MA8701 Advanced methods in statistical inference and learning

L1: Introduction and core concepts

Mette Langaas

~~1/8/23~~

09.01.2023 Class notes

ME

Write an enthusiastic email inviting students to the first lecture today in S21 at 10.15 in a phd course in statistical learning and inference.

ChatGPT helps me!

ME

For the same course, write my welcome introduction to the physical class

Good morning everyone, and welcome to the first lecture of our PhD course in Statistical Learning and Inference. My name is [Your Name], and I will be your instructor for this course.

I am thrilled to have all of you in this class and am looking forward to a semester filled with engaging discussions, hands-on exercises, and, of course, a lot of learning.

and more

As some of you may already know, this course will cover a wide range ~~of~~ topics in some statistical modeling, including linear regression, classification algorithms, and advanced Bayesian methods. We will be using a variety of tools and techniques, so be prepared to get your hands dirty with some data analysis.

I believe that the best way to learn is through active participation, so I encourage you all to ask questions, contribute to discussions, and collaborate with your peers. I am here to support you every step of the way and am excited to see what we can accomplish together.

So, without further ado, let's dive in and get started!

Course philosophy

"Important new algorithms often arise outside the world of professional statisticians: neural nets, support vector machines, and boosting are three famous examples. None of this is surprising. New sources of data, satellite imagery for example, or medical microarrays, inspire novel methodology from the observing scientists. The early literature tends toward the enthusiastic, with claims of enormous applicability and power."

In the second phase, statisticians try to locate the new methodology within the framework of statistical theory. In other words, they carry out the statistical inference part of the cycle, placing the new methodology within the known Bayesian and frequentist limits of performance This is a healthy chain of events, good both for the hybrid vigor of the statistics profession and for the further progress of algorithmic technology."

Efron and Hastie (2016) (page 10).

the aim of this course

Learning

Learning outcome

Will write down such aspects for each topic!

1. Knowledge

- Understand and explain the central theoretical aspects in statistical inference and learning.
- Understand and explain how to use methods from statistical inference and learning to perform a sound data analysis.
- Be able to evaluate strengths and weaknesses for the methods and choose between different methods in a given data analysis situation.

Take home message in MA8701 in 2021 according to the students

LINK (see www)

in groups

one compulsory data analysis project

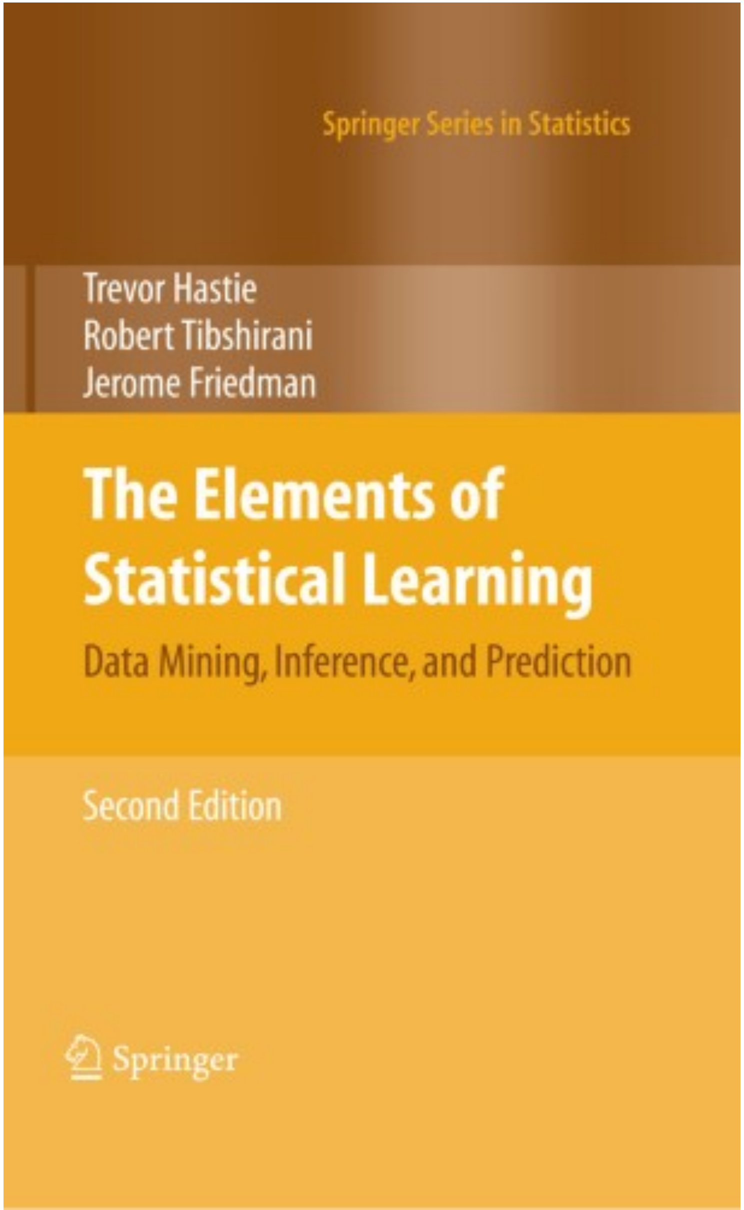
2. Skills

Be able to analyse a dataset using methods from statistical inference and learning in practice (using R or Python), and give a good presentation and discussion of the choices done and the results found.

one compulsory article presentation!

3. Competence

- The students will be able to participate in scientific discussions, read research presented in statistical journals.
- They will be able to participate in applied projects, and analyse data using methods from statistical inference and learning.

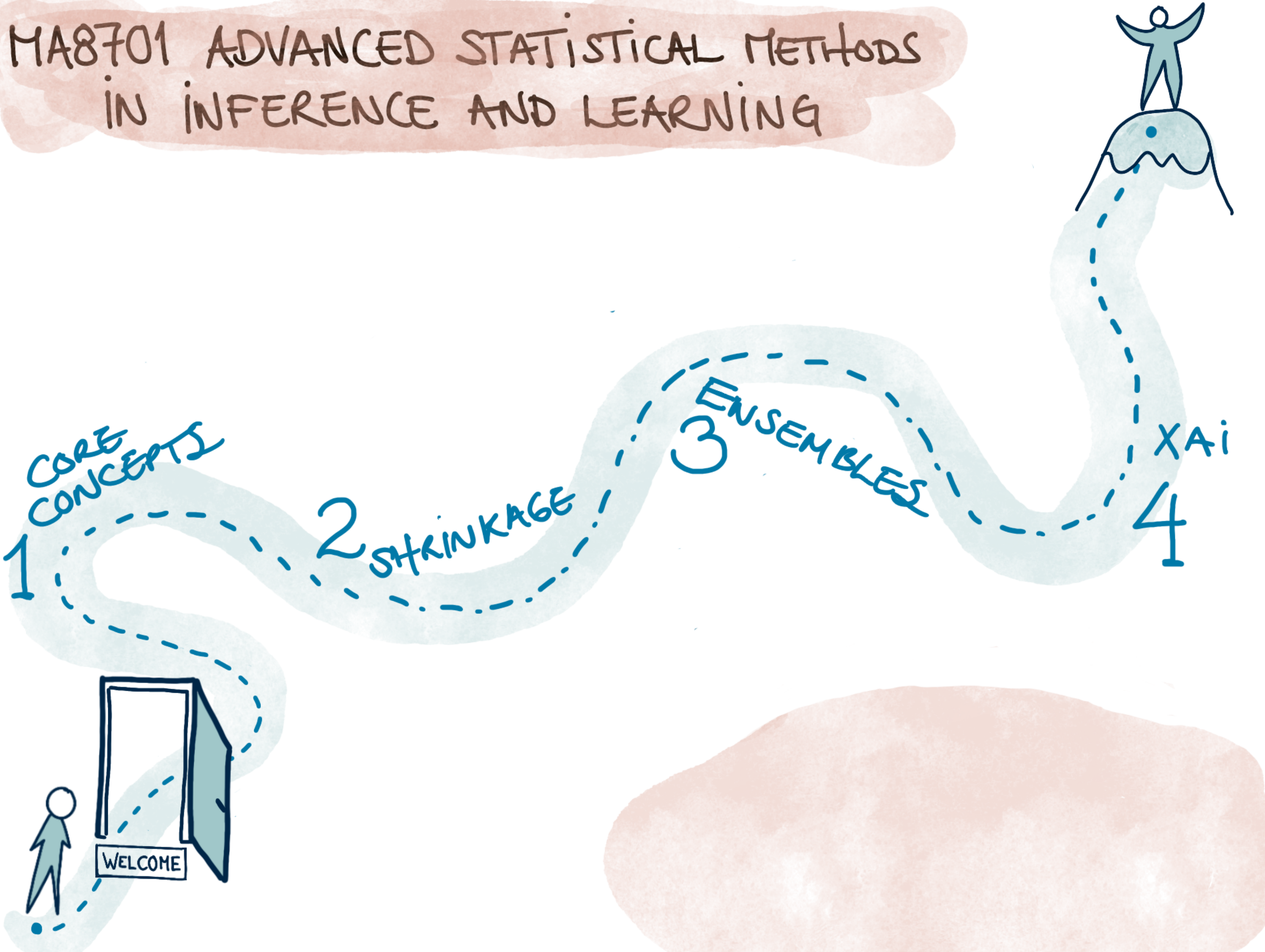


OUR BIBLE

but is from 2001 / 2009
2nd
ed

so also
need to add more
recent literature

MA8701 ADVANCED STATISTICAL METHODS IN INFERENCE AND LEARNING



Part 1: Core concepts [3 weeks]

Sort out assumed background knowledge, and learn something new

- Notation
- Repetition of core concepts (regression and classification)
- Statistical decision theoretic framework (partly new) ESL 2.4
- Model selection and model assessment - including bias-variance trade-off (mostly new) ESL 7
- Handbook of Missing Data Methology (parts of Chapters 11-12, partly 13) and Flexible Imputation of Missing Data (parts of Chapters 2-4)

NEW

missing from our other stats courses

not decided on literature yet!

L1, L2 ++

Part 2: Shrinkage and regularization [3 weeks]

or "Regularized linear and generalized linear models", with focus on the ridge and lasso regression (in detail).

- ESL 3.2.3,3.4, 3.8, 4.4.4.
- Hastie, Tibshirani, Wainwright (HTW): "Statistical Learning with Sparsity: The Lasso and Generalizations". Selected sections from Chapters 1,2,3,4,6.
- Selective inference (articles)

Important that we know linear models (LM) and GLM (in particular logistic regression) BEFORE we add more in this part!

Part 3: Ensembles [4 weeks]

- trees, bagging and random forests
- xgboost
- general ensembles (including super learner) autoML
- hyper-parameter tuning I need help to make sure new stuff included!

Selected Chapters in ESL (8.7, 8.8, 9.2, parts of 10, 15, 16) and several articles.

Part 4: XAI [2 weeks]

Lectured by Kjersti Aas https://www.nr.no/~kjersti/.
Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, Molnar (2019), with the following topics:

- LIME,
- partial dependence plots,
- Shapley values,
- relative weights and
- counterfactuals.

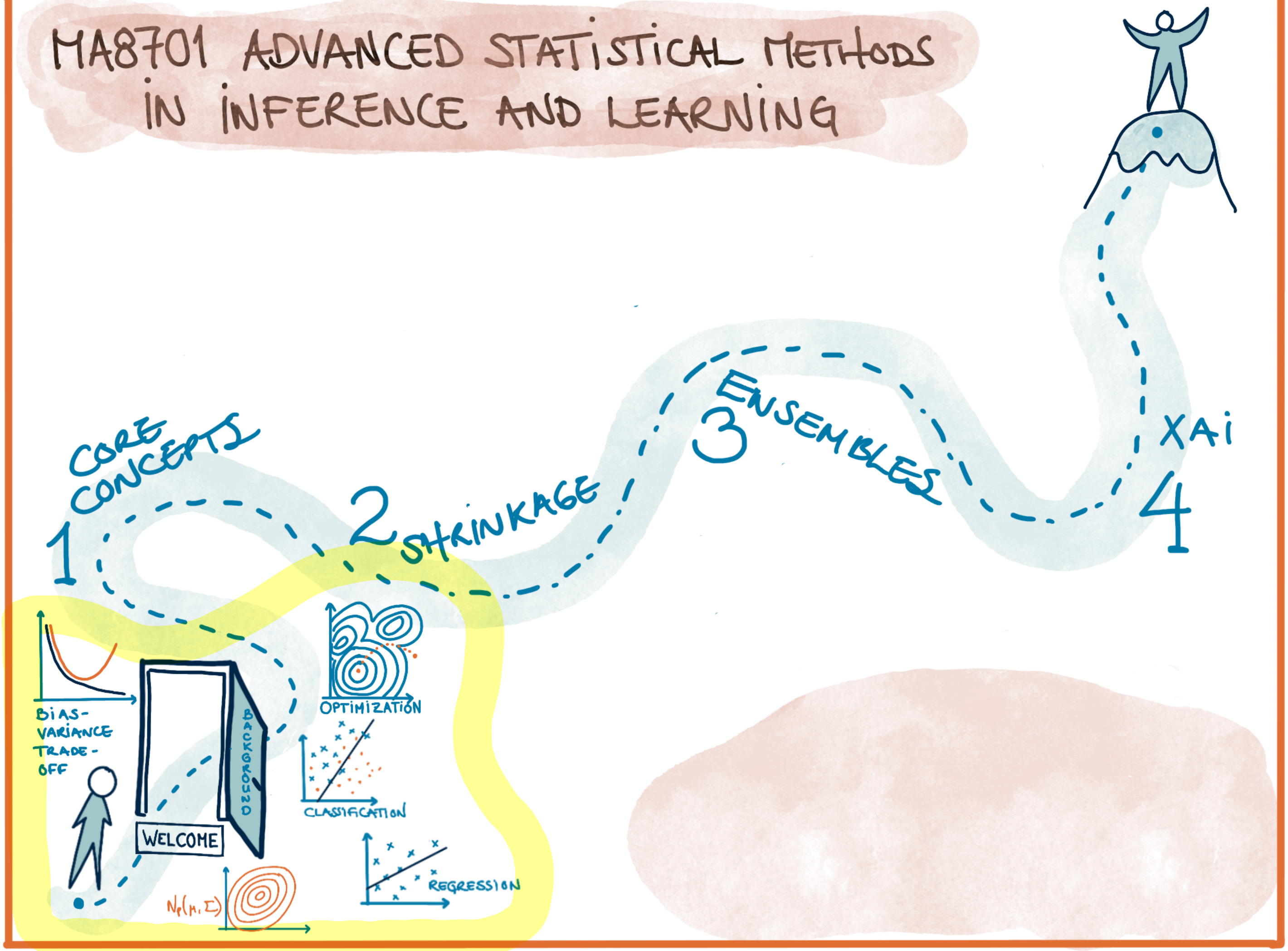
Part 5: Closing [2 weeks]

buffer and article presentations

Some observations about the course

- Mainly a frequentist course, but some of the concepts and methods have a Bayesian version that might give insight into why and how the methods work, then Bayesian methods will be used.
- Focus is on regression and classification, and unsupervised learning is not planned to be part of the course.
- The required previous knowledge is listed because this is a PhD-course designed for statistics students. The background make the students go past an overview level of understanding of the course parts (move from algorithmic to deep understanding).

MA8701 ADVANCED STATISTICAL METHODS IN INFERENCE AND LEARNING



In particular

"Required" previous knowledge

- TMA4267 Linear statistical methods Linear algebra & multivariate normal
- TMA4268 Statistical learning ← lingo
- TMA4295 Statistical inference ← properties of param. est, CI & hyp.
- TMA4300 Computer intensive statistical methods Bayes + bootstrap Gibbs sampling
- TMA4315 Generalized linear models ← logistic regression
- Good understanding and experience with R, or with Python, for statistical data analysis.
- Knowledge of markdown for writing reports and presentations (Rmarkdown/Quarto, Jupyther).
- Skills in group work - possibly using git or other collaborative tools.

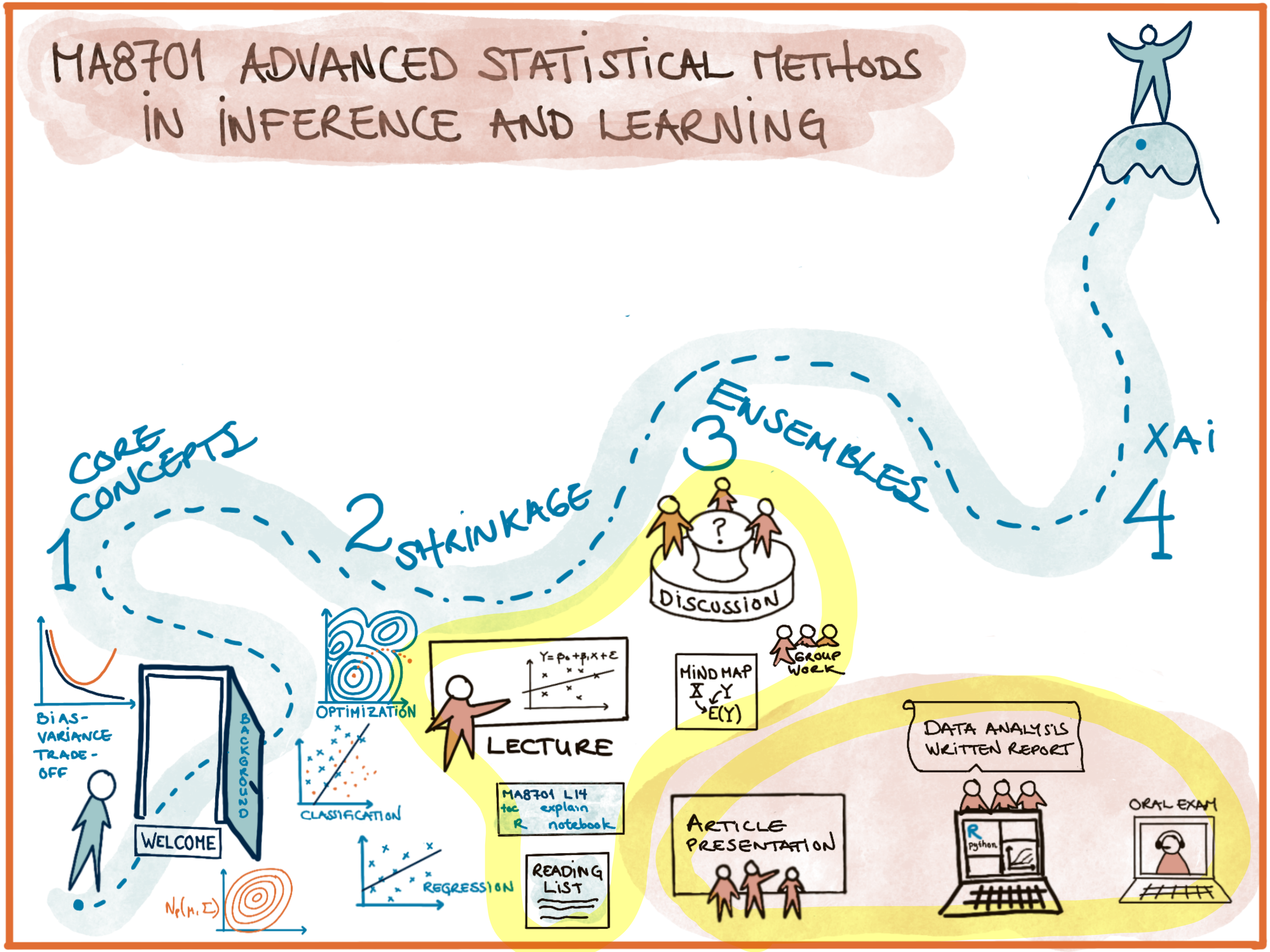
Course elements

Course wiki at https://wiki.math.ntnu.no/ma8701/2023v/start

- Lectures
- Problem sets to work on between lectures.
- Office hours and/or mattelab.math.ntnu.no?
- Study techniques (share)
- Ethical considerations
- Compulsory work
- Final individual oral exam in May

FORGOT TO ASK!

The learning material is also available at https://github.com/mettelang/MA8701V2023.



MA8701 ADVANCED STATISTICAL METHODS
IN INFERENCE AND LEARNING
1
CORE CONCEPTS
2
SHRINKAGE
3
ENSEMBLES
4
XAI
BIAS-VARIANCE TRADE-OFF
WELCOME
BACKGROUND
$N_p(\mu, \Sigma)$
OPTIMIZATION
CLASSIFICATION
REGRESSION
$Y = \beta_0 + \beta_1 x + \varepsilon$
LECTURE
MA8701 L14
explain
R notebook
READING LIST
?
DISCUSSION
MIND MAP
X Y
E(Y)
GROUP WORK
ARTICLE PRESENTATION
DATA ANALYSIS WRITTEN REPORT
R
python
ORAL EXAM

Class activity

Aim: get to know each other - to improve on subsequent group work!

```
while (at least one student not presented)
   lecturer give two alternatives, you choose one.
   lecturer choose a few students to present their view
   together with giving their name and study programme
   (and say if they are looking for group members)
```

We were Kenneth, Nora, Elias, Jane, Jacob,
WaiYen, Nils, Caroline, Philip, Jonas
Fanny, Øyvind, Sebastian.

Learning methods and activities

Herbert A. Simon (Cognitive science, Nobel Laureate): *Learning results from what the student does and thinks and only from what the student does and thinks. The teacher can advance learning only by influencing what the student does to learn.*

Tentative plan for part 1

(progress may be faster or slower than indicated)

L1

Notation, regression and statistical theoretic framework

- Notation (ESL Ch 2.2)
- Regression - should not be new (ESL Ch 3, except 3.2.3, 3.2.4, 3.4, 3.7, 3.8)
- Statistical decision theoretic framework for regression (ESL 2.4)

For 18.01
homework: exercises in the end of the L1.html file
+ browse through the L2.html

L2

Continue with the same framework but for classification, if time also bias-variance trade-off

- Classification - should not be new (ESL Ch 4.1-4.5, except 4.4.4)
- Statistical decision theoretic framework for classification (ESL 2.4)
- and the bias-variance trade-off ← and maybe more from ch 7

W2

L3-4: Then, cover new aspects for

- Model selection and assessment (ESL Ch 7.1-7.6, 7.10-7.12)

W3

L5-6

- How to handle missing data in data analyses

Core concepts

Notation

(mainly from ESL 2.2)

We will only consider supervised methods.

- Response Y (or G): dependent variable, outcome, usually univariate (but may be multivariate)
  - quantitative Y: for regression
  - qualitative, categorical G: for classification, some times dummy variable coding used (named one-hot coding in machine learning)
- Covariates $X_1, X_2, \ldots, X_p$: "independent variables", predictors, features
  - continuous, discrete: used directly
  - categorical, discrete: often dummy variable coding used

We aim to construct a rule, function, learner: $f(X)$, to predict Y (or G).

$\hat{G}(X)$ for classification

Random variables and (column) vectors are written as uppercase letters X, and Y, while observed values are written with lowercase (x, y). (Dimensions specified if needed.)

Matrices are presented with uppercase boldface: X, often $N \times (p + 1)$.

ESL uses boldface also for x_j being a vector of all N observations of variable j, but in general vectors are not boldface and the vector of observed variables for observation i is just x_i.

Random variables: joint, conditional and marginal distributions

regression $\downarrow$ classification $\downarrow$

Aim: construct $f(X)$ to predict Y or G

Both X and Y are random variables and drawn from some joint distribution

$$p(x_1, x_2, \ldots, x_p, Y)$$

$$p(x,y) = p(y|x) \cdot p(x)$$

popular to use conditional · marginal

Training set

(ESL 2.1)

A set of size N of independent pairs of observations (x_i, y_i) is called the *training set* and often denoted $\mathcal{T}$. Here x_i may be a vector.

The training data is used to estimate the unknown function f.

Validation and test data

Validation data is used for *model selection* (finding the best model among a candidate set).

Test data is used for *model assessment* (assess the performance of the fitted model on future data).

We will consider theoretical results, and also look at different ways to split or resample available data.

More in ESL Chapter 7.

Q: Why did we not use a train/validation/test split in other courses?

Group discussion

Two core regression methods are multiple linear regression (MLR) and k-nearest neighbour (kNN).

For the two methods

- Set up the formal definition for f, and model assumptions made
- What top results do you remember? Write them down.
- What are challenges?

MLR

Classical

$Y_i = f(\mathbb{X}_i) + \varepsilon_i, \quad i = 1, \dots, N$

$\mathbb{X}_i^T \beta$

$1 \dots \quad \beta_0 \dots \beta_p$

$E(\varepsilon_i) = 0$

$Var(\varepsilon_i) = \sigma^2$

$\varepsilon_i, \varepsilon_j$ independent $\quad i \neq j$

$\hat{f}(\mathbb{X}_i) = \hat{Y}_i$

$\underset{N\times 1}{Y} = \underset{N\times(p+1)}{\mathbb{X}} \; \underset{(p+1)\times 1}{\beta} + \underset{N\times 1}{\varepsilon}$

$E(\varepsilon) = 0$

$Var(\varepsilon) = \sigma^2 \underset{N\times N}{I}$

$Cov(\varepsilon)$

Normal MLR: $\varepsilon_i \sim N(0, \sigma^2)$ —— $\varepsilon \sim N_N(0, \sigma^2 I)$

$\hat{\beta} = (\mathbb{X}^T\mathbb{X})^{-1} \mathbb{X} Y$

$E(\hat{\beta}) = \beta, \quad Cov(\hat{\beta}) = \sigma^2 (\mathbb{X}^T\mathbb{X})^{-1}$

$\hat{\sigma}^2 = \frac{1}{N} SSE$ MLE

REML $(N-p-1)$

$\hat{\varepsilon} = Y - \hat{Y}$

$SSE = \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2 = \hat{\varepsilon}^T \hat{\varepsilon}$

k-NN

$$\hat{Y}_0 = \underbrace{\frac{1}{k} \sum_{i \in N_k(x_0)} Y_i}_{\hat{f}(x_0)}$$

N_k often Euclidean distance

$(x_1, y_1), \ldots, (x_p, y_p)$

training data independent pairs

Choose k

Difficult to use for large p

Exercise

RESOURSES

Regression and MLR

See also the exercises!

Resources

(mostly what we learned in TMA4267, or ESL Ch 3, except 3.2.3, 3.2.4, 3.4, 3.7, 3.8)

- From TMA4268: Overview and in particular Module 3: Linear regression
- From TMA4315: Overview and in particular Module 2: MLR

For kNN see also Problem 1 of the TMA4268 2018 exam with solutions

Statistical decision theoretic framework NEW

(ESL Ch 2.4, regression part)

is a mathematical framework for developing models f - and assessing optimality.

First, regression:

- $X \in \Re^p$
- $Y \in \Re$
- $P(X, Y)$ joint distribution of covariates and respons

Aim: find a function $f(X)$ for predicting Y from some inputs X.

Ingredients: Loss function $L(Y, f(X))$ - for *penalizing errors in the prediction*.

Criterion for choosing f: Expected prediction error (EPE)

$$EPE(f) = \mathop{E}_{X,Y}\left(L(Y, f(X)) \right)$$

$$= \int \cdots \int_{x,y} L(y, f(x))\, p(y, x_1, \ldots, x_p)\, dy\, dx_1 \ldots dx_p$$

Choose f to minimize $EPE(f)$!

$$EPE(f) = \int_x \underbrace{\int_y L(y, f(x))\, p(y|x)\, dy}_{\text{solve for each } x}\, p(x)\, dx$$

Squared error loss

$$\mathsf{EPE}(f) = \mathsf{E}_{X,Y}[L(Y, f(X))] = \mathsf{E}_X \mathsf{E}_{Y|X}[(Y - f(X))^2 \mid X]$$

We want to minimize EPE, and see that it is sufficient to minimize $\mathsf{E}_{Y|X}[(Y - f(X))^2 \mid X]$ for each $X = x$ (pointwise):

$$f(x) = \operatorname{argmin}_c \mathsf{E}_{Y|X}[(Y - c)^2 \mid X = x]$$

This gives as result the conditional expectation - the best prediction at any point $X = x$:

$$f(x) = \mathsf{E}[Y \mid X = x]$$

Proof: by differentiating and setting equal 0. ← Exercise
But, do we know this conditional distribution? In practice: need to estimate f.

kNN and conditional expectation

Local conditional mean for observations in T close to $\mathbf{x}_0$:

$$\hat{f}(\mathbf{x}_0) = \frac{1}{k} \sum_{i \in N_k(\mathbf{x}_0)} Y_i$$

in x_0, we form a local mean

$\hat{E}(Y \mid X = x_0)$

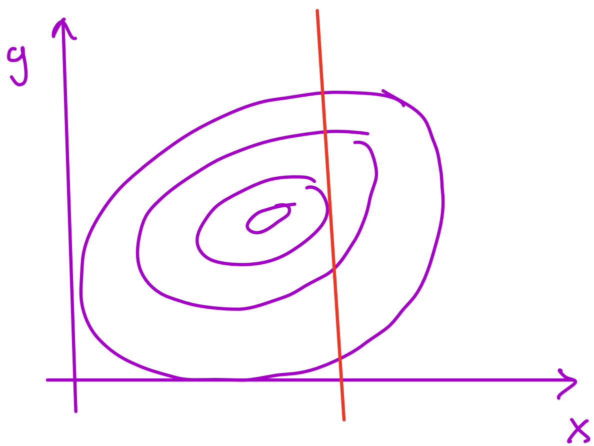
What i~~s~~f the joint distribution is multivariate normal?

Conditionally (known from before): if we assume that $(X, Y) \sim N_{p+1}(\mu, \Sigma)$ then we have seen (TMA4267) that $\mathrm{E}(Y \mid X)$ is linear in X and $\mathrm{Cov}(Y \mid X)$ is independent of X.

Properties of the mvN

↑ Link, check out the $Y|X$ distribution!

$p=1$



We will look more into L1
in part 2

Absolute loss

Regression with absolute (L1) loss: $L(Y, f(X)) = |Y - f(X)|$ gives $\hat{f}(x) = \text{median}(Y \mid X = x)$.
Proof: for example pages 8-11 of
https://getd.libs.uga.edu/pdfs/ma_james_c_201412_ms.pdf

Exercises ← HOMEWORK!

1: Law of total expectation and total variance

This is to get a feeling of the joint and conditional distributions, so that we understand expected value notation with joint, conditional and marginal distributions.
Give a derivation of the law of total expectation:

$$\mathsf{E}[X] = \mathsf{E}[\mathsf{E}(X \mid Y)]$$

and the law of total variance:

$$\mathsf{Var}[X] = \mathsf{EVar}[X \mid Y] + \mathsf{VarE}[X \mid Y]$$

(There is also a law of total covariance.)

2: Quadratic loss and decision theoretic framework

Show that $f(x) = \mathsf{E}[Y \mid X = x]$ for the quadratic loss.

Discussion and conclusions

• core concept: statistical decision theory

$EPE(f) = E_{XY}(L(f(X), Y))$

optimal solutions: $f(x) = E(Y|X)$ if $L(f,Y) = (Y-f)^2$

+ add $(X,Y) \sim mvN$: $f(X) = X^T\beta$

+ kNN is a local version of $\hat{E}(Y|X=x_0)$

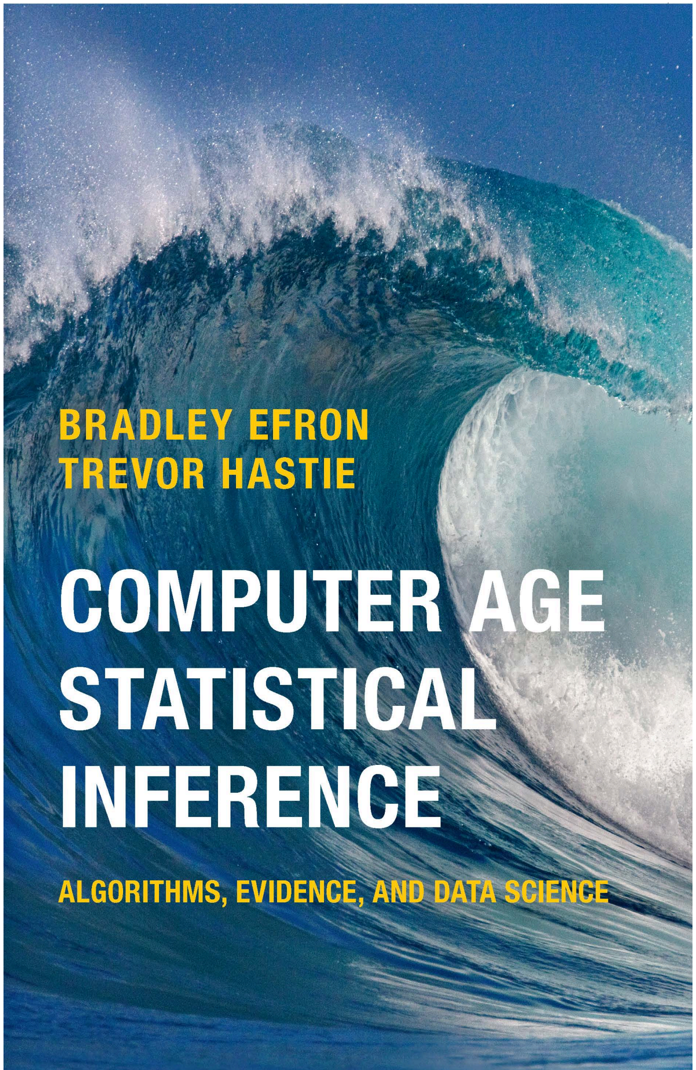
- What are key take home messages from today´s teaching session?
- What do you plan to do before the next teaching session?
- Feedback on today´s teaching session?

welcome!

Exercises from L1.html?

do we want mattelab.math.ntnu.no?

or officehrs?



← good support literature
looks at important
chronological developments
in statistics!