TMA4268 Statistical Learning V2018

Module 12: SUMMING UP

Mette Langaas, Department of Mathematical Sciences, NTNU

week 17, version 22.04.2018

Overview

- course content and learning outcome
- reading list
- overview of modules and core course topics (with exam type questions)
    - incoming questions (no incoming)

through examples of questions

- exam: different types of exam questions and exam preparation
- suggestions for statistics-related courses
- questionnaire

Some of the figures in this presentation are taken from (or are inspired by) "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani.

Course content

In january - just words →
now : real meaning!

Statistical learning, multiple linear regression, classification, resampling methods, model selection/regularization, non-linearity, support vector machines, tree-based methods, unsupervised methods, neural nets.

Learning outcome

1. **Knowledge.** The student has knowledge about the most popular statistical learning models and methods that are used for *prediction* and *inference* in science and technology. Emphasis is on regression- and classification-type statistical models.
2. **Skills.** The student knows, based on an existing data set, how to choose a suitable statistical model, apply sound statistical methods, and perform the analyses using statistical software. The student knows how to present the results from the statistical analyses, and which conclusions can be drawn from the analyses.

And you got to be an expert in using the R language and writing R Markdown reports?

Final reading list

CURRICULUM

Textbook: James, Witten, Hastie, Tibshirani (2013): "An Introduction to Statistical Learning".

- the whole textbook (436 pages)
- the 12 module pages (remark: module 11 not in book, and module 1+12 no "new" material)
- the 3 compulsory exercises
  - Compulsory1
  - Compulsory2
  - Compulsory3

1. Introduction
2. Statistical learning*
3. Linear regression*
4. Classification*
5. Resampling methods
6. Linear model selection and regularization
7. Moving beyond linearity
8. Tree-based methods
9. Support vector machines
10. Unsupervised learning
11. Neural networks*
12. Summing up

Remark: * means that some material is added as compared to the textbook

Core of the course

time, space, images ...

build toolbox: how to analyse data (that are not too complex)

Y, X $\quad$ X

- supervised and unsupervised learning
- supervised: regression and classification

  Y continuous $\quad$ $Y = k$

  - examples of regression and classification type problems
  - how complex a model to get the best fit? flexiblity/overfitting/underfitting.
  - the bias-variance trade-off
  - how to find the perfect fit - validation and cross-validation (or AIC-type solutions) model selection
  - how to compare different solutions - for regression and for classification theory
  - how to evaluate the fit - on new unseen data model assessment
- unsupervised: how to find structure or groupings in data?

and of cause all **the methods** (with underlying models) to perform regression, classification and unsupervised learning. Deep(er) theoretical background and understanding of the models is provided in other statistics courses.

The modules

Here we list *topics* and possible exam related *questions/problems*. In addition, the recommended exercises are useful to work on - and also the exercises in the textbook.

1. Introduction

Topics in Module 1

not so relevent for the exam ...

- Examples, the modules, required background in statistics and
- introduction to R
  - Rbeginner
  - Rintermediate
  - Rintermediate-solutions

2. Statistical learning

and solutions to RecEx

Topics in Module 2

- Model complexity
  - Prediction vs. interpretation.
  - Parametric vs. nonparametric.
  - Inflexible vs. flexible.
  - Overfitting vs. underfitting
- Supervised vs. unsupervised.
- Regression and classification.
- Loss functions: quadratic and 0/1 loss. Fist time we heer this!
- Bias-variance trade-off (polynomial example): mean squared error, training and test set. more in M3
- Classification: the Bayes and KNN-classifier
- * Vectors and matrices, rules for mean and covariances, the multivariate normal distribution. not in text-book
- Model complexity and the bias-variance trade-off is important in “all” subsequent modules.

Questions/Problems:

"Essaytype" questions

- Compulsory1: Problem 1.
- What are differences between a supervised and an unsupervised method? List one method of each type and explain briefly which problem they can solve.
- What are the two main types of supervised methods discussed in this course, and how do they differ? List two methods of each type and explain briefly how the two methods are related.

understanding & overview

3. Linear regression

and solutions to RecEx

Topics in Module 3

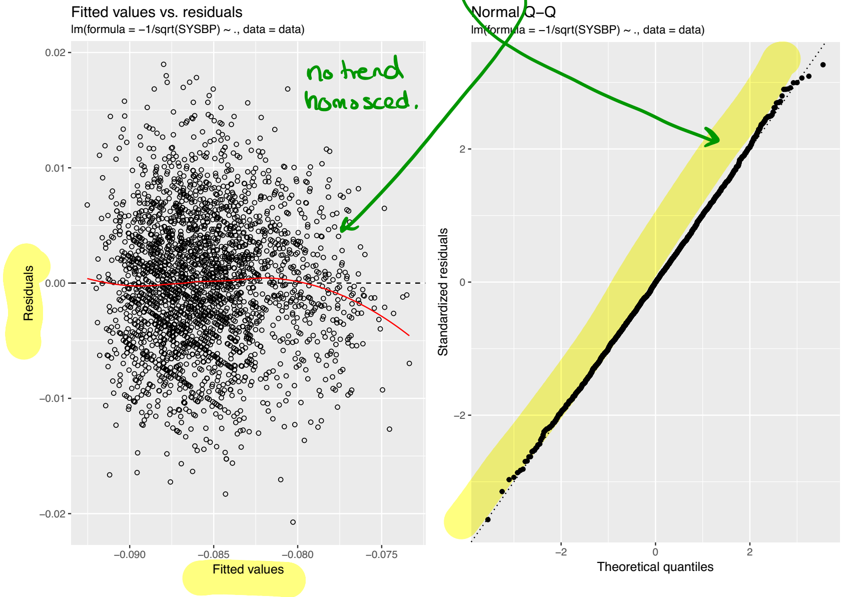
- Examples: Munich rent index, ozone, SLID, Framingham heart disease, Boston housing prices, auto.
- The classical normal linear regression model on vector/matrix form.
- Parameter estimators and distribution thereof. Model fit.
- Confidence intervals, hypothesis tests, and interpreting R-output from regression.
- Qualitative covariates, interactions.
- This module is a stepping stone for all subsequent uses of regression in Modules 6, 7, 8, and 11.

Questions/Problems:

- Compulsory1: Problem 2.
- Theoretical questions are referred to TMA4267 Linear statistical models, but basic knowledge and interpretation of `lm` output important.
- Write down the classical normal multiple regression model in vector and matrix notation. Specify dimensions and explain your notation. Also write down the estimator for the regression coefficients. What is the distribution of this estimator?
- Output from `summary.lm` presented - maybe with question marks in place of number - and you explain and calculate. Interpret two top residual plots!
- What is the "bias-variance decomposition"? Is it applicable to all choices of loss functions? Write down the derivation for quadratic loss for a regression problem at $\mathbf{x}_0$. Explain your notation.

```
##
## Call:
## lm(formula = -1/sqrt(SYSBP) ~ ., data = data)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0207366 -0.0039157 -0.0000304  0.0038293  0.0189747
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.103e-01  1.383e-03 -79.745  < 2e-16 ***
## SEX         -2.989e-04  2.390e-04  -1.251 0.211176
## AGE          2.378e-04  1.434e-05  16.586  < 2e-16 ***
## CURSMOKE    -2.504e-04  2.527e-04  -0.991 0.321723
## BMI          3.087e-04  2.955e-05  10.447  < 2e-16 ***
## TOTCHOL      9.288e-06  2.602e-06   3.569 0.000365 ***
## BPMEDS       5.469e-03  3.265e-04  16.748  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.005819 on 2593 degrees of freedom
## Multiple R-squared:  0.2494, Adjusted R-squared:  0.2476
## F-statistic: 143.6 on 6 and 2593 DF,  p-value: < 2.2e-16
```

$\varepsilon \sim \boxed{N}(0, \sigma^2 I)$



Fitted values vs. residuals

lm(formula = −1/sqrt(SYSBP) ~ ., data = data)

Normal Q−Q

lm(formula = −1/sqrt(SYSBP) ~ ., data = data)

no trend
homosced.

The bias-variance trade-off in the regression setting

Assume that we have fitted a *regression* curve $Y = f(x) + \varepsilon$ to our training data, which consist of independent observation pairs $\{x_i, y_i\}$ for $i = 1, .., n$.

We assume that ε is an unobserved random variable that adds noise to the relationship between the response variable and the covariates and is called the random error, and that the random errors have mean zero and constant variance σ^2 for all values of x.

This noise is used as a substitute for all the unobserved variables that is not in our equation, but that influences Y.

The fitted curve is denoted by $\hat{f}$.

We want to use $\hat{f}$ to make a prediction for a new observation at x_0, and are interested in the error associated with this prediction. The predicted response value is then $\hat{f}(x_0)$.

Prerequisites →

The *expected test mean squared error (MSE) at* x_0 is defined as:

$$\mathrm{E}[Y - \hat{f}(x_0)]^2$$

This expected test MSE can be decomposed into three terms

$$
\begin{aligned}
\mathrm{E}[Y - \hat{f}(x_0)]^2 &= \mathrm{E}[Y^2 + \hat{f}(x_0)^2 - 2Y\hat{f}(x_0)] \\
&= \mathrm{E}[Y^2] + \mathrm{E}[\hat{f}(x_0)^2] - \mathrm{E}[2Y\hat{f}(x_0)] \quad \rightarrow \text{ Y and } \hat{f}(x_0) \text{ independent} \\
&= \mathrm{Var}[Y] + \mathrm{E}[Y]^2 + \mathrm{Var}[\hat{f}(x_0)] + \mathrm{E}[\hat{f}(x_0)]^2 - 2\mathrm{E}[Y]\mathrm{E}[\hat{f}(x_0)] \\
&= \mathrm{Var}[Y] + f(x_0)^2 + \mathrm{Var}[\hat{f}(x_0)] + \mathrm{E}[\hat{f}(x_0)]^2 - 2f(x_0)\mathrm{E}[\hat{f}(x_0)] \\
&= \mathrm{Var}[Y] + \mathrm{Var}[\hat{f}(x_0)] + (f(x_0) - \mathrm{E}[\hat{f}(x_0)])^2 \\
&= \mathrm{Var}(\varepsilon) + \mathrm{Var}[\hat{f}(x_0)] + [\mathrm{Bias}(\hat{f}(x_0))]^2.
\end{aligned}
$$

$$\mathrm{E}[(Y-\hat{f}(x_0))^2] = \cdots = \mathrm{Var}(\varepsilon) + \mathrm{Var}[\hat{f}(x_0)] + [\mathrm{Bias}(\hat{f}(x_0))]^2$$

- First term: irreducible error, $\sigma_{\cancel{\varepsilon}}^2$ and is always present unless we have measurements without error. This term cannot be reduced regardless how well our statistical model fits the data.
- Second term: variance of the prediction at x_0 or the expected deviation around the mean at x_0. If the variance is high, there is large uncertainty associated with the prediction.
- Third term: squared bias. The bias gives an estimate of how much the prediction differs from the true mean. If the bias is low the model gives a prediction which is close to the true value.

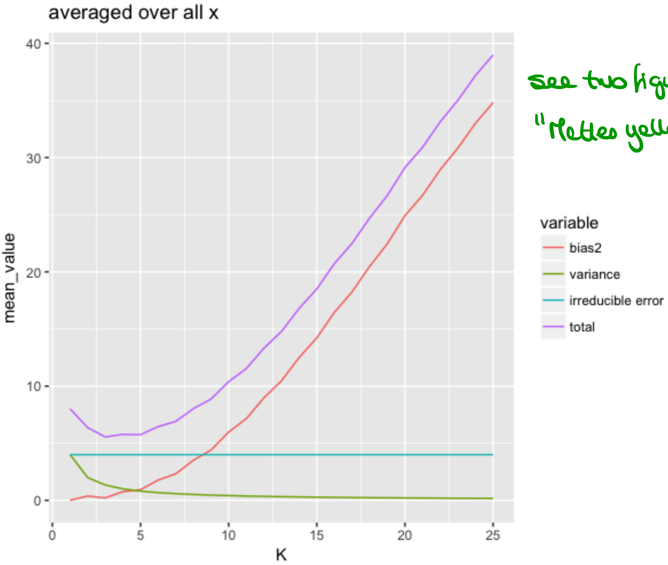


Figure 1: Compulsory 2, Problem 1

see two figures on "Mettes yellow sheet"

4. Classification

and solutions to RecEx (~~Mainly~~ mostly discussed the two-class problem in this course)

Topics in Module 4

- Examples: South African heart disease, wine, German credit data, IMDB movie review, MNIST digit classification, iris plants.
- Bayes classifier: classify to the most probable class gives the minimize the expected 0/1 loss. We usually do not know the probability of each class for each input. The Bayes optimal boundary is the boundary for the Bayes classifier and the error rate (on a test set) for the Bayes classifier is the Bayes error rate. Related to the irreducible error (but bias-variance decomposition is for quadratic loss).

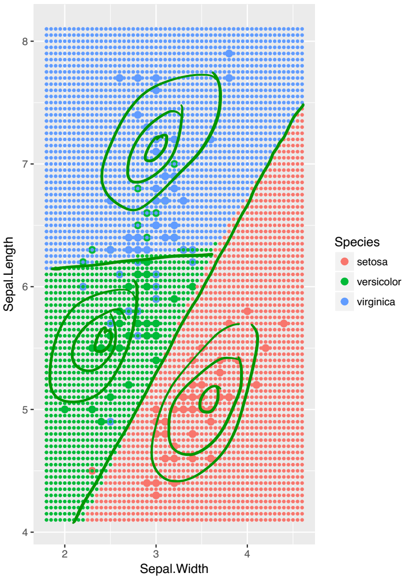
CompEx3, 3a

useful when we simulate data and can calculate the Bayes error rate

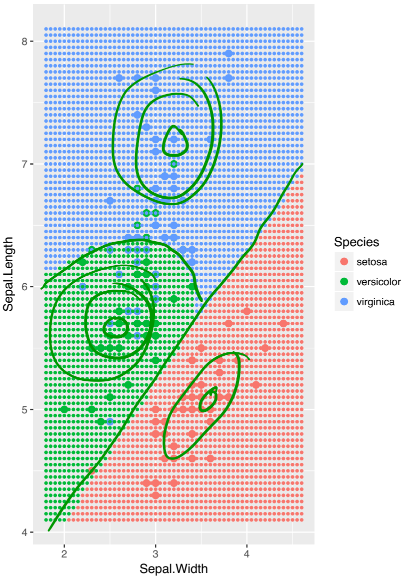
- Two paradigms (not in textbook):
  - $P(Y=k \mid X=x)$

    diagnostic (directly estimating the posterior distribution for the classes)
  - $\pi_k = P(Y=k)$ $f_k(x)$

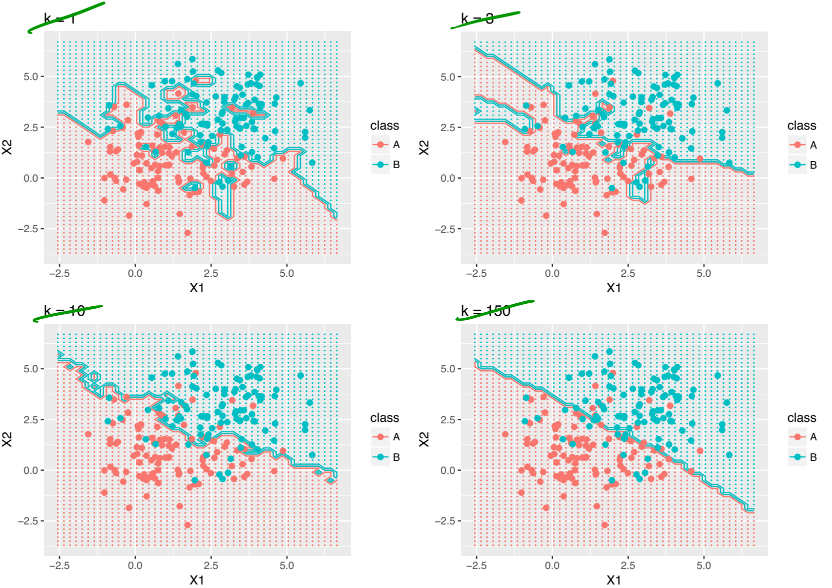
    sampling (estimating class prior probabilities and class conditional distribution and then putting together with Bayes rule)
- LDA and QDA: sampling paradigm. Multivariate normal class densities with common covariance (LDA) or class specific covariance (QDA). Class boundaries will be linear (LDA) or quadratic (QDA). Handle easily more than two classes.
- KNN: diagnostic paradigm. Formula for posterior class probability. Overfitting/underfitting and flexibility of class boundary as a function of K. Non-linear class boundaries. Handle easily more than two classes.


Sepal.Length
8
7
6
5
4
2
3
4
Sepal.Width
Species
setosa
versicolor
virginica

LDA


Sepal.Length
8
7
6
5
4
2
3
4
Sepal.Width
Species
setosa
versicolor
virginica

QDA



Are you able to deduce which is $k = 1, 3, 10, 150$?
why?

$P(Y=k \mid X=x)$

- Logistic regression: not regression but classification method. Diagnostic paradigm. Logistic (sigmoid) function and linear predictor. Interpretation of regression coefficients using odds. Linear class boundaries. Two classes.
- Evaluation with confusion matrix, ROC-curve and AUC.
- This is one of the few modules where we have done some theoretical work (both LDA and logistic regression), and this is also a stepping stone module for subsequent use of classification in Modules 8, 9 and 11.

Questions/Problems:

Types of exam questions:

essay theory hands-on

- Compulsory1: Problem 3.
- Compulsory3: Problem 2a.
- Which are the two paradigms we have presented for classification? Explain briefly how these two differ and identify which of the classification methods that we have discussed in this course belongs to which paradigm. Describe one method from each paradigm briefly.
- Assume we have two classes (class 1 and class 2) and a bivariate input variable (covariate) $\mathbf{x}$. We now assume that each class conditional distribution is bivariate normal with the same covariance matrix for the two classes. Write down the posterior probability for class 1 (explaining all the parameters that are involved). Then show (yes, derive the result) that the class boundary between class 1 and 2 is linear in the two components of $\mathbf{x}$. What is the name of this classification method? LDA
- Given parameter estimates for class means and common covariance matrix (numerical values), use LDA to predict the class of a new observation. (Then pdf in multivariate normal must be given.)

- Logistic regression is a classification method for two classes, where the classes are coded 0 and 1. Assume we have fitted a logistic regression to a data set with covariates x_1 and x_2 and that the fitted model is written

$$\hat{p} = \frac{\exp(1 + 2 \cdot x_1 + 3 \cdot x_2)}{1 + \exp(1 + 2 \cdot x_1 + 3 \cdot x_2)}$$

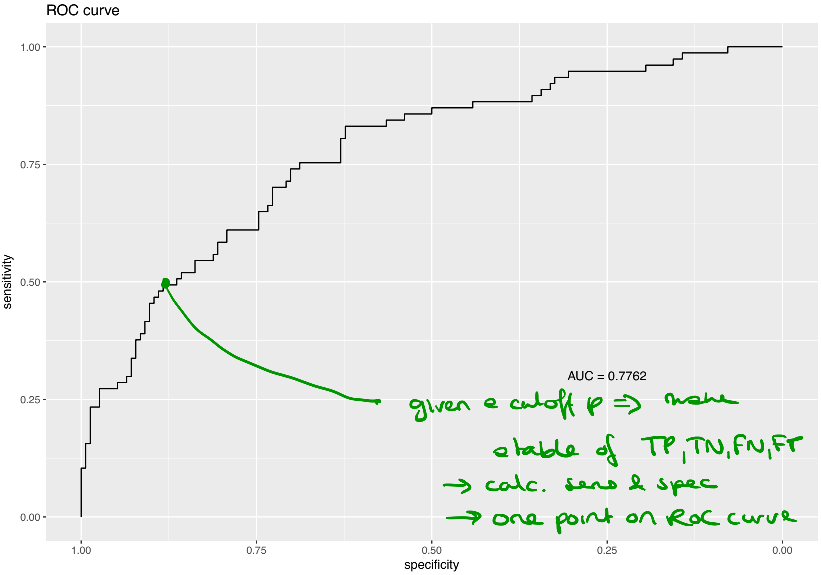
[x_1 incr by 1
odds $\frac{p}{1-p}$ mult by $e^{\hat{\beta}_1}$]

  What is the interpretation of p (left side) here? What is the interpretation of the regression coefficient $\hat{\beta}_1 = 2$?
- What is a confusion matrix? What is it used for? How is the misclassification rate defined?
- We have actively used *receiver-operator-curve* (ROC) and the *area under the curve* (AUC) in this course. In which types of problems are these used? Explain how a ROC-curve is constructed. If a method gives a AUC of 0.5 when used on a data set, what can you say about this method?
- Output from fitting a method is presented - you explain and evaluate output, evaluate classification boundaries, interpret ROC-curve and compare methods.

```
##
## Call:
## glm(formula = chd ~ ., family = "binomial", data = heartds)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7781  -0.8213  -0.4387   0.8889   2.5435
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -6.1507209  1.3082600  -4.701 2.58e-06 ***
## sbp             0.0065040  0.0057304   1.135 0.256374
## tobacco         0.0793764  0.0266028   2.984 0.002847 **
## ldl             0.1739239  0.0596617   2.915 0.003555 **
## adiposity       0.0185866  0.0292894   0.635 0.525700
## famhistPresent  0.9253704  0.2278940   4.061 4.90e-05 ***
## typea           0.0395950  0.0123202   3.214 0.001310 **
## obesity        -0.0629099  0.0442477  -1.422 0.155095
## alcohol         0.0001217  0.0044832   0.027 0.978350
## age             0.0452253  0.0121298   3.728 0.000193 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 596.11  on 461  degrees of freedom
## Residual deviance: 472.14  on 452  degrees of freedom
## AIC: 492.14
##
## Number of Fisher Scoring iterations: 5
```

asymptotic results

more in TMA4315 GLM



given a cutoff p => make a table of TP, TN, FN, FP

→ calc. sens & spec

→ one point on ROC curve

The bias-variance trade-off in the classification setting?

- Bias-variance trade-off is for quadratic loss.
- Generalizations exists - but not covered in this course.
- For classification we tend to think of the Bayes error rate as some kind of lowest possible error rate - similar to the irreducible error.
- In classification we are also focussed on over/under-fitting, and refer to a method that fits the classification boundary closely as having small bias.

5. Resampling methods ← mainly for model selection

and (handwritten) solutions to RecEx

Topics in Module 5

fit & select

model assessment

- Data rich situation: Training-validation and test set.
- Validation set approach
- How is cross-validation performed? For regression and for classification.
- LOOCV, 5 and 10 fold CV
- good and bad issues with validation set, LOOCV, 10-fold CV
- bias and variance for k-fold cross-validation - end up with k=5 or k=10 fold as good balance?
- selection bias - the right and wrong way to do cross-validation
- bootstrapping to estimate uncertainty in statistic (warming up to Module 8)

Questions/Problems:

- Compulsory 2: Problem 1cd
- In a setting where you have access to unlimited amounts of data explain the role of the training set, validation set, and test set. Point to advantages/disadvantages of making such a division of the data set. Your answer should include the words: model complexity, tuning parameters, overfitting, model fit/parameters.
- In a setting where you have access to limited amount of data explain how *k*-cross-validation can be used for model assessment and model selection. A drawing might be useful.
- (From MA8701 exam): Explain what is meant by cross-validation. Discuss its use in practice. How does cross-validation relate to the use of training/validation/test sets?

- Explain how a bootstrap sample is drawn. What is the probability that an observation in our data set will be a part of a given bootstrap sample?
- Assume that we want to fit a regression model. Explain how se can use bootstrapping to estimate the standard deviation of parameters estimates in our model.

6. Linear model selection and regularization:

- Lecture 1 and
- Lecture 2
- and solutions to RecEx

Topics in Module 6:

- Model selection: estimate performance of different models to choose the best one.
- Model assessment: having chosen a final model, estimate its performance on new data
- Model selection by penalizing the training error: AIC, BIC, C_p, Adjusted R^2.
- Cross-validation can be used for model selection and assessment.
- Subset selection:
  - best subset selection
  - stepwise model selection

$\hat{Y} = x^T \hat{\beta}$

$(y - \hat{y})^2$

loss + λ penalty β_j^2 $|\beta_j|$

- Shrinkage methods
  - ridge regression: quadratic L2 penalty added to RSS
  - lasso regression: absolute L1 penalty added to RSS
  - no penalty on intercept, not scale invariant: center and scale covariates
- Dimension reduction methods:

  p large compared to n
  multicoll. ← $(X^TX)^{-1}\sigma^2$

  - principal component analysis: eigenvectors, proportion of variance explained, scree plot
  - principal component regression
  - parti~~c~~al least squares (lightly covered) PLS
    not on exam
- High dimensionality issues: multicollinearity, interpretation.

Questions/Problems:

- Compulsory 2: Problem 1
- Compulsory 2: Problem 2
- Print-out from best subset selection, explain how this is done and what the best model is if you use BIC. Explain how you instead (of using BIC) can use cross-validation.
- We have discussed parametric methods where the parameters are found by minimizing the sum of a loss function and a penalty. Choose one such method, write down the loss and penalty used, and explain how this is related to the bias-variance trade-off.
- Interpret figures, explain what you see. What do we call this method?
- Best subset and lasso: Exam TMA4267V2016 Problem 2d with solutions
- Best subset and lasso: Exam TMA4267V2014 Problem 2c with solutions

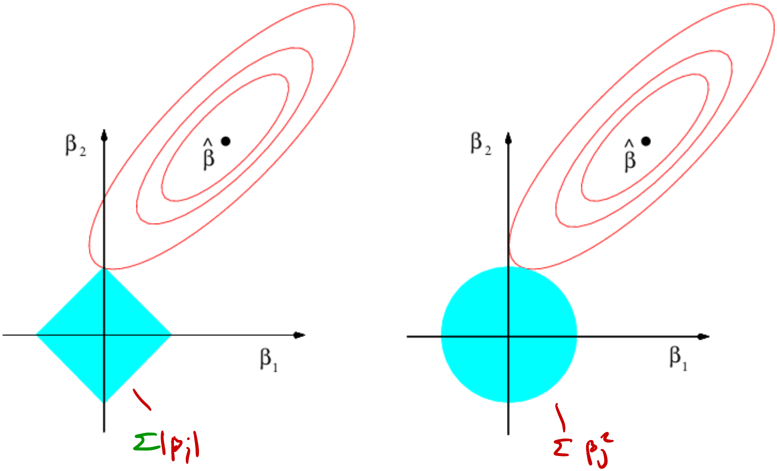


Figure 2: ISL 6.7

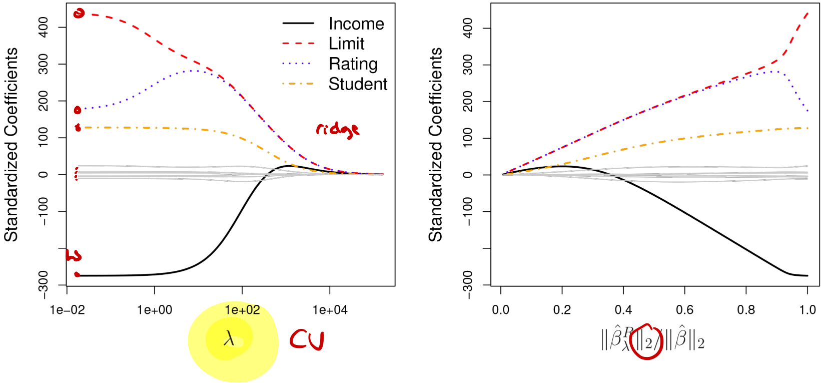
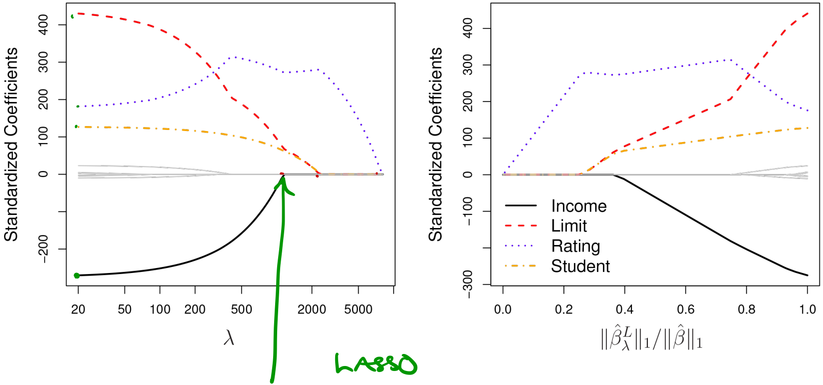


Figure 3: ISL 6.4

loss + λ. penalty



LASSO

goes to zero

Figure 4: ISL 6.6

- Explain how you find the principal components for a given data set, and how these are used in regression. Assume you have p covariates and n observations (where $n >> p$) and you fit a regression model with the first p principal components as regressors. How does this compare to fitting a multiple linear regression to the original covariates? What if you instead only use the first q principal components, where $q < p$.
- MLR, overfitting and principal component regression: Exam TMA4267K2014 Problem 2 with solutions

7. Moving beyond linearity

and solutions to RecEx

Topics in Module 7

- Modifications to the multiple linear regression model - when a linear model is not the best choice. Similar techniques can be used for classification, but we only looked at regression. First look at one covariate, combine in "additive model".
- Basis functions: fixed functions of the covariates (no parameters to estimate)
- Polynomial regression: multiple linear regression with polynomials as basis functions.
- Step functions - piece-wise constants. Like our dummy variable coding of factors.
- Regression splines: regional polynomials joined smoothly - neat use of basis functions. Cubic splines very popular.

design matrix
X

- Smoothing splines: smooth functions - minimizing the RSS with an additional penalty on the second derivative of the curve. Results in a natural cubic spline with knots in the unique values of the covariate. Complexity parameters chosen by AIC (with degrees of freedom) or cross-validation. (UiO mainly AIC, we mainly cross-validation.)
- Local regressions: smoothed K-nearest neighbour with local regression and weighting. In applied areas `loess` is very popular.
- Additive models: combine the above. Sum of (possibly) non-linear instead of linear functions.

Questions/Problems:

- Compulsory 2: Problem 3.
- What is the difference between a cubic spline and a natural cubic spline? What would you prefer?
- A smoothing spline is a function minimizing the RSS and an additional penalty. What type of penalty is this? There is a tuning parameter involved - hos can that be chosen? (Details on the smoother matrix and relationship to ridge is beyond the scope here.)
- UiO 2017 Problem 1c with solutions- but we did not focus on degrees of freedom.
- Remark: the methods of this module will be elaborated on in the MA8701 Statistical learning phd course V2019. Then connections to the radial basis functions in Module 9 would become more clear.

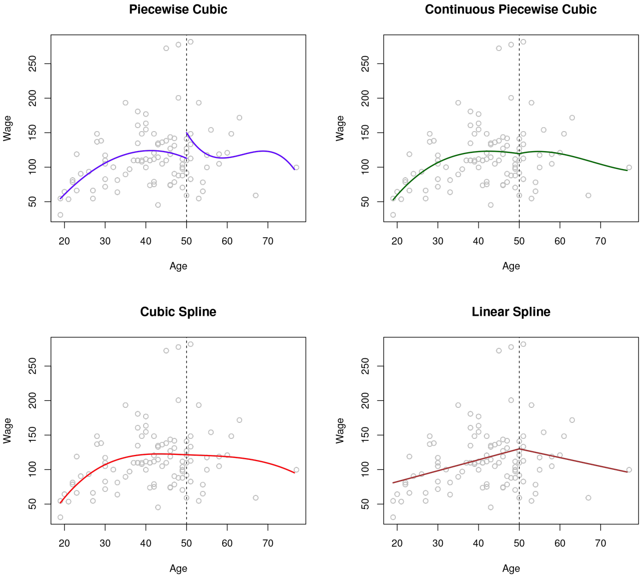


Figure 5: ISL 7.3

8. Tree-based methods

and solutions to RecEx

Topics in Module 8

- Method applicable both to regression and classification (K classes) and will give non-linear covariate effects and include interactions between covariates. Based on binary splits of each covariate at a time.
- Glossary: root, branches, internal nodes, terminal (leaf) nodes. Tree drawn upside down.
- A tree can also be seen as a division of the covariate space into non-overlapping regions.
- We build a tree from binary splits in one covariate at the time, chosen to improve some measure of error or impurity. The tree is created by not looking ahead - only at the current best split - thus a *greedy strategy*.
- Criterion to minimize
  - Regression: residual sums of squares
  - Classification: Gini or cross entropy impurity measure or deviance

- When to stop: decided stopping criterion - like minimal decrease in RSS or less than 5 observations in terminal node.
- Prediction in terminal nodes:
  - Regression: $\hat{y} = \frac{1}{N_j} \sum_{i:x_i \in R_j} y_i$
  - Classification: majority vote or fraction of each class in a node - and cut-off on probabiity.
- Grow full tree, and then prune back using pruning strategy: cost complexity pruning= cost function + penalty times number of terminal notes (hot handled in detail).

- From one tree to many trees= forest. Why? To improve prediction (but this will give worse interpretation).
- Bagging (bootstrap aggregation): draw B bootstrap samples and fit one full tree to each, used the average over all trees for prediction.
- Random forest: as bagging but only m (randomly) chosen covariates (out of the p) are available for selection at each possible split. Rule of thumb for m is $\sqrt{p}$ for classificaton and $\overset{p}{m}/3$ for regression.
- OOB: out-of-bag estimation can be used for model selection - no need for cross-validation.
- Variable importance plots: give the total amount of decrease in RSS or Gini index over splits of a predictor - averaged over all B trees. May also be calculated over randomization of OOB.
- Boosting: fit one tree with d splits, make residuals and fit a new tree, adjust residuals partly with new tree - repeat. Three tuning paramteers chosen by cross-validation.

MA8701 ← more there

Questions/Problems:

- Compulsory 3: Problem 1.
- What does it mean that a method is *greedy*? Mention one greedy method that we have studied and explain why it is greedy.
- How do we choose that we perform a split in a tree? What is the natural cost function for regression? For classification we focus on node impurity - explain one possible cost function for node impurity.
- Image of tree, explain what you see. Predict the value for a new observation with numerical value given.
- Show full tree and pruned tree and results on test set: compare and argument for which of the models to choose.

- How do we choose the number of bootstrap samples B to be used in bagging and random forest? What about boosting?
- Why do we not have to use cross-validation to estimate error rates for bagging and random forest? What do we instead use, and how do we estimate error rates?
- (MA871 exam): What is boostrapping? We have looked at boostrapping for finding the standard error of an estimator and for bagging and random forest. What is the main idea behind bagging? What is the connection between bagging and random forests?
- For regression trees - how is a simple way to perform boosting?

9. Support vector machines

and solutions to RecEx - stolen from other source.

- SVM is a method for both classification and regression, but we have only studied two-class classification (classes are coded -1 and 1).
- Aim: find high dimensional hyperplan that separates two classes $f(\mathbf{x}) = \beta_0 + \mathbf{x}^T\beta = 0$. If $y_i f(\mathbf{x}_i) > 0$ observation $\mathbf{x}_i$ is correctly classified.
- Central: maximizing the distance (on both sides) from the class boundary to the closes observations= the margin M (maximal marginal classifier) - which is relaxed with slack variables (support vector classifiers), and to allow nonlinear functions of $\mathbf{x}$ by extending an inner product to kernels (support vector machine).
- Support vectors: observations that ~~like~~ lie on the margin or on the wrong side of the margin.

- Kernels: generalization of an inner product to allow for non-linear boundaries and to speed up calculations due to inner products only involve support vectors. Most popular kernel is radial $K(x_i, x_i') = \exp(-\gamma \cancel{)} (\sum_{j=1}^{p}(x_{ij} - x_{i'j})^2.)$
- Tuning parameters: cost and parameters in kernels - chosen by CV.
- Sad: not able to present details since then a course in optimization is needed - more in MA8701.
- Nice connection to non-linar and ridged version of logistic regression - comparing hinge loss to logistic loss - but then without the computational advanges of the kernel method.

Questions/Problems:

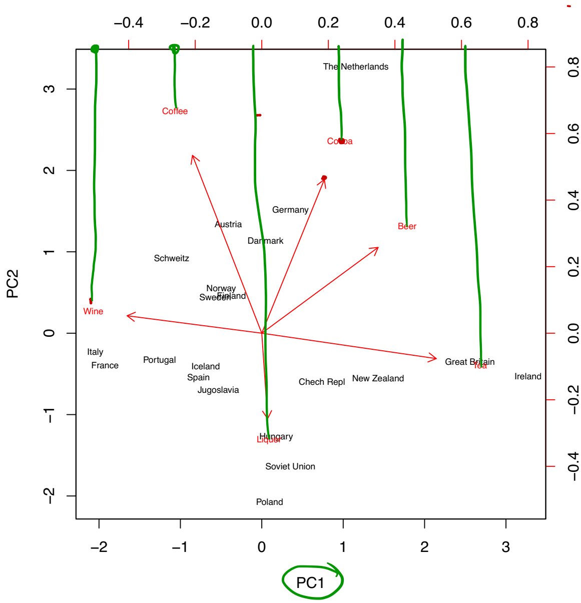
- Compulsory 3: Problem 2b.
- What is a support vector?
- What are differences between a maximal margin classifier and linear discriminant analysis classifier?
- What are the main differences between the maximal margin classifier and the support vector classifier? Explain the concept of a slack variable.
- What are important aspects of the support vector machine?

10. Unsupervised learning: 6 files

- Lecture 1 with Lab1 and New York times stories.
- Lecture 2 with Lab2 and Lab3
- and solutions to RecEx

Topics in Module 10

- Principal component analysis:
  - mathematical details (eigenvectors corresponding to covariance or correlation matrix) also in TMA4267.
  - understanding loadings and scores and a biplot, choosing the number of principal components from proportion of variance explained or scree-type plots (elbow)
- Clustering:
  - *k*-means: number of clusters given, iterative algorithm to classify to nearest centroid and recalculate centroid
  - hierarchical clustering: choice of distance measure, choice of linkage method (single, average, complete),

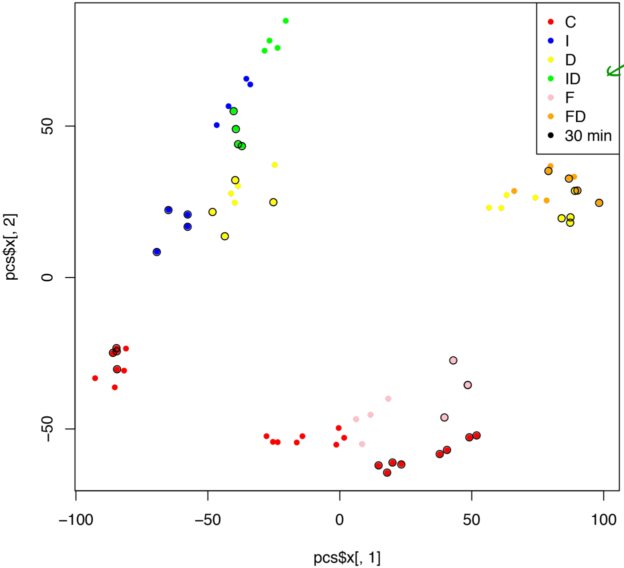


$PC1 = -0.5\,\text{wine} - 0.25\,\text{coffee}$

$-0.1 \cdot \text{Liq} + 0.21 \cdot \text{Cocoa}$

$+ 0.4\,\text{Beer} +$

$0.6\,\text{tea}$



groups

$p = 30\,000$ gene expressions

Figure 6: PCA for quality control

to see if groups cluster = all ok.

else - something wrong?

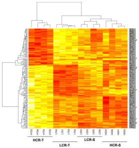


Figure 7: Hierarchical clustering for visualization

Questions/Problems:

do this by hand - please

- Compulsory 3: Problem 3
- Principal component analysis is both used as an unsupervised method and in a supervised regression setting. Explain briefly how we define the principal components (loadings and scores) and how the principal components are used in the two settings.
- Could also have small numerical task to show that you have understood how to construct a dendrogram (see under Exam below).

11. Neural networks: 13 files

- 1-examples_introduction
- 2-data_representation_for_nn
- 3-keras
- 4-deep_learning_models
- 5-compiling_deep_learning_models
- 6-model_inference
- 7-prevent_overfitting
- 8-neural_networks_mnist
- 9-neural_networks_imbd
- 10-neural_networks_neuters
- 11-neural_networks_boston_housing
- 12-neural_networks_convolution_mnist
- 13-neural_networks_recurrent_imbd
- no solutions to RecEx - but many examples to study!

Recommended *further* reading is (soon available at the NTNU library): Deep learning with R

Topics in Module 11:

- Feedforward network architecture: mathematical formula - layers of multivariate transformed (`relu`, `linear`, `sigmoid`) multiple linear regression - sequentially connected.
- What is the number of parameters *p* that need to be estimated? Intercept term (for each layer) is possible and is referred to as "bias term".
- Loss function to minimize (on output layer): regression (quadratic), classification binary (binary crossentropy), classification multiple classes (categorical crossentropy).
- not covered How to minimize the loss function: gradient based (chain rule) back-propagation - many choices: review
- How to avoid overfitting: reduce network size, collect more observations, regularization, drop-out.
- Technicalities: `keras` in R. Use of tensors. Piping sequential layers, piping to estimation and then to evaluation (metrics).
- Recurrent network: with feedback loop (not central)
- Convolution networks: with filters - especially for images (not central)

Questions/Problems:

see also "Mettes yellow sheet"

- Compulsory 3: Problem 4
- Given a data set (inputs and outputs, problem explained) - what are possible feedforward network architectures that you would explore?
- What are the similarities and differences beween a feedforward neural network with one hidden layer with `linear` activation and `sigmoid` output (one output) and logistic regression?
- In a feedforward neural network you may have 10000 weights to estimate but only 1000 observations. How is this possible?
- Which network architecture and activation functions does this formula give?

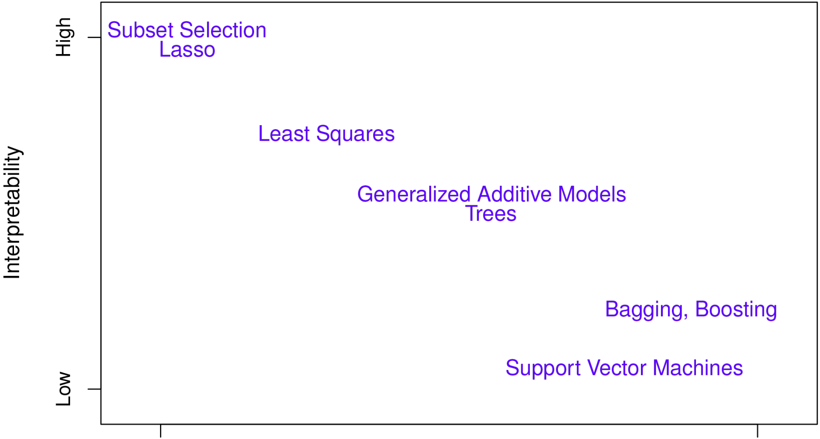
$$\hat{y} = \sum_{j=1}^{5} w_{2j} max(\sum_{i=1}^{10} w_{1ij} x_j, 0)$$

- What is the most interesting aspect of neural network (your opinion)? How would you compare how feedforward neural networks are fitted as compared to fitting multiple linear regression and logistic regression?
- In a regression setting: Consider
  - a sum of non-linear functions of each covariate in Module 7
  - a sum of many non-linear functions of sums of covariates in feedforward neural networks (one hidden layer, non-linear activation in hidden layer) in Module 11.
  - Explain how these two ways of thinking differ?

12. Summing-up (this module)

Questions/Problems - overall level

- This is Figure 2.7 from James et al. (2013). Explain what is the message of this figure.



model complexity

- Make a graph with “horisontal axis: Model complexity” and “vertical axis: Interpretability” and position the classification methods we have covered in this course in the graph.
- For many of the methods we have studied, the models are fitted minimizing a sum of a loss function and a tuning parameter times a penalty. Choose one method from regression and one from classification and explain what is the loss function and what is the penalty. Explain what the goal of the penalty is, and how the tuning parameter can be chosen.

Exam and exam preparation

We take a look at the information posted at Blackboard, left margin "Exam".

- About the exam: when, what to bring.
- Previous exams (none, except UiO) - but possible questions listed (above) for each module.
- Supervision before the exam - see dates -
- and maybe use the Discussion forum on Bb?

Then, a few (more) word on exam set-up and question types

The planned exam set-up

We have 30% on the compulsory exercies, and 70% on the written exam. These 70% is 70 points on the written exam.

- Problem 1: Regression - 20-50 points
- Problem 2: Classificaiton - 20-50 points
- Problem 3: Unsupervised learning - 0-20 points
- inherently: overfitting and bias-variance trade-off, train/validate/test and cross-validation, assumptions and reasoning behind models and methods

Regression problem

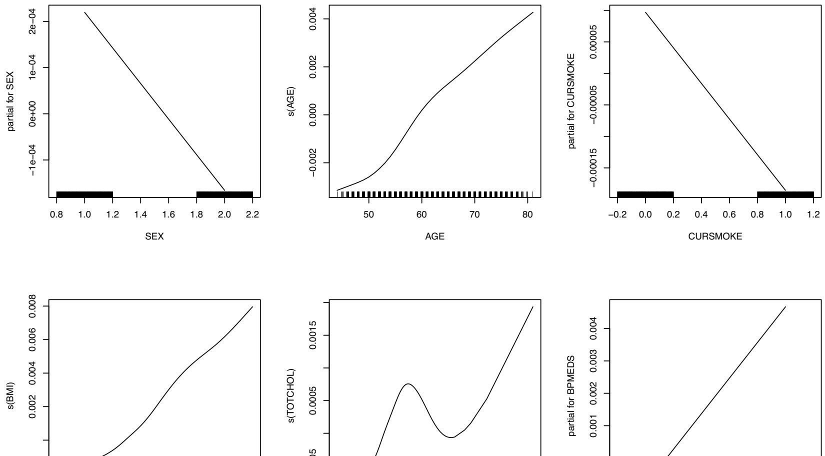
- Explain about a data set and show print-out and residual plots from fitting a multiple linear regression model:
    - interpret, write down formulas, assess model fit.

For example with the Framingham data set from Compulsory exercise 1: Problem 2.

```
##
## Call:
## lm(formula = -1/sqrt(SYSBP) ~ ., data = data)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0207366 -0.0039157 -0.0000304  0.0038293  0.0189747
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.103e-01  1.383e-03 -79.745  < 2e-16 ***
## SEX         -2.989e-04  2.390e-04  -1.251 0.211176
## AGE          2.378e-04  1.434e-05  16.586  < 2e-16 ***
## CURSMOKE    -2.504e-04  2.527e-04  -0.991 0.321723
## BMI          3.087e-04  2.955e-05  10.447  < 2e-16 ***
## TOTCHOL      9.288e-06  2.602e-06   3.569 0.000365 ***
## BPMEDS       5.469e-03  3.265e-04  16.748  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.005819 on 2593 degrees of freedom
```

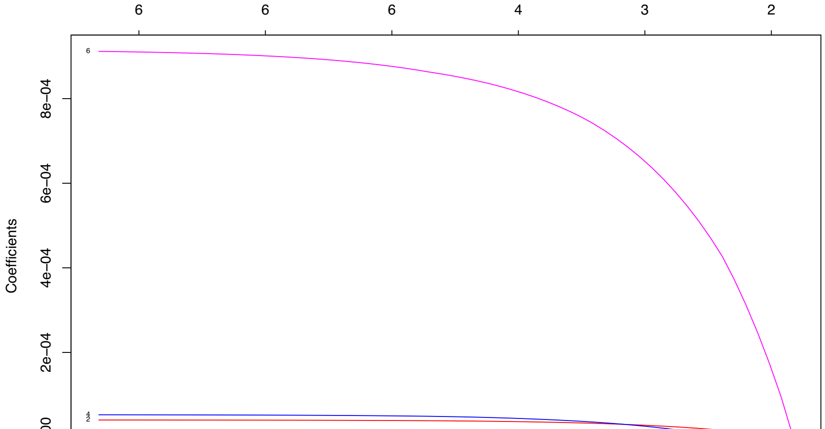
- Expand into investigating non-linearities of one or more covariates, present print-out and plots
    - explain, interpret, write down formulas, maybe compare linear vs. non-linear fit

```
library(gam)
m2=gam(-1/sqrt(SYSBP) ~ SEX+s(AGE)+CURSMOKE+s(BMI)+s(TOTCHOL)+BPMEDS,data = data)
par(mfrow=c(2,3))
plot(m2)
```



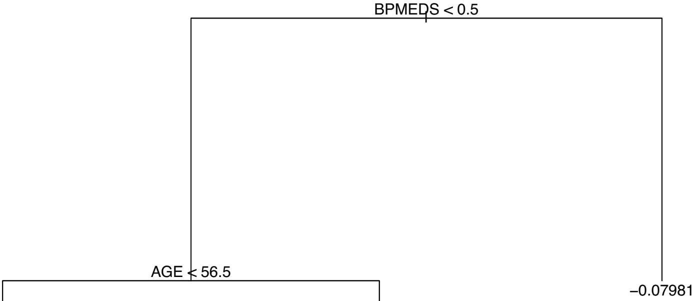
- Fitting a penalized solution might improve on prediction accuracy. Explain what is done below and what is the suggested regression model. Write down the function that is minimized to fit this model, and explain how the tuning parameter is chosen.

```
library(glmnet)
x <- model.matrix(modelA)[,-1]
y<- -1/data$SYSBP
fit.lasso=glmnet(x,y)
plot(fit.lasso,xvar="lambda",label=TRUE)
```



- Then move on to a tree, and look at full (and possibly pruned) tree. Prediction and interpretation. Theoretical questions on the fitting. Predict value for a new observation (numerically): given that a patient has `BPMED=1`, is 30 years of age and has a BMI of 27, what is the predicted value for the `-1/sqrt(SYSBP)`?

```r
library(tree)
m3=tree(-1/sqrt(SYSBP) ~ .,data=data)
plot(m3)
text(m3)
```



Then a test set should mysteriously appear to be part of a testing regime (but skip that for this example). Other questions might be:

- Improving on trees by bagging or random forest: why would we do that?
- Can this problem be solved with the use of feedforward neural networks? Suggest a possible architecture (the number of nodes in input and output layer must match the problem, and activiation functions chosen accoringly). What would be your choice of loss function?
- Comparing different regression solutions with respect to interpretability and prediction accuracy.

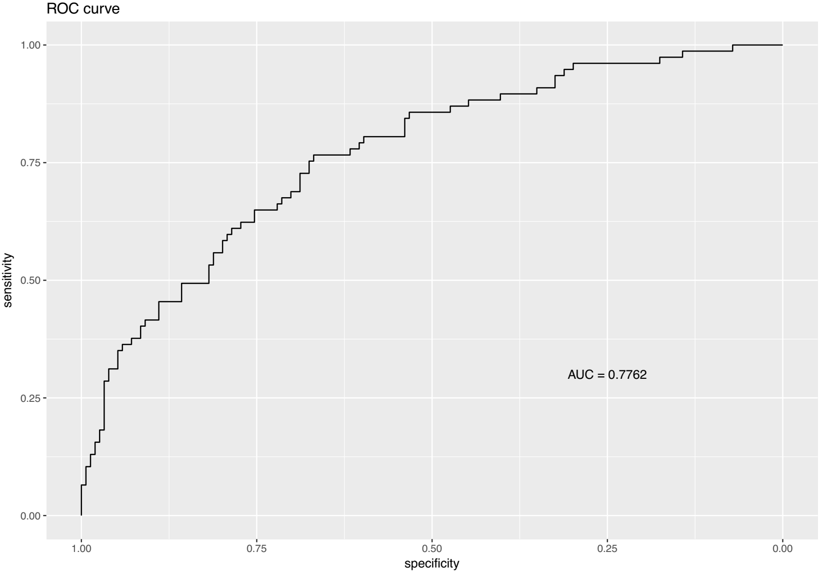
Classifiation problem

Explain about a problem with 2 or more classes, for example the South African heart disease data set (from module 4). Traning and test set.

- Write down the fitted model. The estimated coefficient for `famhist` is 1.047. How can do explain the effect `famhist`? How would you evaluate the fit of this model?

```
##
## Call:
## glm(formula = chd ~ tobacco + famhist + typea + obesity + age,
##     family = "binomial", data = train_SA)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1165  -0.8491  -0.4142   0.9481   2.2283
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -6.452432   1.535920  -4.201 2.66e-05 ***
## tobacco         0.078834   0.035911   2.195  0.02815 *
## famhistPresent  1.047384   0.323689   3.236  0.00121 **
## typea           0.044812   0.017458   2.567  0.01026 *
## obesity        -0.003855   0.036471  -0.106  0.91581
## age             0.060659   0.014736   4.116 3.85e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

- To evaluate the model fit on a new test set an ROC curve is made. How is this curve constructed? Explain what you see and evaluate the goodness of the model.



Unsupervised learning problem

- Similar to the two tasks in Compulsory 3: Problem 3a and 3b – which is to comment on and recognize method used.
- Could also have small calculation task to show that you have understood concepts: Consider the following four observation of a two-dimensional random vector $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$

$$\mathbf{a} = \begin{pmatrix} 5 \\ 4 \end{pmatrix}, \mathbf{b} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \mathbf{c} = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \mathbf{d} = \begin{pmatrix} 4 \\ 0 \end{pmatrix}$$

  Calculate the matrix of pairwise Euclidean distances between the points. Use hierarchical clustering with single, complete and average linkage to cluster the points. Draw dendrograms. Assume that we want two clusters. Which two groups will then the two dendrograms give? [M10, exam TMA4270 1995 3a]

After TMA4268 - what is next?

Do you want to learn more about the methods we have looked at in this course? And also methods that are more tailored towards specific types of data? Then we have many statistics courses that you may choose from.

- V=spring semester
- H=autumn semester

Year 3 courses

- (H)TMA4265 Stochastic processes Important models for multivariate random variables: Markov chains, Poisson processes, birth-and-death processes in continuous time, Brownian motion and Gaussian processes. Approaches for stochastic simulation of random variables.
- (V)TMA4267 Linear statistical models Multiple linear regression. Analysis of variance. Experimental design. Multivariate normal distribution. Multiple testing.

Year 4/5 courses

At IMF

(V)TMA4180 Optimization 1.

First and second order necessary and sufficient (Karush-Kuhn-Tucker) optimality conditions for unconstrained and constrained optimization problems in finite-dimensional vector spaces. Basics of convex analysis and Lagrangian duality theory and their application to optimization problems and algorithms. An overview of modern optimization techniques and algorithms for smooth problems (including line-search/trust-region, quasi-Newton, interior point and active set methods, SQP and augmented Lagrangian approaches). Basic derivative-free and non-smooth optimization methods.

(V)TMA4250 Spatial statistics.

Parameter estimation, simulation and applications of Gaussian random fields, point fields and discrete Markov random fields. Examples from image analysis, and environmental and natural resource applications.

(V)TMA4275 Lifetime analysis.

Basic concepts in lifetime modelling. Censored observations. Nonparametric estimation and graphical plotting for lifetime data (Kaplan-Meier, Nelson-plot). Estimation and testing in parametric lifetime distributions. Analysis of lifetimes with covariates (Cox-regression, accelerated lifetime testing). Modelling and analysis of recurrent events. Nonhomogeneous Poisson-processes. Nelson-Aalen estimators.

(H)TMA4285 Time series models

Autoregressive and moving average based models for stationary and non-stationary time series. Parameter estimation. Model identification. Forecasting. ARCH and GARCH models for volatility. State space models (linear dynamic models) and the Kalman filter.

(H)TMA4295 Statistical inference.

Transformations and moments of random variables. Families of distributions. Inequalities and convergence theorems. Sufficient statistics. Frequentist and Bayesian estimators. Methods of constructing point estimators, interval estimators and hypothesis tests, and optimality of these. Asymptotic properties of estimators and hypothesis tests.

(V)TMA4300 Computational statistics.

Classical and Markov chain methods for stochastic simulation. Hierarchical Bayesian models and inference in these. The expectation maximisation (EM) algorithm. Bootstrapping, cross-validation and non-parametric methods.

(H)TMA4315 Generalized linear models.

Univariate exponential family. Multiple linear regression. Logistic regression. Poisson regression. General formulation for generalised linear models with canonical link. Likelihood-based inference with score function and expected Fisher information. Deviance. AIC. Wald and likelihood-ratio test. Linear mixed effects models with random components of general structure. Random intercept and random slope. Generalised linear mixed effects models. Strong emphasis on programming in R. Possible extensions: quasi-likelihood, over-dispersion, models for multinomial data, analysis of contingency tables, quantile regression.

Outside IMF

- (V)KLMED8005 Analysis of repeated measurements. Intensive schedule only a part of the semester. 5STP.
- (V+H?)SMED8002 Epidemiology 2. Intensive schedule only a part of the semester.
- (V)TDT4300 Datavarehus og datagruvedrift: need background on Algorithms and data structures, and data bases.
- (V)TDT4173 Maskinlæring og case-based reasoning Overlap in topics with TMA4268, but not in philosophy - and not in exam questions (previous years).
- (V)NEVR3004 Nevrale nettverk: if you want to work within neuroscience. Intensive schedule only a part of the semester.
- (?) [Computer vision] New course by Anette Stahl, Cybernetics.
- (?) [Censor systems] New course by Edmund F. Brekke, Cybernetics.

PhD courses

- (H)MA8704 Asympotic methods every autumn. Requires TMA4295.
- (V)MA8701 General statistical methods next time spring of 2019. We go in more detail into related topics to TMA4268. Requires TMA4295 and TMA4300. Nice to have TMA4180, TMA4285 and TMA4315.
- (V)MA8702 Computational statistics 2 next time spring of 2020. Requires TMA4300 and TMA4250.

Course evaluation in TMA4268

- Many topics covered, mostly understanding - not so much of mathematical proofs.
- Elements: Course set-up, modules (12), textbook, plenary lectures (18), interactive lectures (4), supervision sessions (19), compulsory exercises (3), quizzes.
- Rich set of resources vs. attendance. Competing with more theoretical courses? EiT conflict.

I missed you at the lectures ☹

Please answer the course evaluation (anonymous): https://kvass.svt.ntnu.no/TakeSurvey.aspx?SurveyID=tma4268v2018

and give feed back to the reference group for the final report of the course.

On behalf of the teaching staff - Thea, Martina, Andreas, Thiago and Mette-

thank you for attending this course - hope to see you for the exam supervision - and good luck on May 24!

— ♡ ♡ ♡ —

References

Efron, Bradley, and Trevor Hastie. 2016. *Computer Age Statistical Inference - Algorithms, Evidence, and Data Science*. Cambridge University Press.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2001. *The Elements of Statistical Learning*. Vol. 1. Springer series in statistics New York.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning*. Vol. 112. Springer.