732A95 Introduction to Machine Learning
Division of Statistics and Machine Learning
Department of Computer and Information Science

Computer lab 3 block 2

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. High-dimensional methods

The data file **data.csv** contains information about 64 e-mails which were manually collected from DBWorld mailing list. They were classified as: 'announces of conferences' (1) and 'everything else' (0) (variable Conference)

1. Divide data into training and test sets (70/30) without scaling. Perform nearest shrunken centroid classification of training data in which the threshold is chosen by cross-validation. Provide a centroid plot and interpret it. How many features were selected by the method? List the names of the 10 most contributing features and comment whether it is reasonable that they have strong effect on the discrimination between the conference mails and other mails? Report the test error.
2. Compute the test error and the number of the contributing features for the following methods fitted to the training data:
    a. Elastic net with the binomial response and $\alpha = 0.5$ in which penalty is selected by the cross-validation
    b. Support vector machine with “vanilladot” kernel.

    Compare the results of these models with the results of the nearest shrunken centroids (make a comparative table). Which model would you prefer and why?
3. Implement Benjamini-Hochberg method for the original data, and use t.test() for computing p-values. Which features correspond to the rejected hypotheses? Interpret the result.

732A95 Introduction to Machine Learning
Division of Statistics and Machine Learning
Department of Computer and Information Science

Assignment 2. Online learning

Implement the budget online SVM. Check the course slides for the pseudo-code. **Use the template below**. Note that you are not using all the attributes and points in the file. Run your code on the spambase.csv file for the (M, beta) values (500,0), (500,-0.05), (20,0) and (20,-0.05). Plot the error rate as a function of the number of training point, and analyze the results. In particular,

- explain why (500,0) gives better results than (500,-0.05).
- explain why (20,-0.05) gives a smoother error rate plot than (20,0).
- explain why (20,0) is the slowest.

```
set.seed(1234567890)
spam <- read.csv2("spambase.csv")
ind <- sample(1:nrow(spam))
spam <- spam[ind,c(1:48,58)]
h <- 1
beta <- # Your value here
M <- # Your value here
N <- 500 # number of training points

gaussian_k <- function(x, h) { # Gaussian kernel
# Your code here
}

SVM <- function(sv,i) { # SVM on point i with support vectors sv
# Your code here
# Note that the labels in spambase.csv are 0/1 and SVMs need -1/+1. Then, use 2*label-1
# to convert from 0/1 to -1/+1
# Do not include the labels when computing the Euclidean distance between the point i
# and each of the support vectors. This is the distance to use in the kernel function
# You can use dist() to compute the Euclidean distance
}

errors <- 1
errorrate <- vector(length = N)
errorrate[1] <- 1
sv <- c(1)
for(i in 2:N) {
# Your code here
}
plot(errorrate[seq(from=1, to=N, by=10)], ylim=c(0.2,0.4), type="o")
length(sv)
errorrate[N]
```

732A95 Introduction to Machine Learning
Division of Statistics and Machine Learning
Department of Computer and Information Science

Submission procedure

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 Study room *Group Y→Documents* and opens file *Password X.txt*. 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.txt
    - 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 *Course Documents→Password X.txt*, 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.