Klassifikasjon

ISTx1003 Statistisk læring og Data Science

Sara Martino, Institutt for matematiske fag

Oktober 28 og 29, 2024

Takk til

Mette Langaas (pensumvideoer, kompendium, slides og inspirasjon) Stefanie Muff (slides og inspirasjon) Ingeborg Hem Sørmoen (slides og inspirasjon)

Mine slides bygger på materiale fra både Stefanie, Mette og Ingeborg, som underviste dette kurset i 2021-2023

Praktisk om denne zoom-forelesningen

- Zoom-webinar (som i fellesdelen)
- Forelesningen blir tatt opp og lagt ut
- Dere kan bruke chatten for å sende spørsmål til meg, jeg skal prøve å følge med så godt jeg kan.
- Jeg oppfordrer dere til å sitte sammen med andre studenter og se forelesningen!
- Alle info ligger i wiki sida: https://wiki.math.ntnu.no/istx100y/2024h/1003

Plan - Forelesninger

- **Mandag 21. oktober 15:15-16:00**: Introduksjon
- **Tirsdag 22. oktober 14:15-16:00**: Oppgave 1 (multippel lineær regresjon)
- **Mandag 28. oktober 15:15-16:00** Oppgave 2 (klassifikasjon) og oppgave 3 (klyngeanalyse)
- **Tirsdag 29. oktober 14:15-16:00** Oppgave 3 fors. (?), Q&A og praktisk informasjon
- Alle de 4 forelesningene vil være hel-digitale
- Alle forelesningene tas opp, tentative slides legges ut før forelesningen og slides med notater legges ut etter forelesningen
- Forelesningene vil i all hovedsak utdype prosjektoppgavene. Pensum dekkes av kompendier og tilhørende videoer, og blir ikke repetert her.

Plan - Pensumvideoer

For å få fullt utbytte av forelesningene bør dere ha sett følgende:

- Enkel lineær regresjon (uke 9) innen mandag 21. oktober (i dag)
- Multippel lineær regresjon innen tirsdag 22. oktober (i morgen)
- Klassifikasjon innen 28. oktober
- Klyngeanalyse innen 28. oktober

Plan - Veiledning

- **Digitalt**:
  - Tirsdag 5/11 og 12/11, 14:15-16:00
  - Torsdag 24/10, 31/11, 7/11, 14/11 14:15-16:00
- **Fysisk**::
  - Trondheim: Torsdag 24/10, 31/11, 7/11, 14/11 14:15-16:00 i S5 (SBII)
  - Gjøvik: Kun digitalt
  - Ålesund: Digital. I tilleg gruppene kan møte Siebe til veiledning, avtal tid med han. Se blackboardsiden ISTA: Campus Ålesund for mer informasjon.
- **Forum**: Åpent 24/7, besvares jevnlig av fagteamet

18 november 12:00: Levere prosjekt i Inspera

Vi har laget et svarark i Word dere **skal** bruke. Der står alle spørsmålene i oppgave 2 og 3, og dere skal skrive svarene rett inn der. Når det i oppgaveteksten står at dere skal legge ved et bilde, kan dere ta et skjermbilde av jupyter notatboka (eller bruke den oppgitte kodesnutten for lagring av plott som oppgis i `oppg_1.ipynb`) og lime inn i Word. Der dere skal levere kode kan dere lime inn koden, eller ta skjermbilde. Her skal kun det vi spør etter være med i svarene, og merk at det ofte står en begrensning på antall setninger dere skal svare med.

Oppgave 2 - Klassifisering

Se Jupyter notatboka som heter '`oppg_2.ipynb`', og svarark for utfylling her.

Oppgave 3 - Klyngeanalyse

Se Jupyter notatboka som heter '`oppg_3.ipynb`', og svarark for utfylling her.

Plan tema "Klassifikasjon"

- Læringsmål og ressurser
- Hva er klassifikasjon?
- Trening, validering og testing (3 datasett)
- k-nærmeste nabo (KNN): en intuitiv metode
- Forvirringsmatrise og feilrate for å evaluere metoden
- Logistisk regresjon

Læringsmål

- kunne forstå hva klassifikasjon går ut på og kjenne situasjoner der klassifikasjon vil være en aktuell metode å bruke
- kjenne begrepene treningssett, valideringssett og testsett og forstå hvorfor vi lager dem og hva de skal brukes til
- vite hva en forvirringsmatrise er, og kjenne til begrepene *feilrate* (error rate) og *nøyaktighet* (accuracy)
- forstå tankegangen bak k-nærmeste nabo-klassifikasjon, valg av k
- kjenne til modellen for logistisk regresjon, og kunne tolke de estimerte koeffisientene
- forstå hvordan vi utfører klassifikasjon i Python
- kunne besvare problem 2 av prosjektoppgaven

Læringsressurser

Tema Klassifikasjon:

- Kompendier (laget av Mette Langaas)
- Tilhørende videoer (laget av Mette Langaas)
- Notat om interaksjoner (laget av Ingeborg Hem Sørmoen)
- Zoom-forelesningene (med slides)

Kompendier og tilhørende videoer ligger allerede på *Wiki sida*. Zoom-forelesningene og slides legges også ut. Mange andre tilgjengelige kilder på nett også, men jeg garanterer ikke at de dekker de riktige delene av temaene.

Prosjektet

Dere skal bruke mye Python, og vi har laget notatbøker til dere. Jeg anbefaler alle å gjøre prosjektet lokalt på egen maskin, da huben er litt ustabil. Hvordan? Se Blackboard.

- Oppgave 1: Regresjon (50%, prosjektrapport)
- Oppgave 2: Klassifikasjon (30%, 4 deloppgaver, 15 spørsmål)
- Oppgave 3: Klyngeanalyse (20%, 4 deloppgaver, 10 spørsmål)

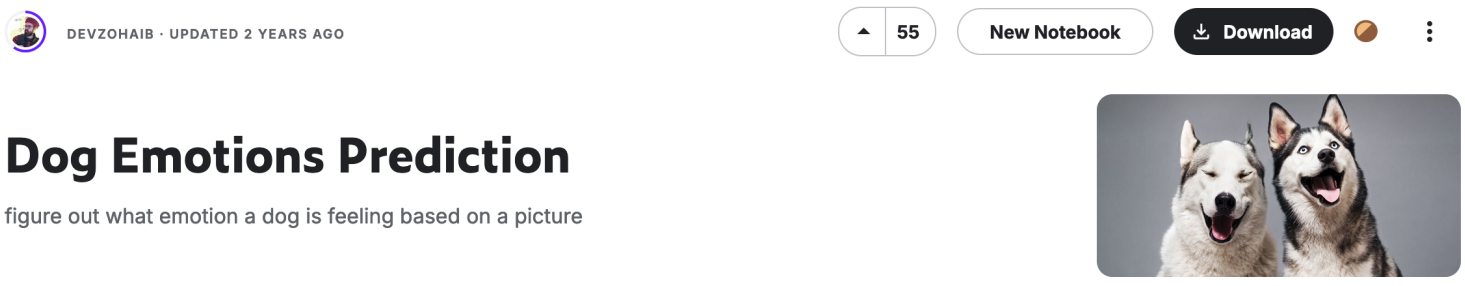
Klassifikasjon – hva er det?

- Mål:
  - tilordne en ny observasjon til en av flere *kjente* klasser
  - lage en klassifikasjonsregel
  - estimere sannsynligheten for at en ny observasjon tilhører de ulike klassene

For hver av de uavhengige observasjonene $i = 1, \ldots, n$ har vi

- Forklaringsvariabler $(x_{1i}, x_{2i}, \ldots, x_{pi})$
- En kategorisk responsvariablel y_i.

Eksempel - Dog emotion Prediction



https://www.kaggle.com/datasets/devzohaib/dog-emotions-prediction

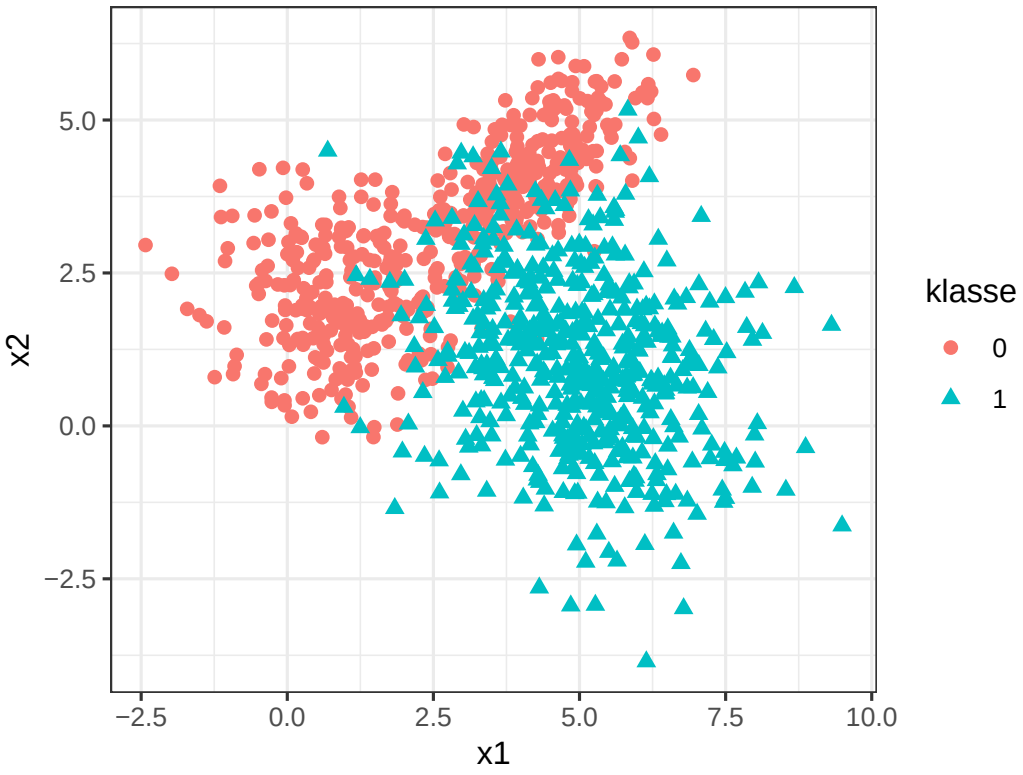
Flere eksempler

- Hvilke tall er handskrevet på brevet? Og hva er sannsynligheten for de ulike tallene (=klasser)?

- Kommer en gitt kunde til å betale tilbake lånet sitt?
- Prognose om noen blir syk (hjertesykdom, kreft...) og sannsynligheten for det.

Syntetisk eksempel

- Et datasett med folgende struktur: (x_{1i}, x_{2i}, y_i).



Spørsmål vi vil besvare: Hva er de beste klassifikasjonsgrensene?
Vi skal laere om to teknikker:

- k-nærmeste-nabo-klassifikasjon
- logistisk regresjon

Data

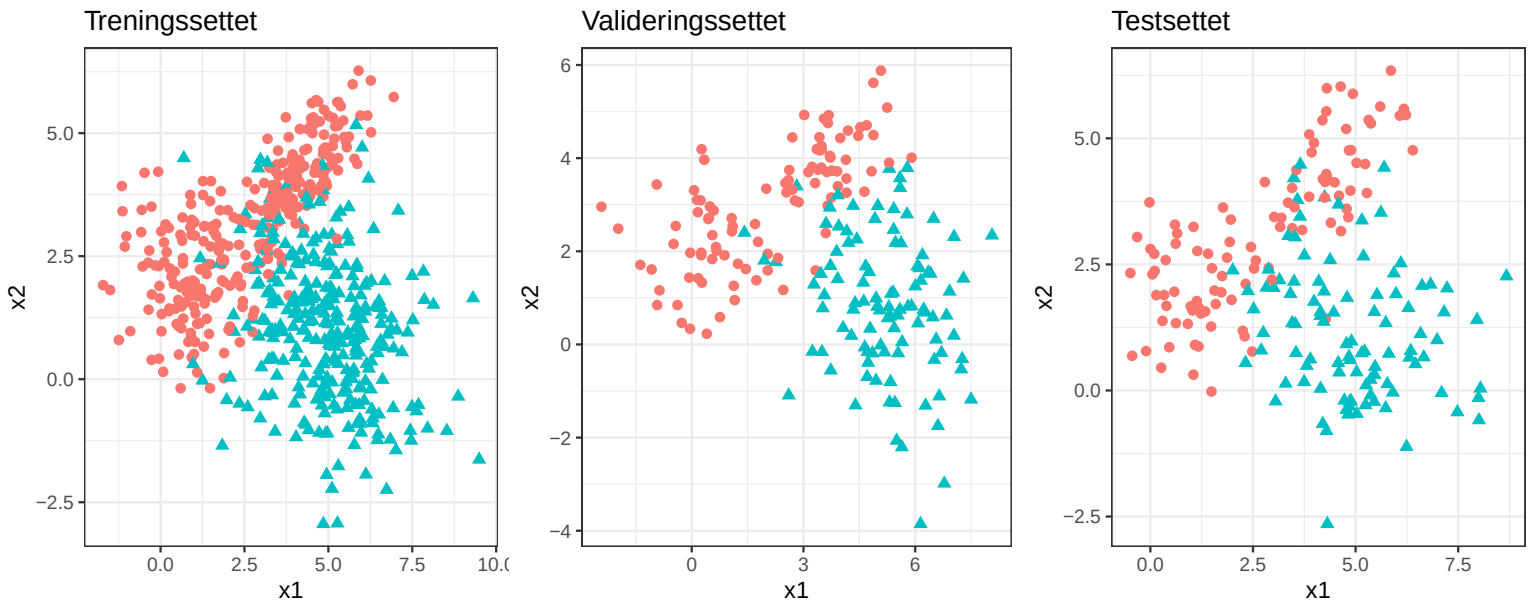
Datasett: $(x_{1i}, x_{2i}, \ldots, x_{pi}, y_i)$

Vi må dele datasettet i tre deler:

- Treningsdata
- Valideringsdata
- Testdata

Hvorfor det?

Trening-, validering- og testsett for syntetiske data



For å *finne* klassifikasjonsreglen bruker vi bare treningsdataene.

k-nærmeste-nabo-klassifikasjon

Algoritmen:

1) Ny observasjon: $x_0 = (x_{1,0}, x_{2,0}, \ldots, x_{p,0})$. Hvilken klasse bør denne klassifiseres til?

2) Finn de k nærmeste naboene til observasjonen i treningssettet.

3) Sannsynligheten for at den nye observasjonen tilhører klasse 1 anslår vi er andelen av de k nærmeste naboene som tilhører klasse 1. Ditto for de andre klassene.

4) Klassen til den nye observasjonen er den som har størst sannsynlighet. Det blir det samme som å bruke flertallsavstemming.

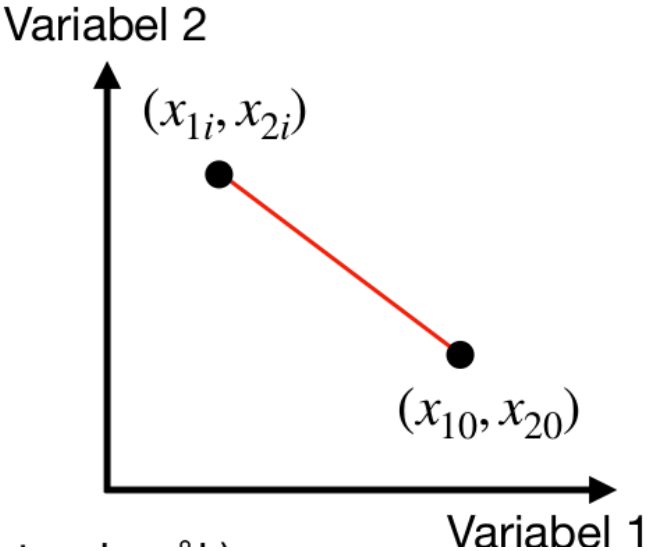
k-nærmeste-nabo-klassifikasjon

Tre spørsmål:

- Hva betyr "nærmest"? Vi trenger en definision av avstand.
- Hvilke verdier kan k ha?
- Hvordan bestemmer vi k?

Hva betyr nærmest?

Nærmest er definert ved å bruke Euklidsk avstand.

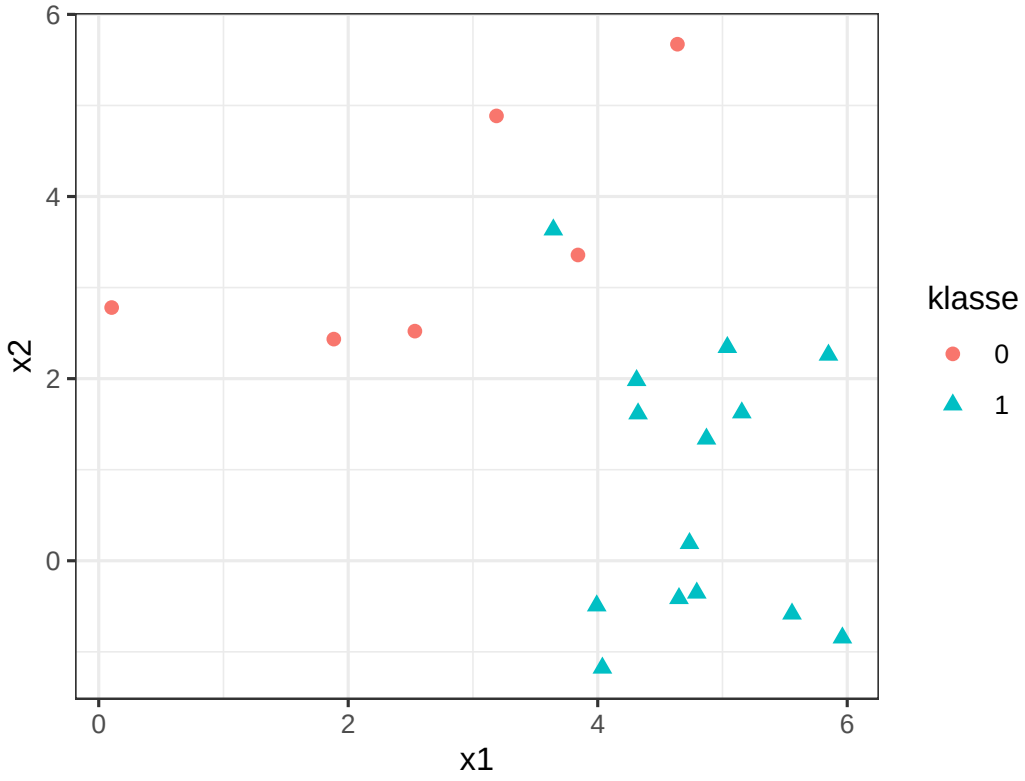


Euklidsk avstand:

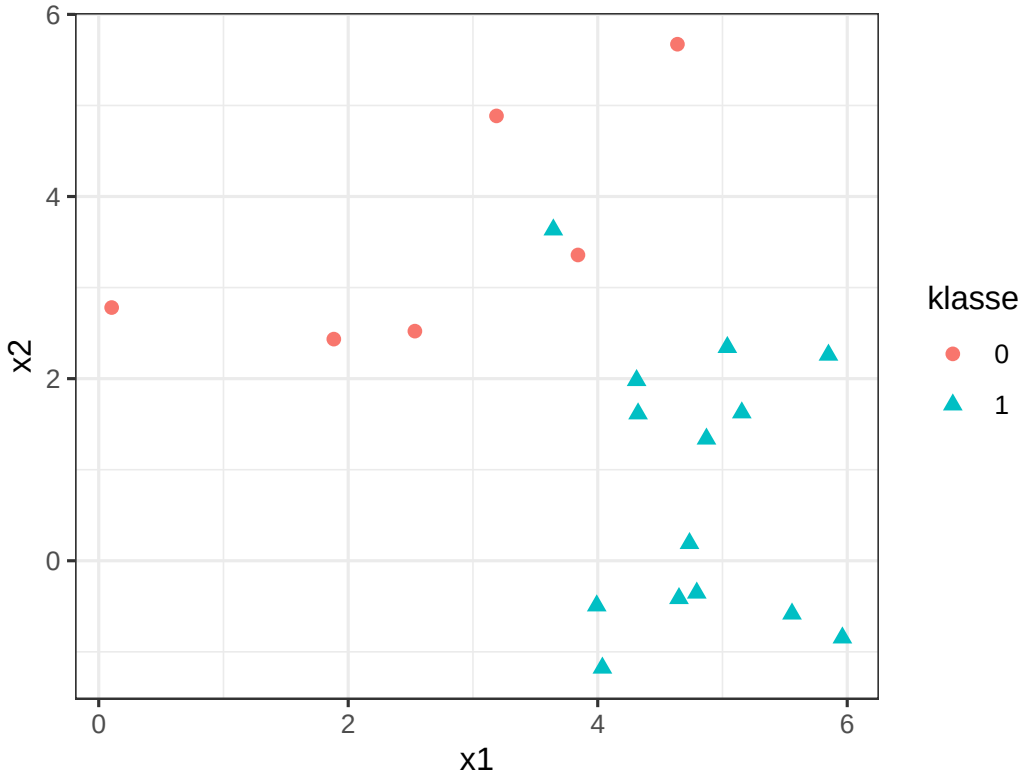
$$D_E(i,0) = \sqrt{\sum_{j=1}^{p}(x_{ji} - x_{j0})^2}$$

Andre avstandsmål kan også brukes, men Euklidsk avstand er mest vanlig.

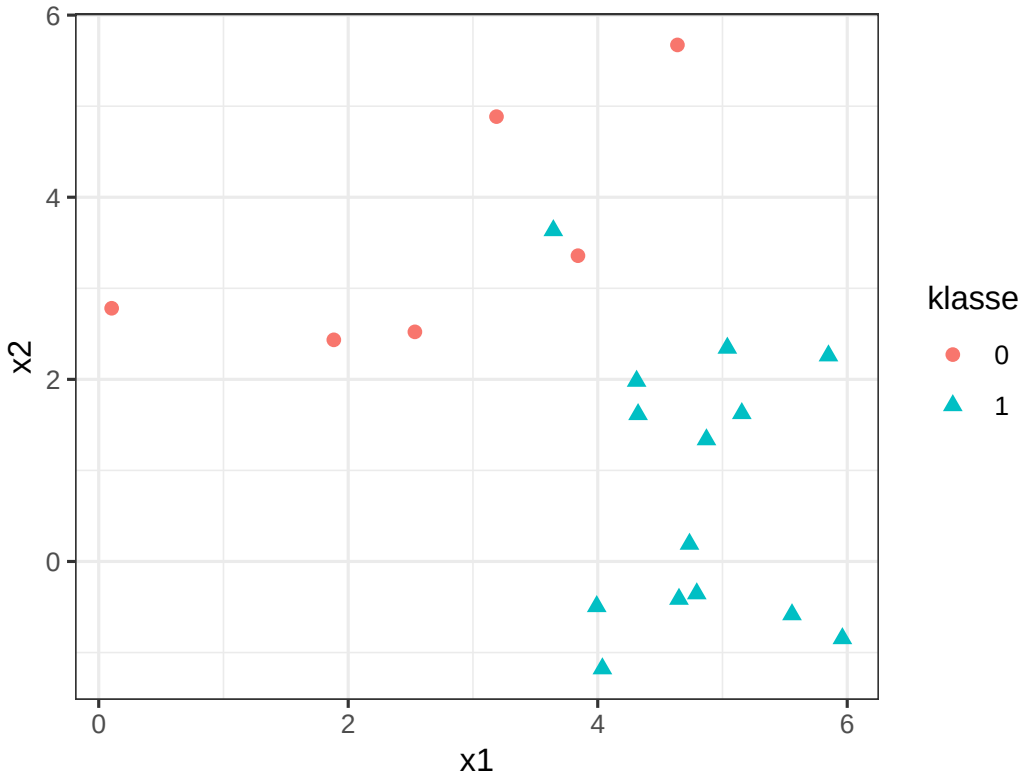
Nærmeste naboer



Nærmeste naboer

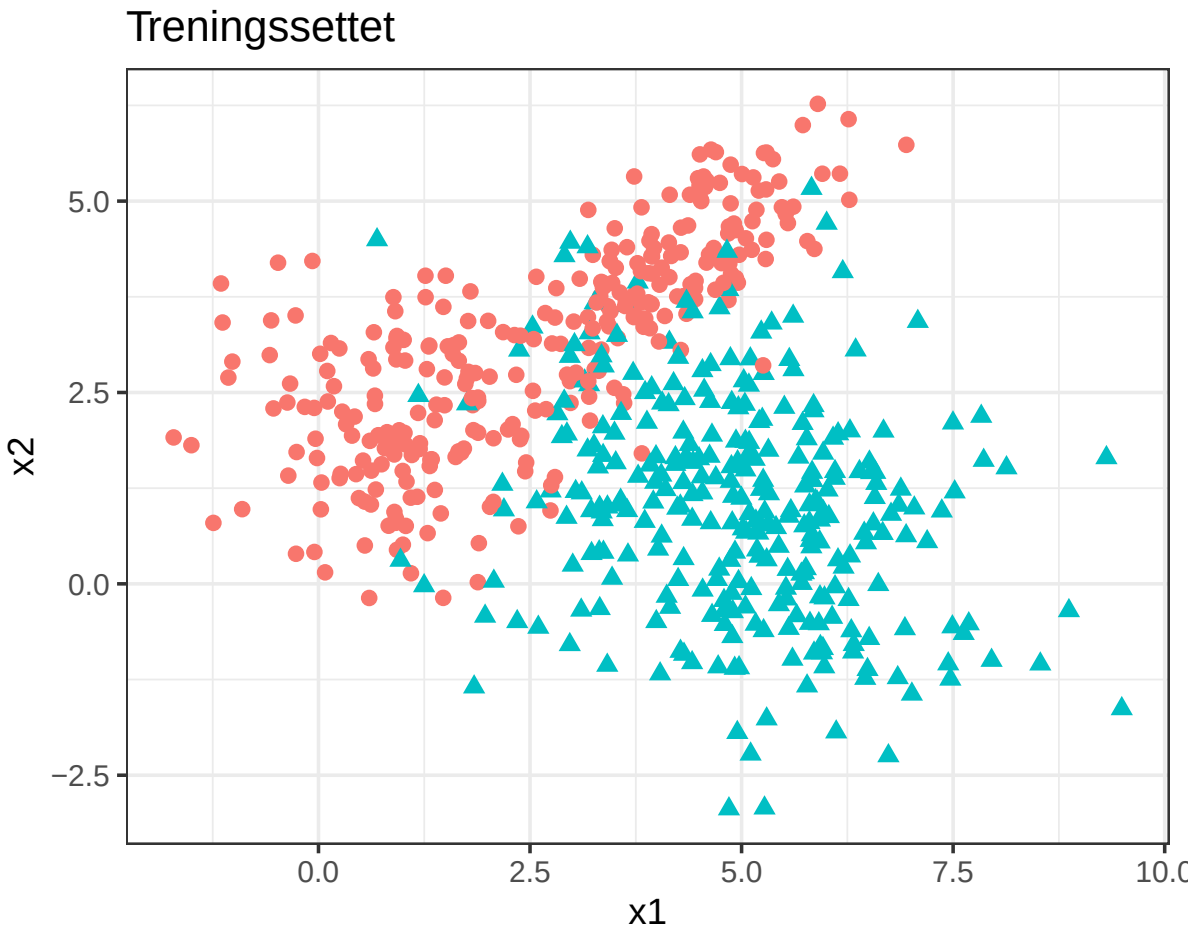


Tegne klassegrenser



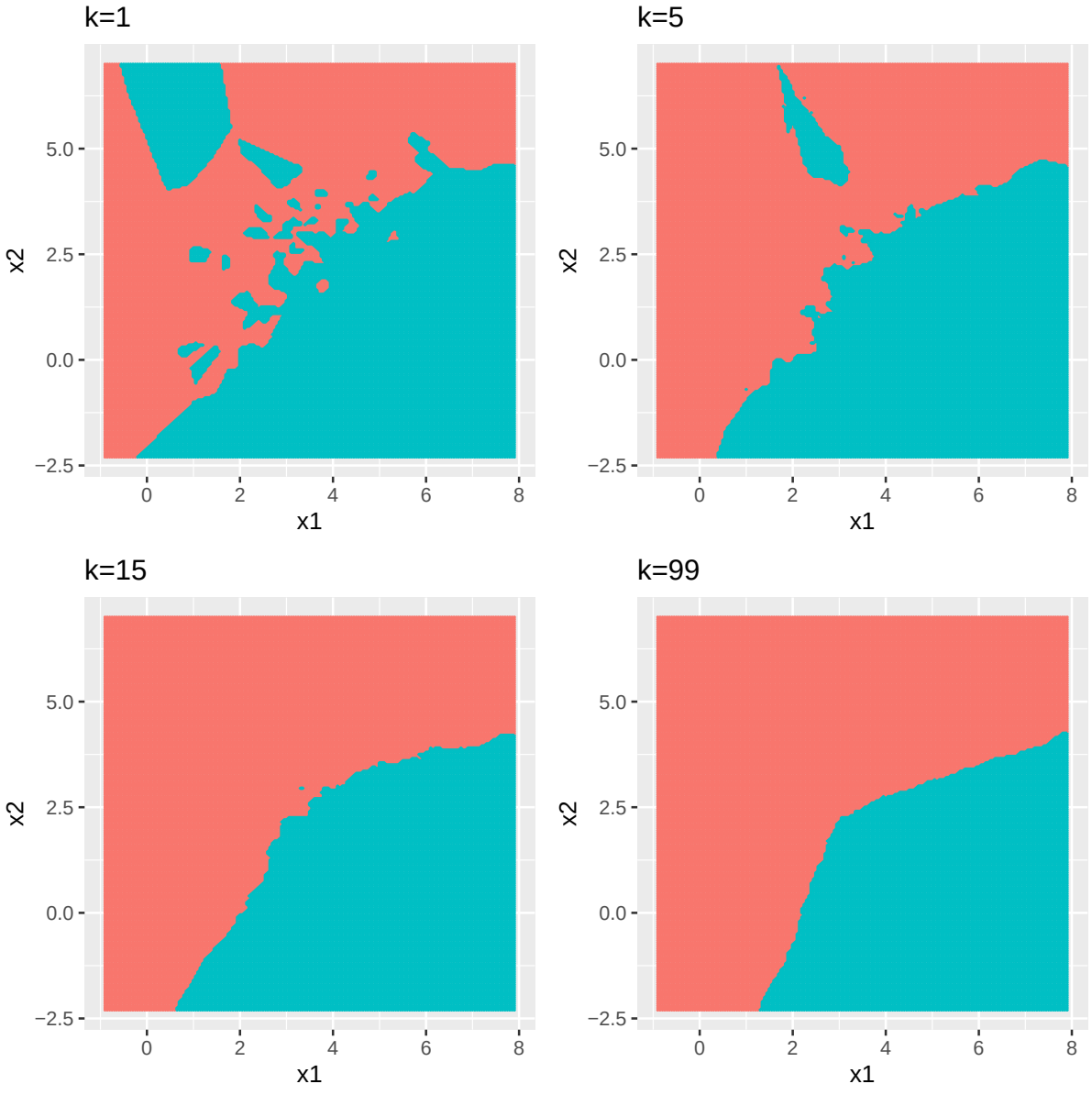
Hvordan ser klassegrensene ut?

Husk: Vi bruker treningssettet for å finne klassifikasjonsreglen:



Hvordan ser klassegrensene ut?

Svar: It depends!



Men vent... hvordan velger vi k da?

Vi så jo at

- hvis man velger k for lite, da blir grensen *for fleksibel*.

Men:

- hvis man velger k for stor, kan grensen bli *for ufleksibel*.

Det er noe som kalles en *trade-off*, og vi skal bruk valideringssettet for å få en idé om kvaliteten i klassifikasjonen.

Forvirringsmatrise

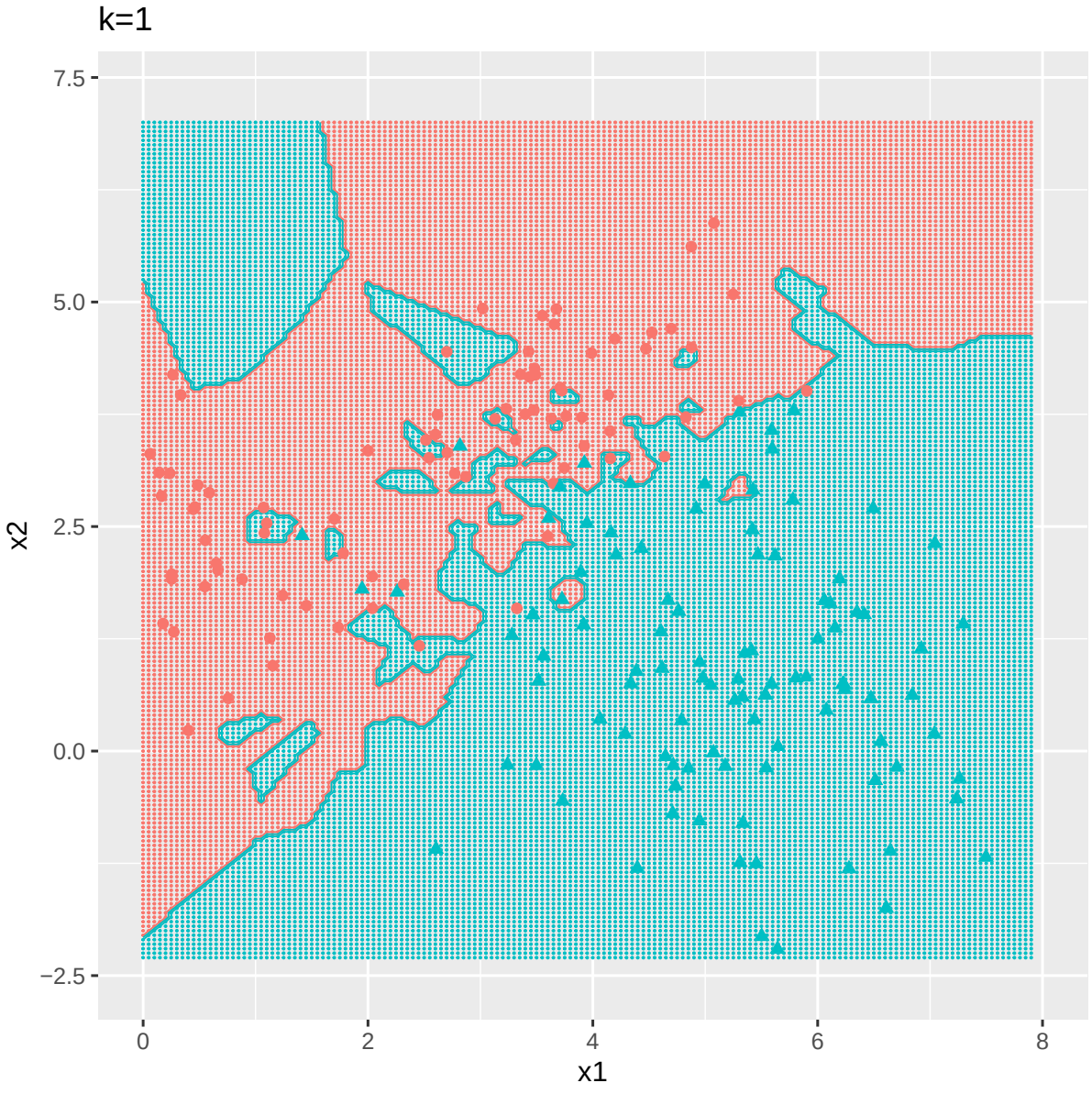
Forvirringsmatrise med to klasser:

	Sann klasse 0	**Sann klasse 1**
Predikert klasse 0	Sann negativ (TN)	Falsk negativ (FN)
Predikert klasse 1	Falsk positiv (FP)	Sann positiv (TP)

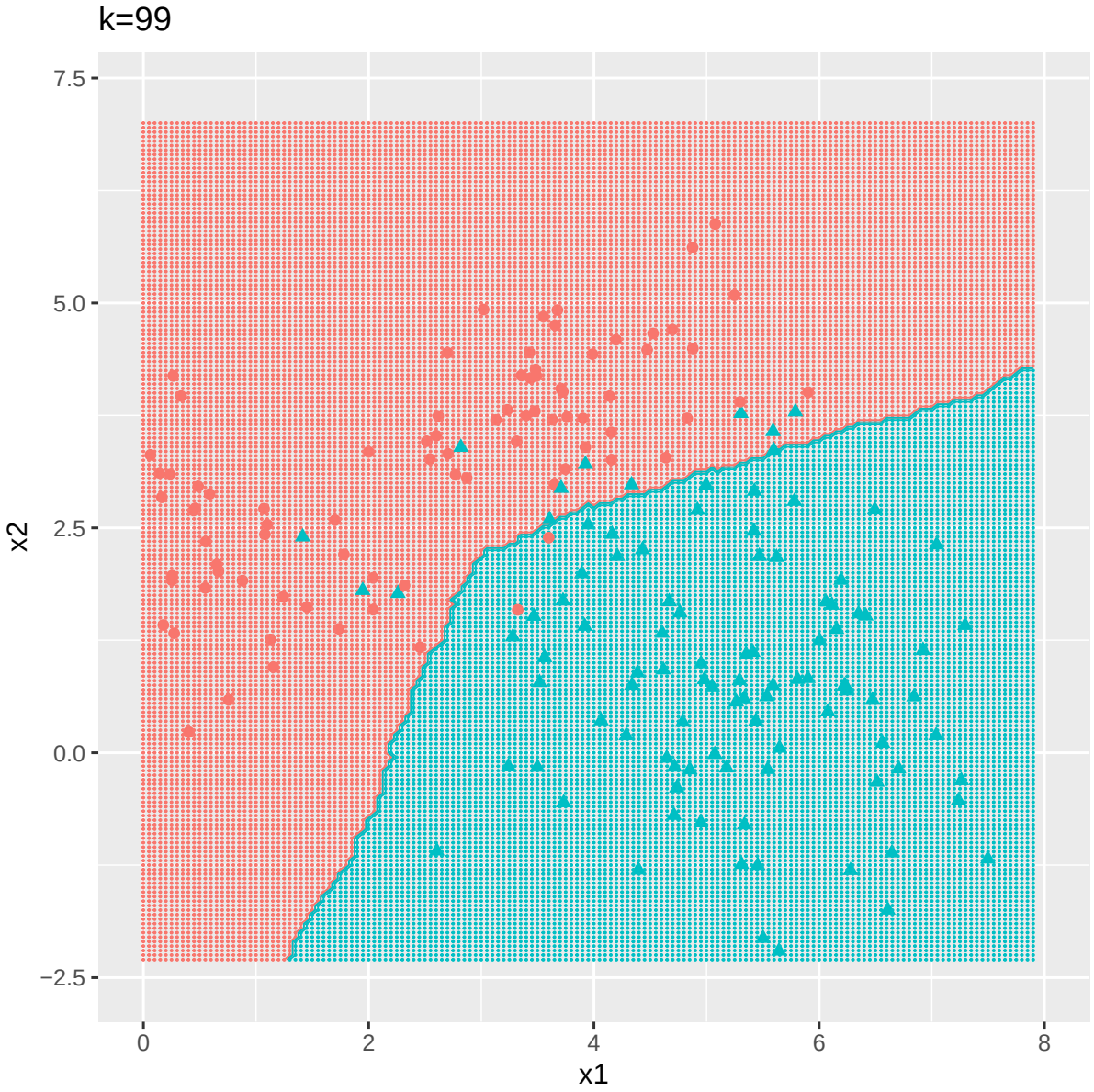
(Fra Mette Langaas)

- **Feilrate**: andel feilklassifiserte observasjoner.
- Derfor: Vi velger den k som minimerer feilraten på *valideringssettet* (ikke trainingssettet!).

Marker og tell antall gale klassifiseringer på valideringssettet ($k = 1$):

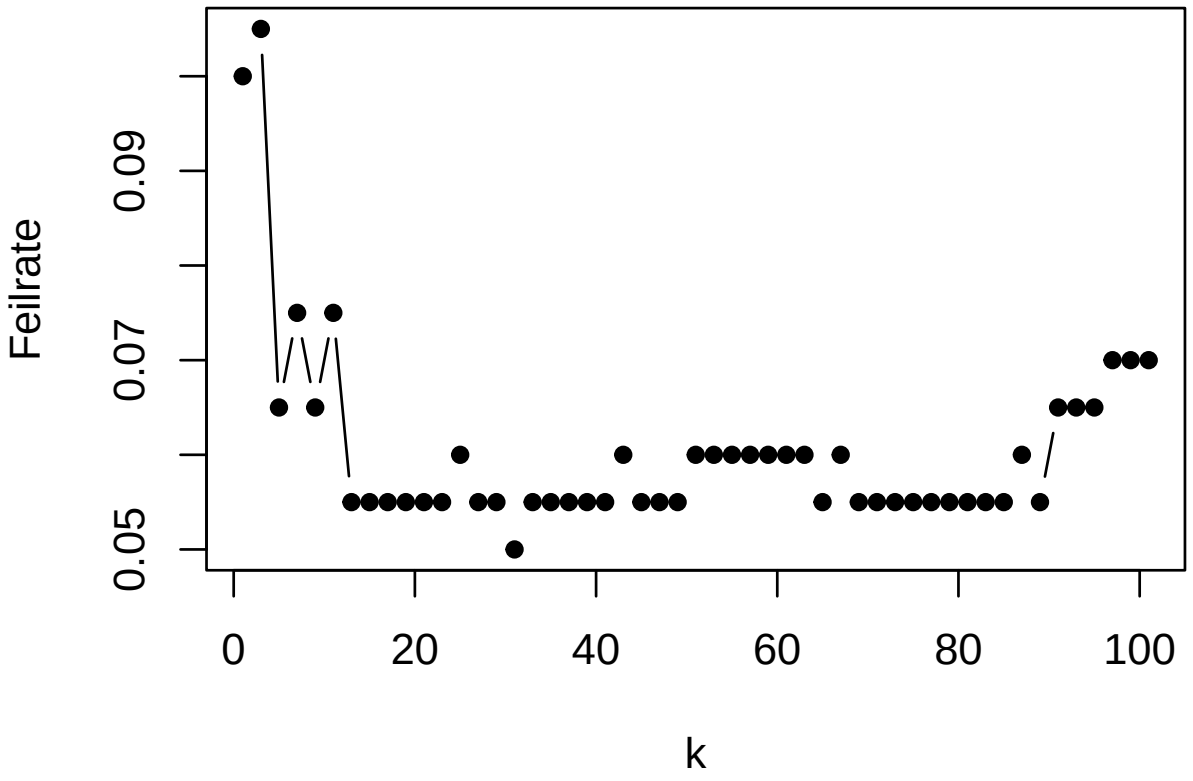


Igjen, men nå med $k = 99$:

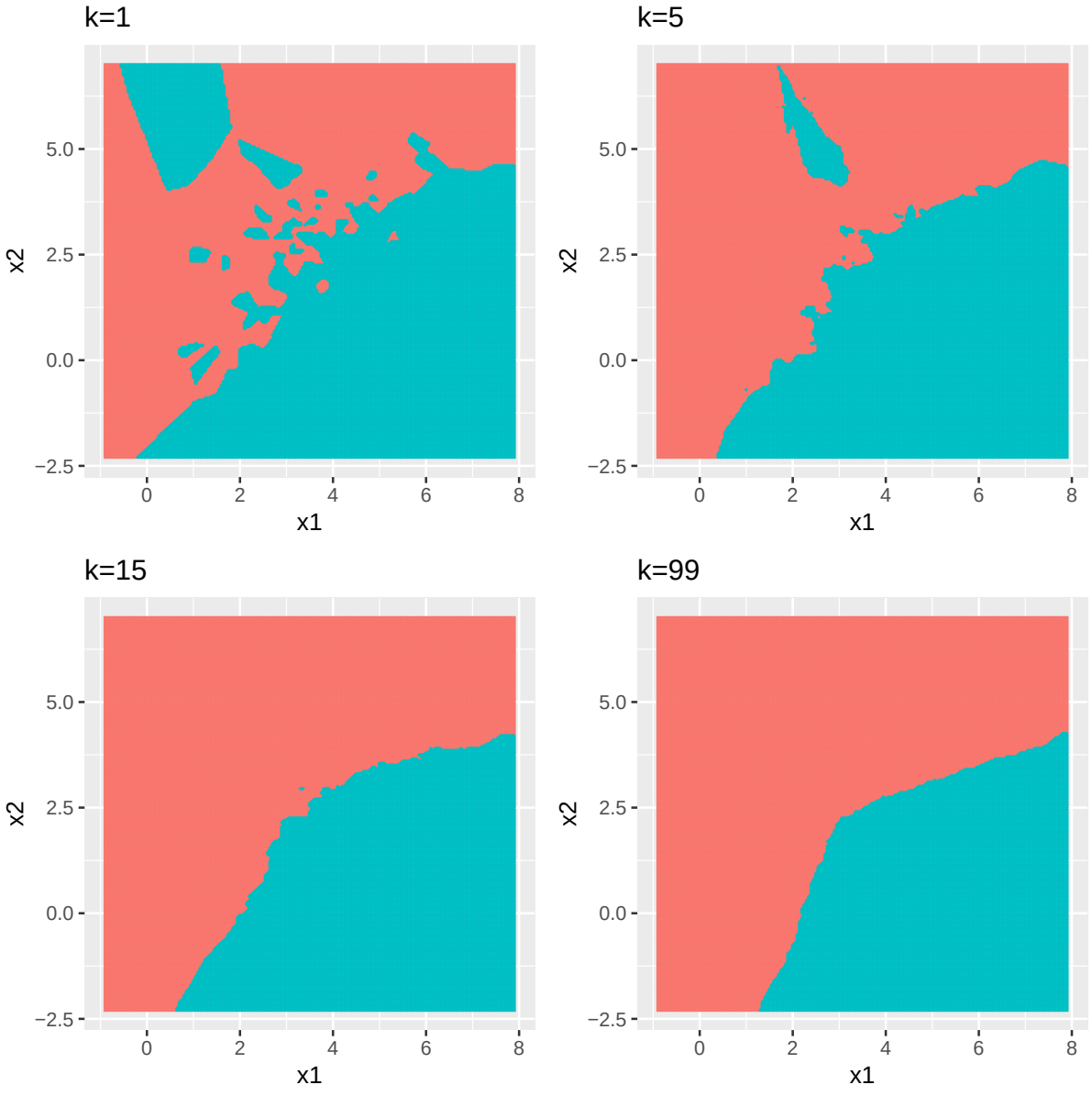


Feilrate i valideringssettet

Det kan vi nå systematisk gjøre med alle $k = 1, 3, ..., k \leq n$:



Klassegrensene ser ganske likt ut for $k \geq 15$:



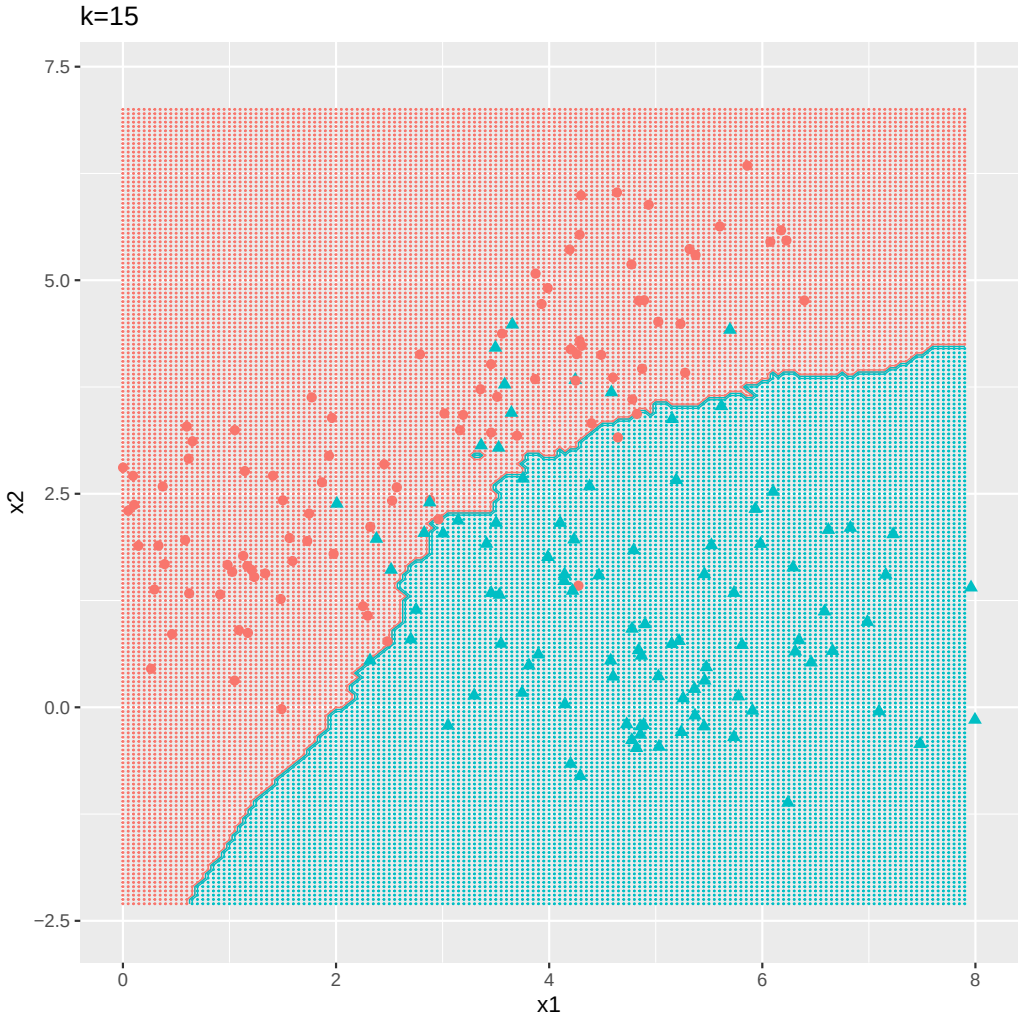
Data

Husk at vi hadde *tre* deler i datasettet:

- **Treningssett**: For å lage en klassifikasjonsregel
- **Valideringssett**: For å finne på optimale hyperparametere
- **Testsett**: For å evaluere regelen på fremtidige data

Nå kan vi bruke testsettet for å *finne feilraten* for fremtidige data.

Feilraten på testsettet for KNN med $k = 15$ (egentlig er $k = 31$ best, men det gjør ikke en stor forskjell):



Feilraten er 0.09.

Logistisk regresjon - mål

- Forstå modellen
- Tolke estimerte koeffisienter
- Hypotesetest og p-verdi
- Velge mellom modeller
- Kunne utføre dette i Python

Logistisk regresjon

- Kan bare handtere *to klasser* $y_i \in \{0, 1\}$.

- Vi antar at Y_i har en **Bernoulli fordeling** med suksessannsynlighet p_i, derfor:

$$y_i = \begin{cases} 1 \text{ med sannsynlighet } p_i, \\ 0 \text{ med sannsynlighet } 1 - p_i. \end{cases}$$

- **Mål**: For forklaringsvariabler $(x_{1i}, x_{2i}, \ldots, x_{pi})$ vi vil estimere $p_i = \Pr(y_i = 1 \mid x_1, \ldots, x_p)$.

Enkel logistisk regresjon

- Vi antar at responsen Y_i er binomisk fordelt med suksessannsynlighet p_i.
- **Ideen**: å koble p_i sammen med forklaringsvariablen med en logistisk funksjon:

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_{1i})}{1 + \exp(\beta_0 + \beta_1 x_{1i})} \ .$$

- Denne verdien p_i er alltid mellom 0 og 1, som den skal være.

Klassifikasjonsregel

I kredittkorteksempelet er x_{1i} balansen til person i.

Klassifikasjonsregel: $p \geq 0.5$ vs $p < 0.5$.

Tolkning av β_1: odds

- Odds er definert som ratioen mellom sannsynlighet for suksess (p) og fiasko ($1-p$):

$$odds = \frac{p}{1-p}$$

- Noen eksempler (og zoom poll):

- I logistisk regresjon: Hvis vi øker verdien til kovariaten fra x_{i1} til $x_{i1}+1$ så blir den nye oddsen lik den gamle multiplisert med e^{β_1}.

Oppgave 2

Data fra de fire øverste divisjonene i engelsk herrefotball, 22-23.

Dere skal predikere om hjemmelaget vinner en gitt kamp basert på antall skudd på mål, cornere og forseelser (regelbrudd som fører til frispark eller straffespark).

Logistisk regresjon og l-nærmeste-nabo.

- **Respons**: 1 om hjemmelaget vant, 0 ellers
- **Forklaringsvariabler**: Forskjell (hjemme-borte) i
  - skudd på mål
  - cornere
  - forseelser

Oppgave 2

15 korte, konkrete spørsmål.

Omtrent all kode dere trenger står i en Jupyter notatbok.

Bruk svarark (enklere for både dere og oss).

2a.1)	Hvorfor ønsker vi å dele dataene inn i trening-, validering- og test-sett?
Svar	

2a.2)	Hvor stor andel av dataene er nå i hver av de tre settene? Ser de tre datasettene ut til å ha lik fordeling for de tre forklaringsvariablene og responsen?
Svar	

2a.3)	La oss si at vi hadde valgt League 1 og 2 som treningssett, Championship som valideringssett, og Premier League som testsett. Hvorfor hadde dette vært dumt?
Svar	

Teller 30% av prosjektet, slik at hvert spørsmål teller 2% hver.

Videre denne uken

- Se på videoene om Klyngeanalyse
- Les i kompendiet.
- Begynn å jobbe med prosjektoppgavene.
- Se her for mer informasjon:
  https://wiki.math.ntnu.no/istx100y/2024h/1003