UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in:	STK2100 — Maskinlæring og statistiske metoder for prediksjon og klassifikasjon
Day of examination:	June 10th, 2020
Examination hours:	9.00 – June 17th, 9:00
This problem set consists of 4 pages.	
Appendices:	None.
Permitted aids:	None.

Please make sure that your copy of the problem set is complete before you attempt to answer anything.

Problem 1 Golub et al. (1999) data

Consider the first dataset of the second mandatory assignment, namely the Golub et al. (1999)'s data containing 7128 gene expressions of 72 patients with leukaemia. As you know, data can be found at `http://web.stanford.edu/~hastie/CASI_files/DATA/leukemia_big.csv`. Split the data into a training and a test set, with the latter containing the following observations (obtained by a stratified random split): `ALL.4, ALL.8, ALL.10, ALL.11, ALL.13, ALL.18, AML, AML.1, AML.4, AML.6, AML.8, ALL.23, ALL.26, ALL.29, ALL.31, ALL.32, ALL.35, ALL.39, ALL.40, ALL.41, ALL.42, AML.16, AML.22, AML.24`. All the other observations belong to the training set. Here the label "`ALL`" denotes patients with acute lymphoblastic leukaemia, "`AML`" those with acute myeloid leukaemia.

a Penalized logistic regression

Use lasso and ridge logistic regression (i.e., logistic regression with an L_1 and L_2 penalties, respectively) to fit two alternative models to classify the patients between AML and ALL. In both cases, use a cross-validation procedure to select the best penalty among these values for λ: $\{e^i, i = -7, -6, \ldots, 1, 2\}$. In particular, choose a deterministic cross-validation procedure.
For this exercise, provide:

a.1 the best value for λ for both the lasso and the ridge models, together with an explanation of the role of this parameter **(8 pt)**;

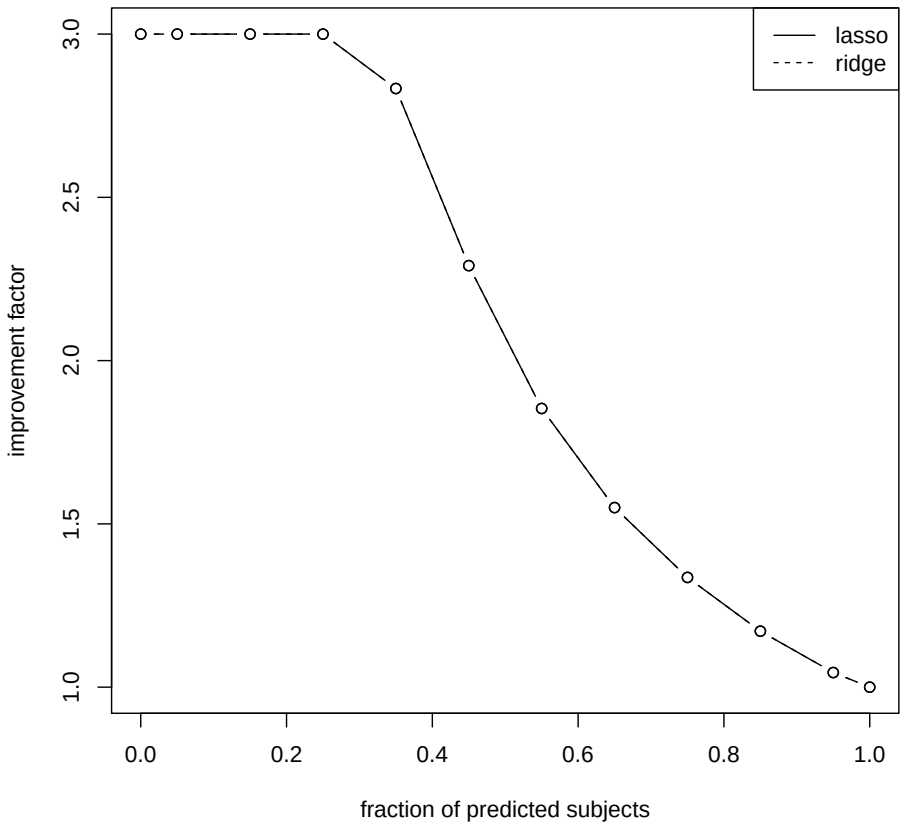
a.2 the misclassification error both in the training and the test set, for both the lasso and the ridge models; briefly explain, in addition, why we expect that, in general, the training errors are smaller than the test errors **(8 pt)**;

(Continued on page 2.)

a.3 a justification for the choice of the particular cross-validation procedure, adding an advantage and a disadvantage with respect to other cross-validation choices **(4 pt)**.

b Models assessment

While evaluating the two models on the test data, the following lift plot has been obtained,



As you can see, the lines related to the lasso and ridge models are indistinguishable:

b.1 how do you explain this situation, which seems in contrast to the different misclassification test errors computed in point **a.2 (13 pt)**?

b.2 what does the point $(0.25, 3)$ in this plot tell us **(7 pt)**?

c Pre-selection

Sometimes, in order to reduce the dimensionality of the problem, a preliminary selection step is performed. One possibility is to perform univariate two-sample t-tests on all the variables, and only keep those with the smallest p-values.

(Continued on page 3.)

Perform the following procedure:

1. perform a two-sample t-test on the whole data set, finding the 9 dimensions (gene expressions) for which the differences in mean between the AML and ALL patients are the highest;
2. split the dataset in a training and a test set (where the sets only consists of the 9 genes selected in step 1), using the same split of exercise **a**;
3. fit a logistic regression model on the training set obtained at step 2;
4. compute the misclassification error on the test set.

For this exercise:

c.1 report the name (variable number) of the 9 genes found in step 1, explain why step 1 is necessary if one wants to implement step 3 as it is written in the procedure above, and report the misclassification error computed in step 4 **(3 pt)**;

c.2 explain why this procedure is wrong, in the sense that we cannot compare the misclassification error of step 4 with those obtained for lasso and ridge logistic regression at point **a.2 (9 pt)**;

c.3 correct the procedure above in order to obtain an estimate of the misclassification test error that allows us to compare the "logistic regression with 9 pre-selected genes" with lasso and ridge logistic regression. Report the number of the correctly selected 9 genes and of the correct estimate of the misclassification error **(8 pt)**.

d Non-hierarchical clustering

Merge again training and test set (i.e., consider the whole dataset). We want to evaluate the two kinds of cluster analysis seen in the course, hierarchical and non-hierarchical. Let us start with the latter: apply a K-mean algorithm with 10 different initialization points for the centroids. In particular, use as initialization points the following set of observations (consider all the possible combinations): `{ALL, ALL.10, ALL.11, AML, AML.2}`.
For this exercise:

d.1 since in this particular case we know the truth, i.e. which are the two real groups, report the misclassification error for all 10 different results. Comment on the results, explaining why it is possible that they are so different **(8 pt)**;

d.2 comment on the choice of K: why it is in general problematic, but it is not in this specific example; why it is not possible to compute its best values by cross-validation; propose a strategy to select it **(12 pt)**.

(Continued on page 4.)

e Hierarchical clustering

Try now a hierarchical cluster analysis method, using, in particular, an agglomerative strategy and a complete link measure.

e.1 Report the dendrogram and comment on the results, given that you know the truth (i.e., which observations belong to the AML and to the ALL groups) **(8 pt)**.

Since we saw in the second mandatory assignment that it is graphically possible to separate the two groups based on the first two principal components, repeat the analysis of point **e.1** using only the two first principal components.

e.2 Provide a plot that shows the percentage of the original variance "contained" in the 72 principal components **(4 pt)**;

e.3 Report the dendrogram for the new analysis (i.e., those only using the first two principal components as input) and comment on the results **(8 pt)**.

References

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C. & Lander, E. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537.

THE END