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

Computer lab 1 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.
- Some exercises that require implementations of methods are provided with a skeleton of the code (see LISAM), you are free to use this code or implement your own code from scratch.
- **Include all your codes as an appendix into your report.**
- 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. Handwritten digit recognition with K-means

The data file **optdigits.csv** contains information about normalized bitmaps of handwritten digits from a preprinted form from a total of 43 people. The data were first derived as 32x32 bitmaps which were then divided into nonoverlapping blocks of 4x4 and the number of on pixels are counted in each block. This has generated the resulting image of size 8x8 where each element is an integer in the range 0..16. Accordingly, each row in the data file is a sequence corresponding to 8x8 matrix, and the last element shows the actual digit from 0 to 9.

1. Import the data into R and divide it into training, validation and test sets (50%/25%/25%) by using the partitioning principle specified in the lecture slides.
2. Use training data to fit 30-nearest neighbor classifier with function *kknn()* and kernel="rectangular" from package *kknn* and estimate
    - Confusion matrices for the training and test data (use table())
    - Misclassification errors for the training and test data

    Comment on the quality of predictions for different digits and on the overall prediction quality.
3. Find any 2 cases of digit "8" in the training data which were easiest to classify and 3 cases that were hardest to classify (i.e. having highest and lowest probabilities of the correct class). Reshape features for each of these cases as matrix 8x8 and visualize the corresponding digits (by using e.g. heatmap() function with parameters Colv=NA and Rowv=NA) and comment on whether these cases seem to be hard or easy to recognize visually.
4. Fit a K-nearest neighbor classifiers to the training data for different values of $K = 1,2, \dots, 30$ and plot the dependence of the training and validation misclassification

errors on the value of K (in the same plot). How does the model complexity change when K increases and how does it affect the training and validation errors? Report the optimal K according to this plot. Discuss this plot from the perspective of bias-variance tradeoff. Finally, estimate the test error for the model having the optimal K, compare it with the training and validation errors and make necessary conclusions regarding the model quality.

5. Fit K-nearest neighbor classifiers to training data for different values of $K = 1,2, \dots, 30$, compute the empirical risk for the validation data as cross-entropy (when computing log of probabilities add a small constant within log, e.g. 1e-15, to avoid numerical problems) and plot the dependence of the empirical risk on the value of K. What is the optimal K value here? Why might the cross-entropy be a more suitable choice of the empirical risk function than the misclassification error for this problem?

Assignment 2. Ridge regression and model selection

The data file **parkinson.csv** is composed of a range of biomedical voice measurements from 42 people with early-stage Parkinson's disease recruited to a six-month trial of a telemonitoring device for remote symptom progression monitoring. The purpose is to predict Parkinson's disease symptom score (motor UPDRS) from the following voice characteristics:

- Jitter(%),Jitter(Abs),Jitter:RAP,Jitter:PPQ5,Jitter:DDP - Several measures of variation in fundamental frequency
- Shimmer,Shimmer(dB),Shimmer:APQ3,Shimmer:APQ5,Shimmer:APQ11,Shimmer:DDA - Several measures of variation in amplitude
- NHR,HNR - Two measures of ratio of noise to tonal components in the voice
- RPDE - A nonlinear dynamical complexity measure
- DFA - Signal fractal scaling exponent
- PPE - A nonlinear measure of fundamental frequency variation

1. Assuming that motor_UPDRS is normally distributed and can be modeled by Ridge regression of the voice characteristics, write down the probabilistic model as a Bayesian model.
2. Scale the data and divide it into training and test data (60/40). Due to this, compute all models without intercept in the following steps.
3. Implement 4 following functions by using basic R commands only (no external packages):
    a. $Loglikelihood$ function that for a given parameter vector $\boldsymbol{w}$ and dispersion σ computes the log-likelihood function $\log P(D|w, \sigma)$ for the model from step 1 for the training data
    b. $Ridge$ function that for given vector $\boldsymbol{w}$, scalar σ and scalar λ uses function from 2a and adds up a Ridge penalty to the minus log-likelihood

   c. $RidgeOpt$ function that depends on scalar λ , uses function from 2b and function optim() with method="BFGS" to find the optimal w and σ for the given λ.
   d. DF function that for a given scalar λ computes the degrees of freedom of the regression model from step 1 based on the training data.
4. By using function RidgeOpt, compute optimal w parameters for $\lambda = 1, \lambda = 100$ and $\lambda = 1000$. Use the estimated parameters to predict the motor_UPDRS values for training and test data and report the training and test MSE values. Which penalty parameter is most appropriate among the selected ones? Why is MSE a more appropriate measure here than other empirical risk functions?
5. Use functions from step 3 to compute AIC (Akaike Information Criterion) scores for the Ridge models with values $\lambda = 1, \lambda = 100$ and $\lambda = 1000$ and their corresponding optimal parameters **w** and σ computed in step 4. What is the optimal model according to AIC criterion? What is the theoretical advantage of this kind of model selection compared to the holdout model selection done in step 4?

Assignment 3. Linear regression and LASSO

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 objective 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 ($\log \lambda$) 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. Present a plot of how degrees of freedom depend on the penalty parameter. Is the observed trend expected?
5. Repeat step 3 but fit Ridge instead of the LASSO regression and compare the plots from steps 3 and 5. Conclusions?
6. Use cross-validation 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. Comment whether the selected λ value is statistically significantly better than $\log \lambda = -2$. 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.
7. Use the feature values from test data (the portion of test data with Channel columns) and the optimal LASSO model from step 6 to generate new target values. (Hint: use rnorm() and compute σ as standard deviation of residuals from train data predictions). Make a scatter plot of original Fat in test data versus newly generated ones. Comment on the quality of the data generation.

Submission procedure

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

Assume that X is the current lab number, Y is your group number.

If you are neither speaker nor opponent for this lab,

- Submit your report using *Lab X* item in the *Submissions* folder before the deadline.
- Make sure that you or some of your group members submits the group report using *Lab X group report* in the *Submissions* folder before the deadline

If you are a speaker for this lab,

- Submit your report using *Lab X* item in the *Submissions* folder before the deadline.
- Make sure that you or some of your group members does the following before the deadline:
    - submits the group report using *Lab X group report* in the *Submissions* folder before the deadline

- Goes to *Sumissions* and opens item *Password X*. Then the student should put your group report into ZIP file *Lab X_Group Y.zip* and protect it with a password you found in Password X
- Uploads the file to *Collaborative workspace* →*Lab X* folder

If you are opponent for this lab,

- Submit your report using *Lab X* item in the *Submissions* folder before the deadline.
- Make sure that you or some of your group members submits the group report using *Lab X group report* in the *Submissions* folder before the deadline
- After the deadline for the lab has passed, go to Collaborative workspace→*Lab X* folder and download the appropriate ZIP file. Open the PDF in this ZIP file by using the password available in *Submission* →*Password X* item, read it carefully and prepare (in cooperation with other group members) **at least three questions/comments/improvement suggestions per lab assignment** in order to put them at the seminar.