732A99/732A68/ TDDE01 Machine Learning
Division of Statistics and Machine Learning
Department of Computer and Information Science

Computer lab 2 block 1

Instructions

- Create a report to the lab solutions in PDF.
- Be concise and do not include unnecessary printouts and figures produced by the software and not required in the assignments.
- **Include all your codes as an appendix into your report.**
- **Use set.seed(12345) for every piece of code that contains randomness**
- A typical lab report should 2-4 pages of text plus some amount of figures plus appendix with codes.
- The lab report should be submitted via LISAM before the deadline.

Assignment 1. Explicit regularization

The **tecator.csv** contains the results of study aimed to investigate whether a near infrared absorbance spectrum can be used to predict the fat content of samples of meat. For each meat sample the data consists of a 100 channel spectrum of absorbance records and the levels of moisture (water), fat and protein. The absorbance is -log10 of the transmittance measured by the spectrometer. The moisture, fat and protein are determined by analytic chemistry.

Divide data randomly into train and test (50/50) by using the codes from the lectures.

1. Assume that Fat can be modeled as a linear regression in which absorbance characteristics (Channels) are used as features. Report the underlying probabilistic model, fit the linear regression to the training data and estimate the training and test errors. Comment on the quality of fit and prediction and therefore on the quality of model.
2. Assume now that Fat can be modeled as a LASSO regression in which all Channels are used as features. Report the cost function that should be optimized in this scenario.
3. Fit the LASSO regression model to the training data. Present a plot illustrating how the regression coefficients depend on the log of penalty factor (log λ) and interpret this plot. What value of the penalty factor can be chosen if we want to select a model with only three features?
4. Repeat step 3 but fit Ridge instead of the LASSO regression and compare the plots from steps 3 and 4. Conclusions?

5. Use cross-validation with default number of folds to compute the optimal LASSO model. Present a plot showing the dependence of the CV score on $\log \lambda$ and comment how the CV score changes with $\log \lambda$. Report the optimal λ and how many variables were chosen in this model. Does the information displayed in the plot suggests that the optimal λ value results in a statistically significantly better prediction than $\log \lambda = -4$? Finally, create a scatter plot of the original test versus predicted test values for the model corresponding to optimal lambda and comment whether the model predictions are good.

Assignment 2. Decision trees and logistic regression for bank marketing

The data file **bank-full.csv** is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

Input variables:
# bank client data:
1 - age (numeric)
2 - job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
3 - marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
4 - education (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')
5 - default: has credit in default? (categorical: 'no','yes','unknown')
6 - housing: has housing loan? (categorical: 'no','yes','unknown')
7 - loan: has personal loan? (categorical: 'no','yes','unknown')
# related with the last contact of the current campaign:
8 - contact: contact communication type (categorical: 'cellular','telephone')
9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
10 - day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
11 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.
# other attributes:
12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)

13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
14 - previous: number of contacts performed before this campaign and for this client (numeric)
15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')
# social and economic context attributes
16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)
17 - cons.price.idx: consumer price index - monthly indicator (numeric)
18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)
19 - euribor3m: euribor 3 month rate - daily indicator (numeric)
20 - nr.employed: number of employees - quarterly indicator (numeric)

Output variable (target):
21 - y - has the client subscribed a term deposit? (binary: 'yes','no')

1. Import the data to R, **remove variable "duration"** and divide into training/validation/test as 40/30/30: use data partitioning code specified in Lecture 2a.
2. Fit decision trees to the training data so that you change the default settings one by one (i.e. not simultaneously):
   a. Decision Tree with default settings.
   b. Decision Tree with smallest allowed node size equal to 7000.
   c. Decision trees minimum deviance to 0.0005.

   and report the misclassification rates for the training and validation data. Which model is the best one among these three? Report how changing the deviance and node size affected the size of the trees and explain why.

3. Use training and validation sets to choose the optimal tree depth in the model 2c: study the trees up to 50 leaves. Present a graph of the dependence of deviances for the training and the validation data on the number of leaves and interpret this graph in terms of bias-variance tradeoff. Report the optimal amount of leaves and which variables seem to be most important for decision making in this tree. Interpret the information provided by the tree structure (not everything but most important findings).
4. Estimate the confusion matrix, accuracy and F1 score for the test data by using the optimal model from step 3. Comment whether the model has a good predictive power and which of the measures (accuracy or F1-score) should be preferred here.
5. Perform a decision tree classification of the test data with the following loss matrix,

$$L = \begin{array}{ccc} & & \textit{Predicted} \\ \textit{Observed} & \begin{array}{c} \textit{yes} \\ \textit{no} \end{array} & \begin{pmatrix} 0 & 5 \\ 1 & 0 \end{pmatrix} \end{array}$$

and report the confusion matrix for the test data. Compare the results with the results from step 4 and discuss how the rates has changed and why.

6. Use the optimal tree and a logistic regression model to classify the test data by using the following principle:
$$\hat{Y} = 1 \; if \; p(Y = 'good'|X) > \pi, otherwise \; \hat{Y} = 0$$
where $\pi = 0.05, 0.1, 0.15, \ldots 0.9, 0.95$. Compute the TPR and FPR values for the two models and plot the corresponding ROC curves. Conclusion? Why precision-recall curve could be a better option here?

Assignment 3. Principal components and implicit regularization

The data file **communities.csv** contains the results of studies of the crime level in the united states based on various characteristics of the given location. The main variable that is studied is ViolentCrimesPerPop which represents the total number of violent crimes per 100K population. The meaning of other variables can be found at:
https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime

1. Scale all variables except of ViolentCrimesPerPop and implement PCA by using function eigen(). Report how many features are needed to obtain at least 95% of variance in the data. What is the proportion of variation explained by each of the first two principal components?
2. Repeat PCA analysis by using princomp() function and make the score plot of the first principle component. Do many features have a notable contribution to this component? Report which 5 features contribute mostly (by the absolute value) to the first principle component. Comment whether these features have anything in common and whether they may have a logical relationship to the crime level. Also provide a plot of the PC scores in the coordinates (PC1, PC2) in which the color of the points is given by ViolentCrimesPerPop. Analyse this plot (hint: use **ggplot2** package).
3. Scale the original data: both features and response, split data into training and test (50/50) and estimate a linear regression model from training data in which ViolentCrimesPerPop is target and all other data columns are features. Compute training and test errors for these data and comment on the quality of model.
4. Implement a function that depends on parameter vector θ and represents the cost function for linear regression without intercept on the training data set. Afterwards, use BFGS method (optim() function without gradient specified)

to optimize this cost with starting point $\theta^0 = 0$ and compute training and test errors for every iteration number. Present a plot showing dependence of both errors on the iteration number and comment which iteration number is optimal according to the early stopping criterion. Compute the training and test error in the optimal model, compare them with results in step 3 and make conclusions.

a. **Hint 1**: don't store parameters from each iteration (otherwise it will take a lot of memory), instead compute and store test errors directly.
b. **Hint 2**: discard some amount of initial iterations, like 500, in your plot to make the dependences visible.

First read 'Course Information.PDF' at LISAM, folder 'Course documents'

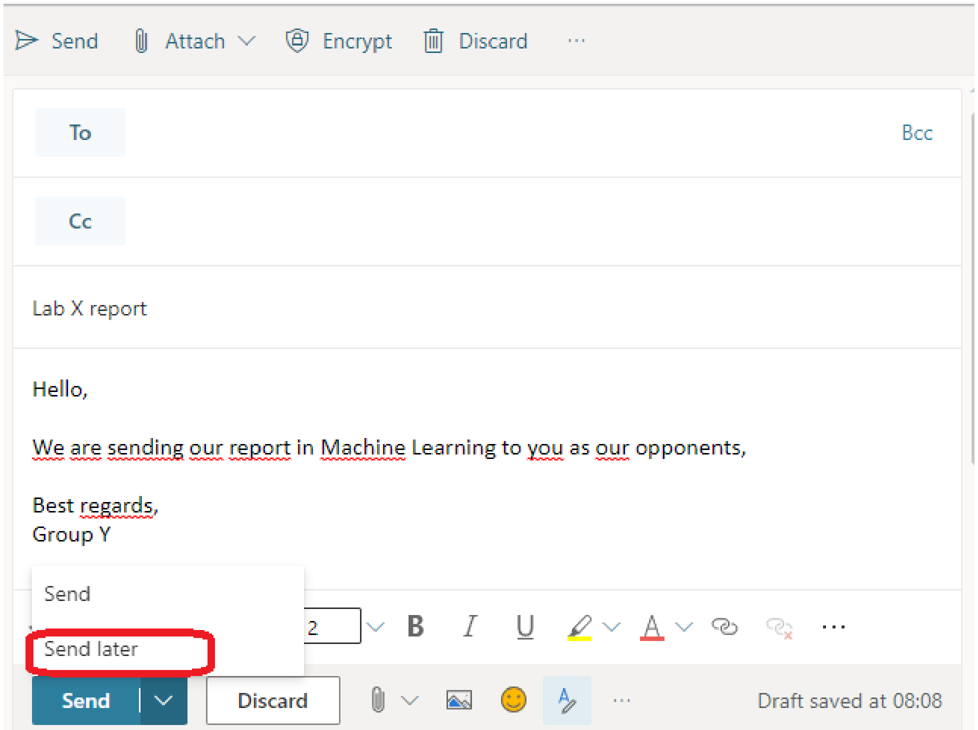
If you are neither speaker nor opponent for this lab,

- Make sure that you or your group mate submits the group report using *Lab X* item in the *Submissions* folder before the deadline. Make sure that the report contains the Statement Of Contribution describing how each group member has contributed into the group report.

If you are a speaker for this lab,

- Make sure that you or your group mate does the following before the deadline:
  1. submits the group report using *Lab X* item in the *Submissions* folder before the deadline. Makes sure that the report contains the Statement Of Contribution describing how each group member has contributed into the group report.
  2. Goes to LISAM→Course Documents→Deadlines.PDF, finds the deadline (date and time) for the current lab.
  3. Goes to LISAM→Course Documents→Seminars.PDF and find the group number of your opponent group
  4. Goes to LISAM→Course Documents→Groups.PDF and finds email addresses of the students in the opponent group
  5. Go to LISAM→Outlook app and in the Outlook web client creates a new message where you
     - Specify Lab X report as a title (X is lab number)
     - Specify email addresses of the opponents in the "To:" field
     - Attach your group PDF report.

- **Important: Click on arrow next to "Send" button, choose "Send Later" and specify the lab deadline as the message delivery time stamp (see figure)**



Lab X report

Hello,

We are sending our report in Machine Learning to you as our opponents,

Best regards,
Group Y

If you are opponent for this lab,

- Make sure that you or your group mate submits the group report using *Lab X* item in the *Submissions* folder before the deadline. Make sure that the report contains the Statement Of Contribution describing how each group member has contributed into the group report.
- After the deadline for the lab has passed you should be able to receive the PDF report of the speakers per email. Compile it, read it carefully and prepare (in cooperation with your group comrade) **at least three questions/comments/improvement suggestions per lab assignment** in order to put them at the seminar.