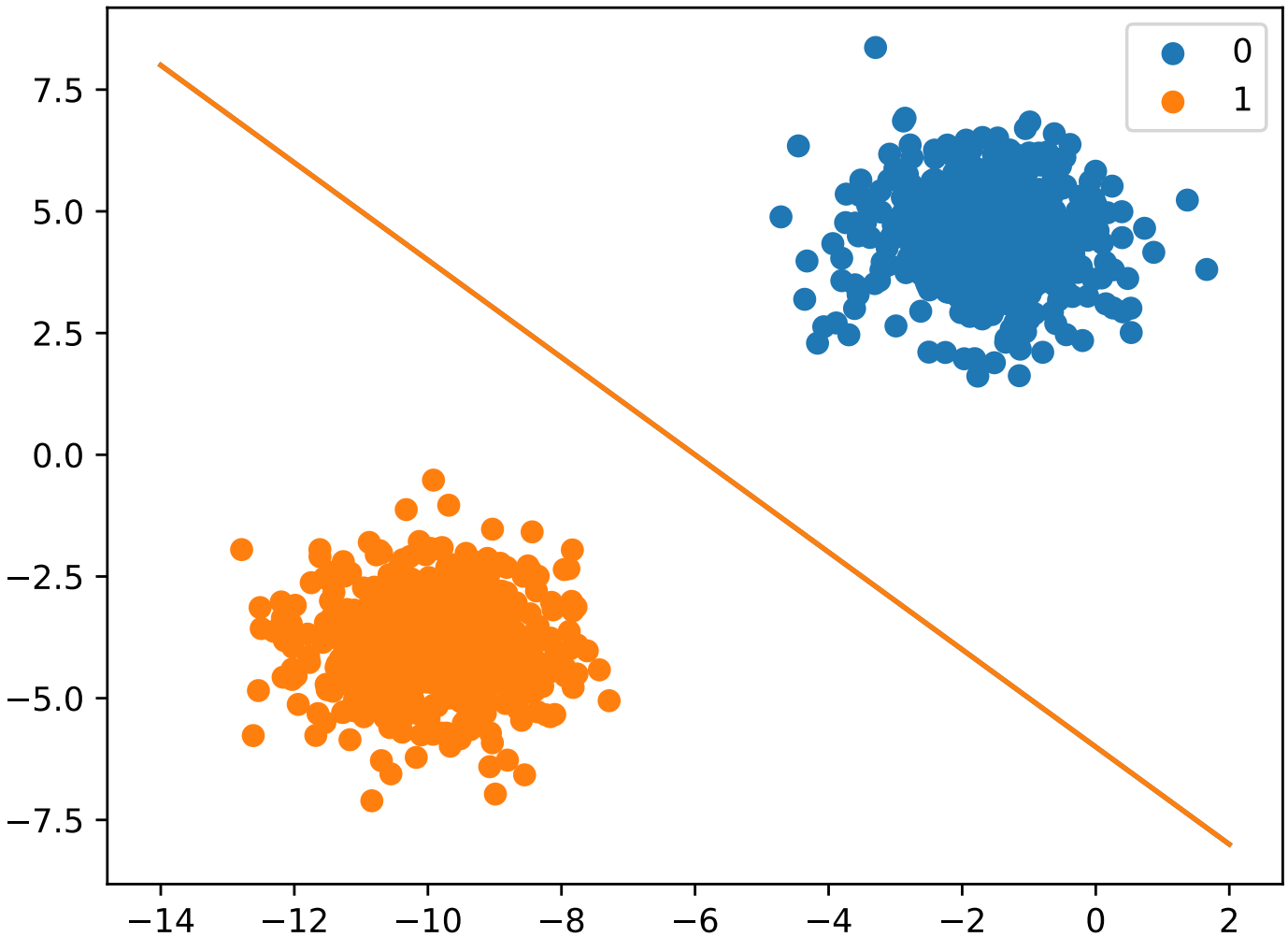
Typer klassifisering: binær



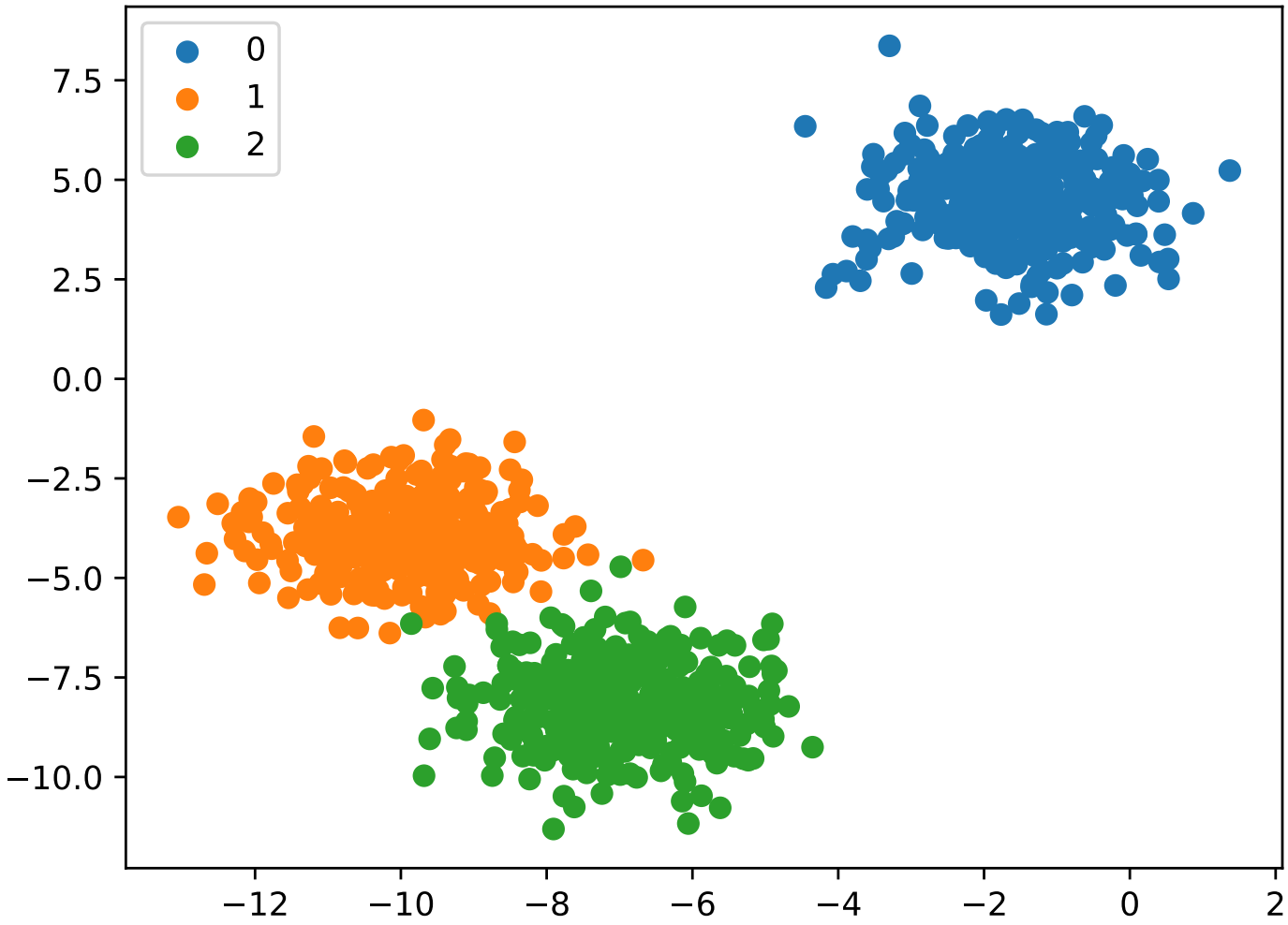
Typer klassifisering: binær

I dette tilfellet: kan skilles med en lineær funksjon:

$$f(x) = ax + b$$



Typer klassifisering: flere klasser



Tekstklassifisering

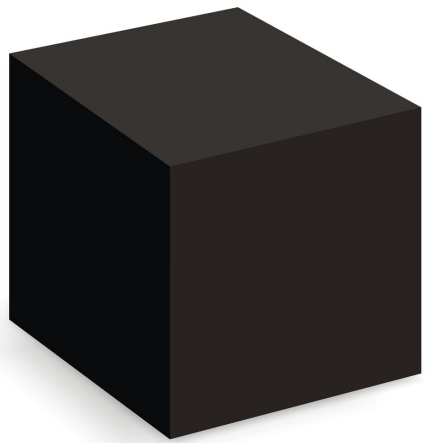
- Tekstklassifisering; problemer der vi ønsker å predikere en klasse for en hel tekst, f.eks et helt dokument, en tweet, eller bare en setning.
- Merk: vi bruker også ofte klassifisering på andre nivåer; f.eks på enkeltord eller sekvenser av ord.
- Eksempler på tekstklassifisering:

Tekstklassifisering

- Tekstklassifisering; problemer der vi ønsker å predikere en klasse for en hel tekst, f.eks et helt dokument, en tweet, eller bare en setning.
- Merk: vi bruker også ofte klassifisering på andre nivåer; f.eks på enkeltord eller sekvenser av ord.
- Eksempler på tekstklassifisering:
- Spamfilter
- Klassifikasjon av tema (*topic classification*)
- Sentimentanalyse (f.eks positiv/negativ polaritet)
- Identifikasjon av forfatterskap (*author identification*)
- Identifikasjon av alder, kjønn, osv. (eksempler på *author profiling*)
- Språk-og dialekt-gjenkjenning
- Identifikasjon av hatefulle ytringer (*hate speech detection*)
- ...

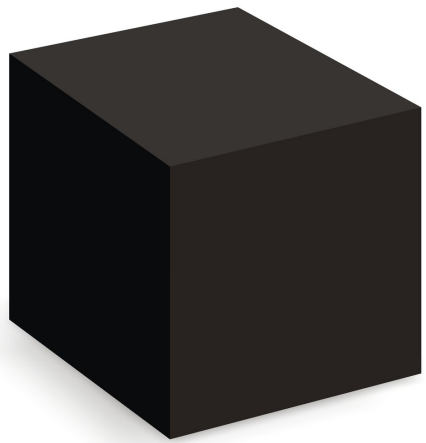
Hva kan utgjøre selve klassifikatoren?

- Regelbasert klassifisering basert på håndskrevne regler:
  "dersom dokumentet inneholder ordene x og y men ikke q og z så..."
- Vanskelig å vedlikeholde og skalere.
- Hvordan vekte ulike 'bevis'?



Hva kan utgjøre selve klassifikatoren?

- Regelbasert klassifisering basert på håndskrevne regler:
  "dersom dokumentet inneholder ordene x og y men ikke q og z så..."
- Vanskelig å vedlikeholde og skalere.
- Hvordan vekte ulike 'bevis'?
- Langt mer vanlig: basert på maskinlæring.
- Finnes mange ulike metoder / modeller.
- Vi skal fokusere på metoden Naive Bayes.
- Enkel men effektiv!
- Men først litt om det generelle oppsettet.



Veiledet læring

- Innen maskinlæring er klassifikasjon et eksempel på en tilnærming som kalles veiledet læring (*supervised learning*).



Veiledet læring

- Innen maskinlæring er klassifikasjon et eksempel på en tilnærming som kalles veiledet læring (*supervised learning*).
- Krever treningsdata: eksempler på input sammenstilt med en 'merkelapp' (*label*) som angir riktig klasse. En 'fasit'.
- Vi sier at dataene er annoterte.
- F.eks en samling av;
  - portretter annotert med navn på person
  - eposter annotert som spam / ikke-spam
  - dokumenter annotert med tema
  - dokumenter annotert med polaritet



Veiledet læring

- Klassifikatoren trenes til fange opp statistiske mønstre i treningsdataene, slik at vi kan anvende den på nye 'usette' datapunkter.
- Modellen kan ikke bare memorere treningsdataene; må kunne generalisere til nye data.

Annoterte data er en flaskehals

- Generelt; jo mere treningsdata, jo bedre modell.
- Gode data (kvalitet + kvantitet) viktigere enn spesifikk metode.
- Men å annotere data krever typisk manuell innsats, ofte av noen med ekspertise for et gitt domene.
- Ressurskrevende i form av tid, penger, og innsats.
- Finnes også ikke-veiledet læring (unsupervised learning) som ikke krever annoterte data, men dette gir forløpig langt mindre presise resultater.
- (Ser eksempler på dette i IN2110)
- Knapphet på data er regelen snarere enn unntaket.

Evaluering

- To viktige trinn ved utvikling av en maskinlæringsmodell:
  1: estimering / læring / trening.
  2: evaluering / validering / testing.
- Vi trenger annoterte data også for å teste klassifikatoren:
- vi sammenlikner prediksjonene til modellen med 'gullstandarden' for å kunne regne ut nøyaktigheten.
- Trenger ulike datasplitter til trening og testing (*held-out test data)*
- Som regel ønsker vi også en tredje 'split' for å evaluere modellen mens den er under utvikling; valideringsdata eller *development*-data

Accuracy

- Accuracy $= \frac{\text{correct predictions}}{\text{total predictions}}$
- Et mål på hvor nøyaktig klassfikatoren er.

Accuracy

- Accuracy $= \frac{\text{correct predictions}}{\text{total predictions}}$
- Et mål på hvor nøyaktig klassfikatoren er.
- Accuracy kan fungere dårlig dersom man har veldig ubalanserte klasser.
- Jf. accuracy for en *majority class baseline* for spam-klassifikasjon.
- Finnes også mange andre mål (f.eks. *precision*, *recall*, og *F1* – disse dekkes i IN2110).

Trekk-representasjon

- Første skritt for å trene en modell er å velge hvordan vi skal representere dataene.
- Gjøres ved såkalte trekk (*features*).
- Bryter ned dataene til forskjellige mindre og tellbare egenskaper som vi tror kan være relevante for å predikere klassene.
- Trinnet der vi henter ut og registrerer trekk for dataene våre kalles trekk-ekstraksjon (*feature extraction*).
- Trekkene defineres med utgangspunkt i treningsdataene; test-dataene bruker samme settet av trekk-typer.

Positiv eller negativ?

En juleklassiker er utsprungen

Albumjul



Guren Hagen med Sølvguttene og Ole Edvard Antonsen
Julekveld hos oss

En hjertevarm, personlig juleplate full av herlig musikalsk variasjon.

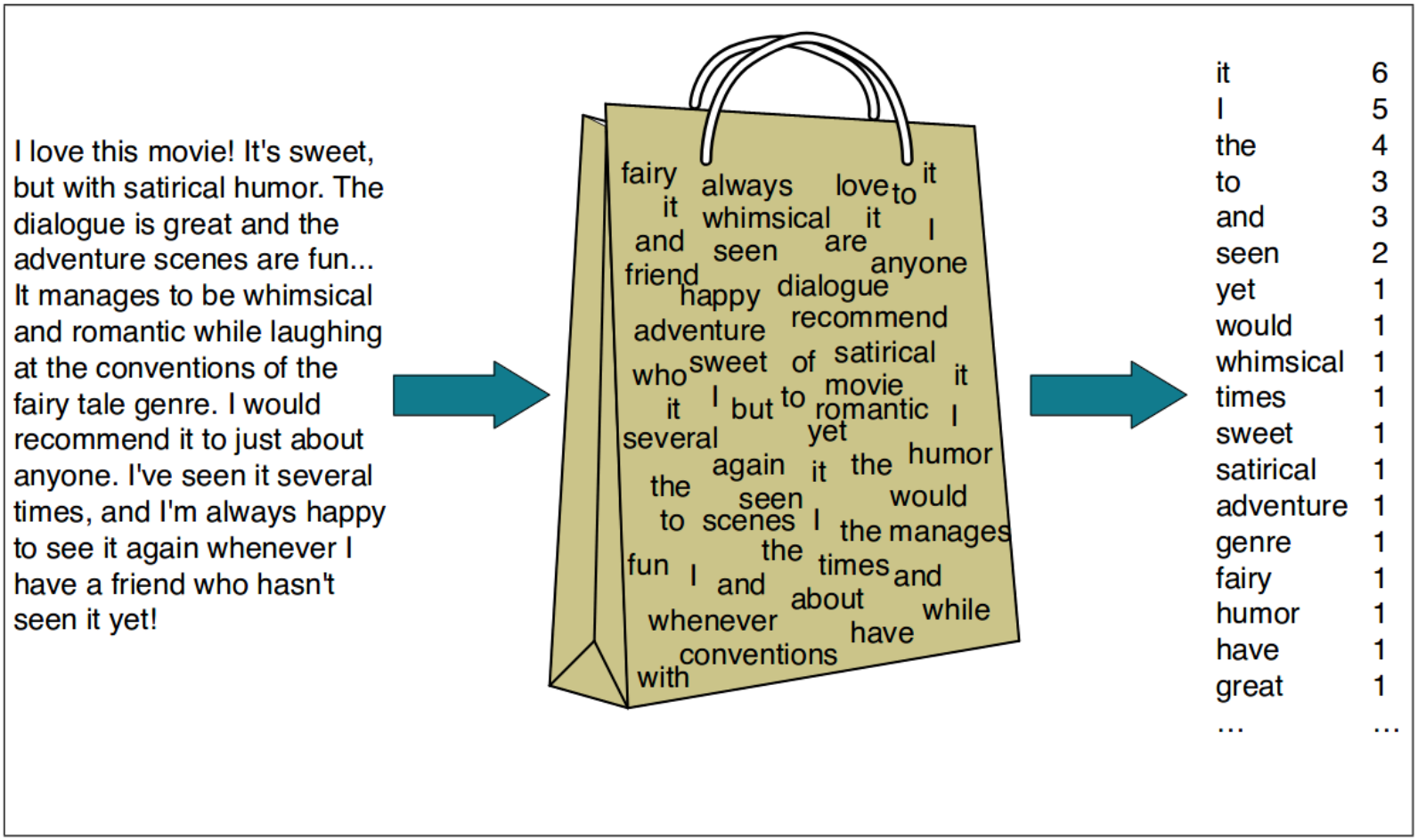
Guren Hagen er en av våre fineste viseartister, og med *Julekveld hos oss* har han skapt ei stemningsfull juleplate med stor troverdighet. Spesielt hans egenkomponerte julesanger gjør dette til julens store stemningsspreder. Hans varme stemme og østerdalsdialekt drar deg inn hans verden, mens Sølvguttenes vakre sang og Ole Edvard Antonsens trompet gir ekstra krydder til en av årets mest vellykkede juleplater. Guren Hagen har på utsøkt vis i tittelmelodien klart å formidle hvordan han opplevde barndommens østerdalsjul slik at vi alle vil være en del av denne kosen. Samtidig gjør han Alf Prøysens «Julekveldsvisa», med hawaiigitar og lapsteel til en sann fryd. I spennet mellom disse to sangene gir han oss troen tilbake på at det finnes overraskende grep også for denne musikktradisjonen. Selv kjente travere som «Det lyser i stille grender» og «Kimer i klokker», sistnevnte i en flott duett med Lise Backstrøm, besitter en nyvunnen nærhet. Dette er blitt en vidunderlig juleplate full av musikalsk overskudd og variasjon. Musikken bærer i seg like porsjoner glede, nærhet og undring.

Svein Andersen

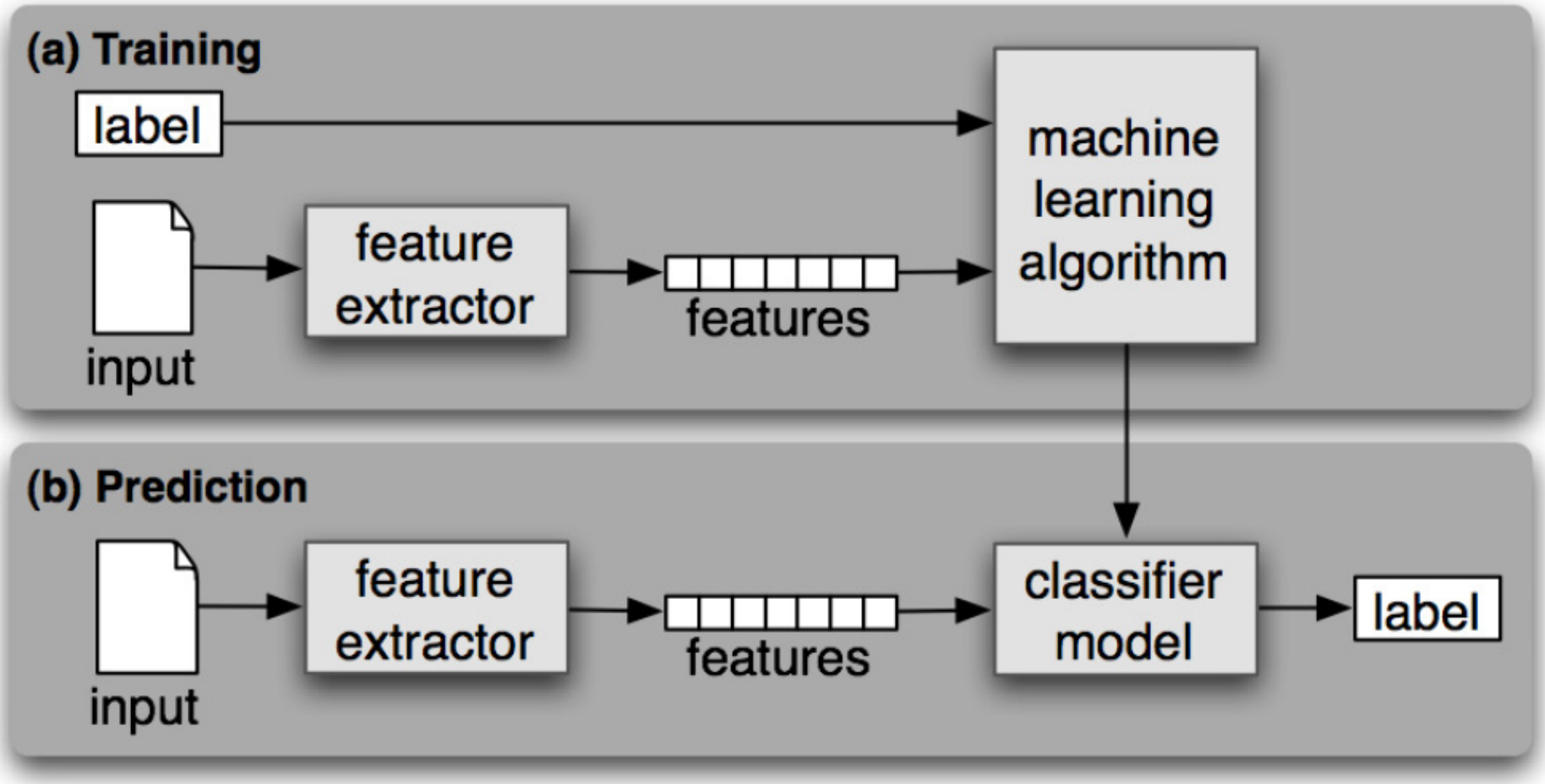
- Norwegian Review Corpus (NoReC): over 43K anmeldelser
- Kan bruke terningkastet som annotasjon
- For oblig 2b: terningkast 1–2 = negativ, terningkast 5–6 = positiv.
- Hva kan være gode trekk?

Bag-of-words

- Vanligste trekk-type for tekstklassifikasjon: bag-of-words / BoW
- Hvert ord som forekommer i dokumentet utgjør et trekk, bryr oss ikke om rekkefølge eller posisjon.
- Unionen av alle trekkene, altså alle ordtypene, kaller vi vokabularet (V).



Klassifikasjon med maskinlæring



Bilde fra NLTK

Litt notasjon og oppsummering

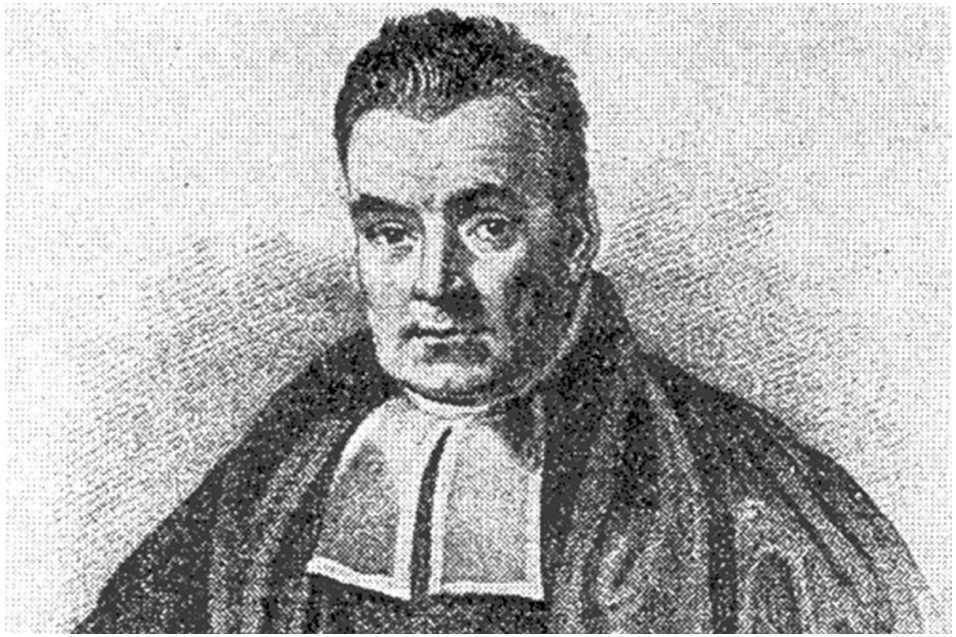
- Vi har et forhåndsdefinert sett med klasser $C = \{c_1, c_2, ..., c_m\}$
- Treningsdataene våre er gitt ved en mengde par av dokumenter og tilsvarende klasse, $\{\langle d_1, c_1\rangle, \langle d_2, c_2\rangle, \ldots, \langle d_n, c_n\rangle\}$
- Ved BoW-trekk lar vi hvert dokument være representert ved ordene som forekommer i det, $d = \langle w_1, w_2, \ldots, w_k\rangle$
- (Merk at lengden k vil variere mellom dokumenter).

Litt notasjon og oppsummering

- Vi har et forhåndsdefinert sett med klasser $C = \{c_1, c_2, ..., c_m\}$
- Treningsdataene våre er gitt ved en mengde par av dokumenter og tilsvarende klasse, $\{\langle d_1, c_1\rangle, \langle d_2, c_2\rangle, \ldots, \langle d_n, c_n\rangle\}$
- Ved BoW-trekk lar vi hvert dokument være representert ved ordene som forekommer i det, $d = \langle w_1, w_2, \ldots, w_k\rangle$
- (Merk at lengden k vil variere mellom dokumenter).
- Naive Bayes er en probabilistisk modell:
- Målet er å trene en modell som gitt et dokument d, gir oss den mest sannsynlige klassen: $\hat{c} = \underset{c \in C}{\operatorname{argmax}}\ P(c|d)$

Bayes' regel (Bayes' teorem)

$$\underbrace{P(H|E)}_{\text{The "Posterior"}} = \frac{\overbrace{P(E|H)}^{\text{The "Likelihood"}}\overbrace{P(H)}^{\text{The "Prior"}}}{\underbrace{P(E)}_{\text{The "Marginal"}}}$$



Naive Bayes

- Utgangspunkt: klassifikasjonen skjer ved å beregne den mest sannsynlige klassen $\hat{c} \in C$ for et gitt dokument $d = \langle w_1, w_2, \ldots, w_k \rangle$:

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} \; P(c|d)$$

Naive Bayes

- Utgangspunkt: klassifikasjonen skjer ved å beregne den mest sannsynlige klassen $\hat{c} \in C$ for et gitt dokument $d = \langle w_1, w_2, \ldots, w_k \rangle$:

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} \; P(c|d)$$

- Etter omskriving med Bayes' regel

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} \; \frac{P(d|c)P(c)}{P(d)}$$

Naive Bayes

- Utgangspunkt: klassifikasjonen skjer ved å beregne den mest sannsynlige klassen $\hat{c} \in C$ for et gitt dokument $d = \langle w_1, w_2, \ldots, w_k \rangle$:

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} \; P(c|d)$$

- Etter omskriving med Bayes' regel

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} \; \frac{P(d|c)P(c)}{P(d)}$$

- Vi kan forenkle denne ytterligere; siden $P(d)$ er uavhengig av c og vil være konstant for hver d kan den droppes:

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} \; P(d|c)P(c)$$

Naive Bayes

- Vi har nå redusert til:

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} \; P(d|c)P(c)$$

- Siden hvert dokument d er representert ved ordene det består av $d = \langle w_1, w_2, \ldots, w_k \rangle$ kan vi tenke på dette som:

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} \; P(w_1, w_2, \ldots, w_k|c)P(c)$$

Naive Bayes

- Vi har nå redusert til:

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} \; P(d|c)P(c)$$

- Siden hvert dokument d er representert ved ordene det består av $d = \langle w_1, w_2, \ldots, w_k \rangle$ kan vi tenke på dette som:

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} \; P(w_1, w_2, \ldots, w_k|c)P(c)$$

- Men det er problematisk å estimere denne distribusjonen direkte, pga dataknapphet (*sparse data*); de fleste sekvenser $w_1, w_2, \ldots, w_k$ vil aldri ha blitt observert i treningsdataene.

Naive Bayes

- Vi gjør derfor en forenkling og antar 'naivt' at alle trekk er uavhengige av andre trekk, gitt klassen.
- Antar også at vi ikke bryr oss om posisjonen til ord-trekkene.
- I stedet for sannsynligheten for alle trekkene samlet, trenger vi bare sannsynligheten for hvert individuelle trekk (betinget på klassen):

$$P(d|c) = P(w_1, w_2, \ldots, w_k|c) \approx \prod_{i=1}^{k} P(w_i|c)$$

Hvordan trene NB-klassifikatoren

- Vi har nå den endelige formuleringen av NB-klassifikatoren:

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} \; P(c) \prod_{w \in d} P(w|c)$$

- For å trene klassifikatoren kan vi beregne sannsynlighetene ved relative frekvenser i treningsdataene, såkalt *Maximum Likelihood Estimation*
- To sannsynligheter:
  1: Prior – sannsynligheten for klassen: $P(c)$
  2: Likelihood – sannsynligheten for individuelle trekk gitt klassen: $P(w_i|c)$

Estimering av NB-sannsynlighetene

- $P(c) = \frac{N_c}{N_{doc}}$
- $N_c =$ antall dokumenter i treningsdataene som er i klassen c og $N_{doc} =$ total antall dokumenter

Estimering av NB-sannsynlighetene

- $P(c) = \frac{N_c}{N_{doc}}$
- $N_c =$ antall dokumenter i treningsdataene som er i klassen c og $N_{doc} =$ total antall dokumenter
- $P(w_i|c) = \frac{count(w_i,c)}{\sum_{w \in V} count(w,c)}$
- Med andre ord: Antall ganger ordet w_i forekommer i alle dokumenter av klasse c, delt på antall forekomster av alle ord i vokabularet V som finnes i alle dokumenter av klasse c.

MLE og “the zero-frequency problem”

- La oss anta at vi prøver å finne sannsynligheten for ordet ‘fantastisk’ gitt klassen ‘positiv’.
- Anta at ordet ‘fantastisk’ ikke brukes i noen av treningsdokumentene markert som positive:

$$P(fantastisk|positiv) = \frac{count(fantastisk, positiv)}{\sum_{w \in V} count(w, positiv)} = 0$$

MLE og "the zero-frequency problem"

- La oss anta at vi prøver å finne sannsynligheten for ordet 'fantastisk' gitt klassen 'positiv'.
- Anta at ordet 'fantastisk' ikke brukes i noen av treningsdokumentene markert som positive:

$$P(fantastisk|positiv) = \frac{count(fantastisk, positiv)}{\sum_{w \in V} count(w, positiv)} = 0$$

- Siden alle sannsynligheter blir multiplisert med hverandre vil dette lede til at sannsynligheten for hele klassen blir 0.
- Når har vi sett dette problemet før, og hvordan løste vi det?

MLE og "the zero-frequency problem"

- La oss anta at vi prøver å finne sannsynligheten for ordet 'fantastisk' gitt klassen 'positiv'.
- Anta at ordet 'fantastisk' ikke brukes i noen av treningsdokumentene markert som positive:

$$P(fantastisk|positiv) = \frac{count(fantastisk, positiv)}{\sum_{w \in V} count(w, positiv)} = 0$$

- Siden alle sannsynligheter blir multiplisert med hverandre vil dette lede til at sannsynligheten for hele klassen blir 0.
- Når har vi sett dette problemet før, og hvordan løste vi det?
- Da vi jobbet med språkmodellering – tok i bruk glatting (*smoothing*)

Glatting (*smoothing*)

- Omfordeling av sannsynlighetsmassen for å unngå noen av problemene med MLE: Sørg for at alle trekk får frekvens > 0.
- Ta fra de rike og gi til de fattige.
- Enkleste metoden; Add-one smoothing (Laplace):

$$P(w_i|c) = \frac{count(w_i, c) + 1}{\sum_{w \in V} (count(w, c) + 1)}$$
$$= \frac{count(w_i, c) + 1}{(\sum_{w \in V} count(w, c)) + |V|}$$

der $|V|$ er antall ordtyper i vokabularet.



Glatting (*smoothing*)

- Omfordeling av sannsynlighetsmassen for å unngå noen av problemene med MLE: Sørg for at alle trekk får frekvens > 0.
- Ta fra de rike og gi til de fattige.
- Enkleste metoden; Add-one smoothing (Laplace):

$$P(w_i|c) = \frac{count(w_i, c) + 1}{\sum_{w \in V} (count(w, c) + 1)}$$
$$= \frac{count(w_i, c) + 1}{(\sum_{w \in V} count(w, c)) + |V|}$$

der $|V|$ er antall ordtyper i vokabularet.



Bonusspørsmål: Hva skjer om vi bruker "add-one million smoothing"? Hva skjer når denne verdien øker?

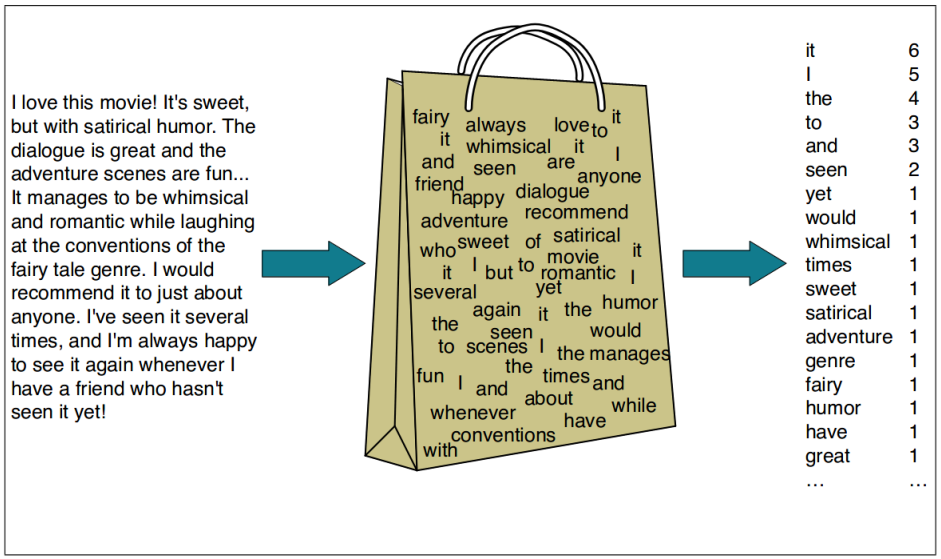
Ukjente ord

- Hva om om vi støter på trekk/ord som ikke finnes i vokabularet V?
- Altså ukjente ord ved testing eller annen anvendelse av modellen – såkalte out-of-vocabulary words (OOV).
- Kan ignoreres – tas ikke med når vi regner ut sannsynlighetene.
- Også mulig å reservere litt sannsynlighetsmasse til ukjente ord;
- f.eks. ved å erstatte alle ord som ikke er i vokabularet eller ord med lav frekvens (f.eks $<= 1$) med en dedikert 'dummy-token' $<unk>$

Kodegjennomgang

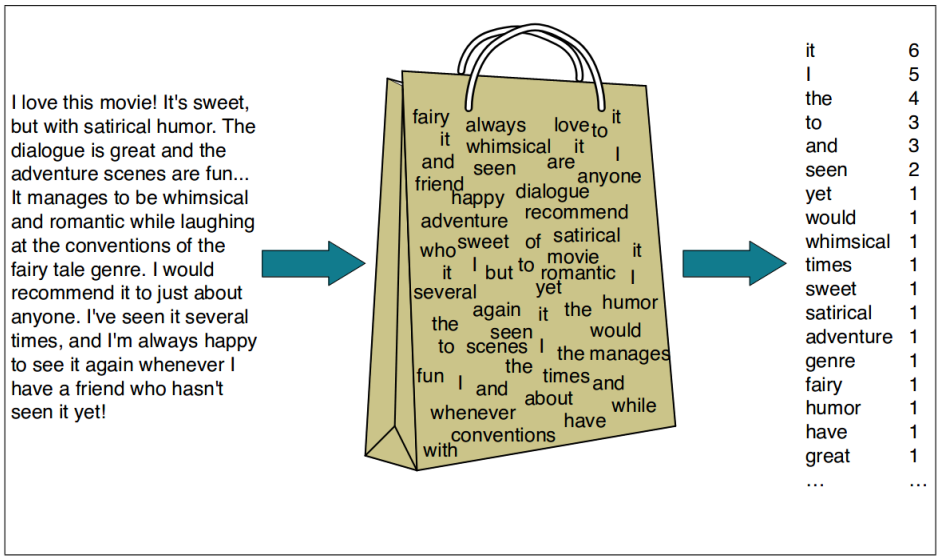
Eksempel: Klassifisere hvilken avis en nyhetsartikkel kommer fra.

Bag-of-words igjen



- Hva er noen mulige problemer med å bruke en BoW-representasjon?

Bag-of-words igjen



- Hva er noen mulige problemer med å bruke en BoW-representasjon?
- Tar ikke høyde for rekkefølge eller kombinasjoner av ord.
- '*Ikke dårlig men bra!*' = '*Ikke bra men dårlig!*'

n-gram

- En enkel men mye brukt måte for å lage trekk med mer kontekst:
- Registrer sekvenser av n ord; n-grammer *n-grams*
- Vanligst å bruke bigram (n=2) eller trigram (n=3)
- Bigrammer for *'Ikke dårlig men bra'*:
  {('*Ikke dårlig*'), ('*dårlig men*'), ('*men bra*') }
- Vanlig BoW tilsvarer unigram (n=1)

Leksikon-trekk

- Spesielt hvis man har lite data kan det fungere å legge til leksikon-baserte trekk i modellen.
- F.eks sentiment-leksikon; lister av ord assosiert med polaritet (positiv/negativ).
- Brukes også ofte alene for å avgjøre dokument-polaritet; bare å telle positive vs. negative treff i leksikonet for ord i dokumentet.
- Ikke uproblematisk, siden polariteten til ord ofte er avhengig av kontekst: '*svake* symptomer' vs. '*svake* prestasjoner'

Negasjon

- Ord og fraser som uttrykker nektelse.
- Viktig og relevant eksempel på effekten av komposisjonalitet.
- Negasjonsord vil ofte reverse polariteten til sentimentbærende ord innenfor dets syntaktiske rekkevidde.
  - I really like this movie. +
  - I really did't like this movie. –
  - Don't dismiss this film, it never got us bored. +

Negasjon

- Mye arbeid innen feltet på å modellere negasjon.
- Men forsøkes også ofte håndtert med enkle heuristikker.
- Kan f.eks konkatenere prefikset `NOT_` på hvert ord som følger et negasjonsord, frem til neste tegnsetting eller et visst antall posisjoner:
  - I didn't like this movie , despite my expectations .
  - I didn't `NOT_like NOT_this NOT_movie` , despite my expectations .

Noen flere alternativer ang. trekk

Eksempler på andre trekk-typer

- n-grams på tegn-nivå
- Ordklasser
- Syntaktiske relasjoner
- Prediksjoner av andre modeller

Noen flere alternativer ang. trekk

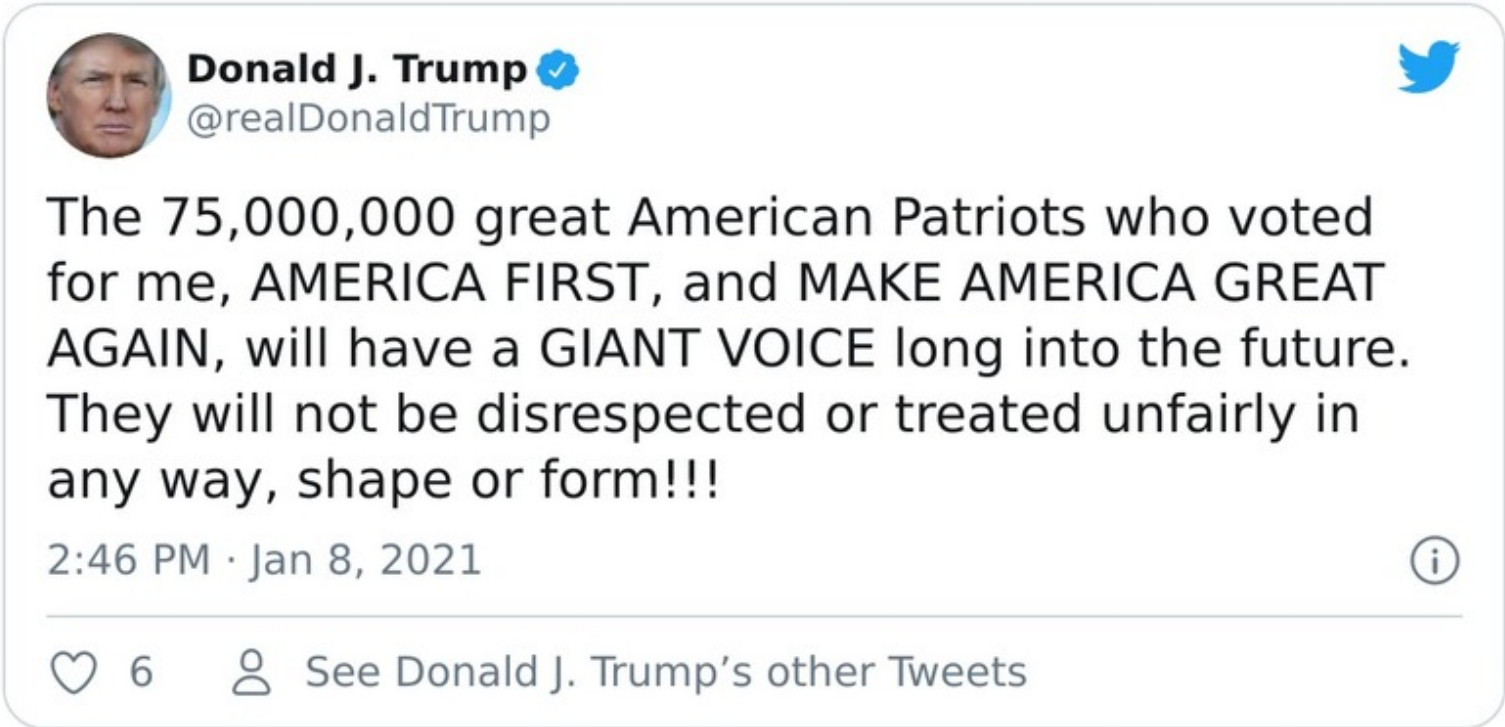
Eksempler på andre trekk-typer

- n-grams på tegn-nivå
- Ordklasser
- Syntaktiske relasjoner
- Prediksjoner av andre modeller

Andre teknikker

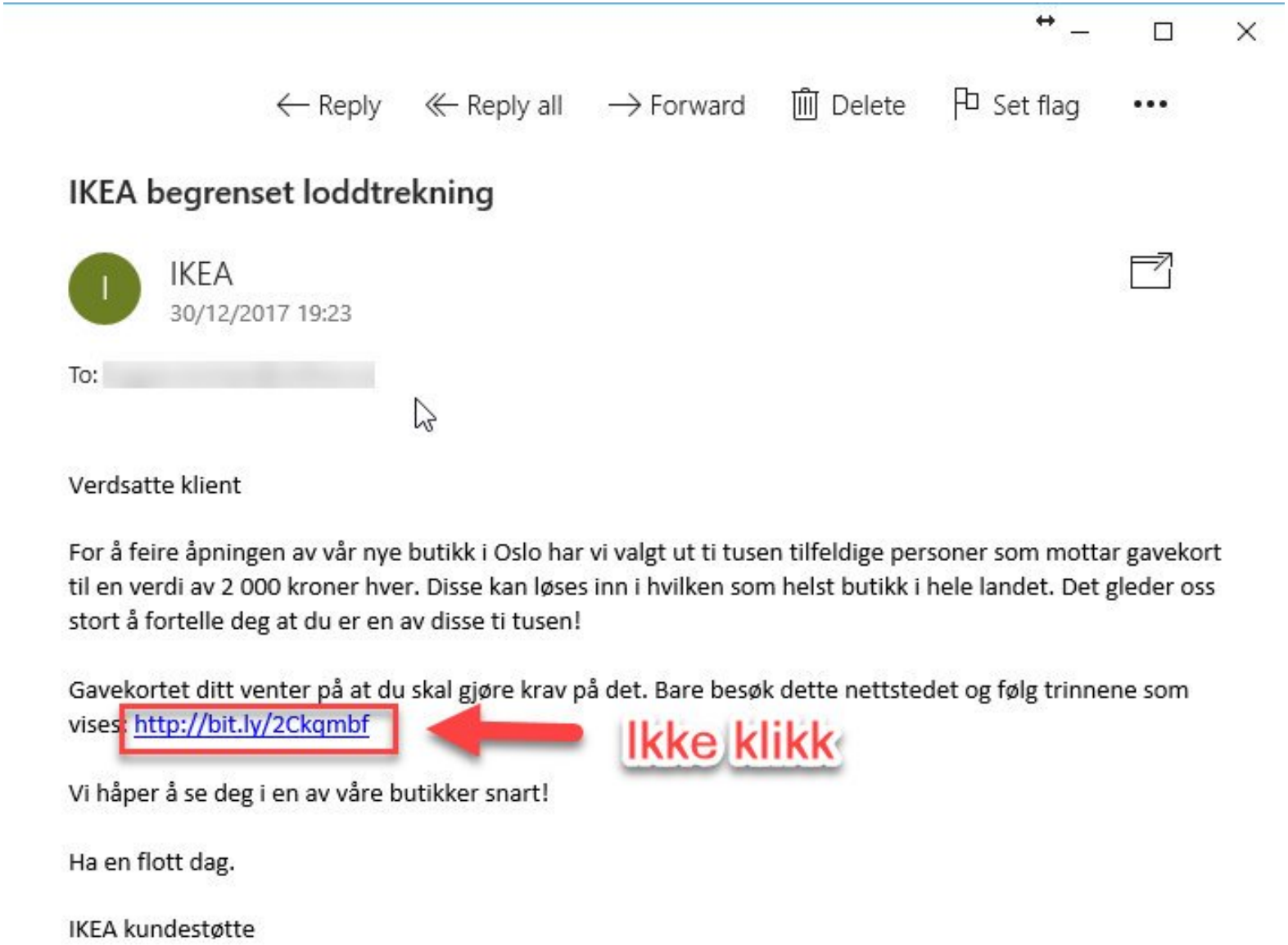
- 'Bevisstyrke': merk at vi kan bruke $P(w_i|c)$ til å få vite noe om hvilke trekk som er mest informative.
- Trekk-vekting (feature weighting); ulike vektingsfunksjoner kan brukes for å transformere rå frekvenser til vektede verdier (dekkes i IN2110).
- Trekk-utvelgelse (feature selection); f.eks bruke de N mest frekvente trekkene.

Trump eller ikke Trump?



- Om *author identification* på tweets fra @realDonaldTrump:
  `https://github.com/raffg/trump-tweet-author-identification/`
- Morsomt eksempel på hva de informative trekkene i en klassifikator kan fortelle oss om et gitt fenomen.

Spam eller ikke-spam?



https://twitter.com/UlvBjerkan/status/947280876479664128

Oblig 2b

- Ut i dag, inn 8/11.
- Naive Bayes for tekstklassifikasjon.
- Sentimentanalyse med to klasser; positiv / negativ.
- Basert på anmeldelser fra NoReC (Norwegian Review Corpus).
- BoW og andre trekk.

Takk for i dag!