UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in:	STK2100 — Machine learning and statistical methods for prediction and classification
Day of examination:	Friday, May 31st, 2019
Examination hours:	9.00 – 13.00
This problem set consists of 6 pages.	
Appendices:	None.
Permitted aids:	Approved calculator and List of formulas for STK1100/STK1110 and STK2100.

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

Problem 1 Classification

Consider the data by Rosenberg et al. (2003, Statistics in Medicine) that you analysed in your first mandatory assignment. They belong to a case-control study on the risk of oral cancer in the African-American population. For those involved in the study, information about two risk factors, `drinks` (average number of ounces of alcoholic drinks per week) and `cigs` (originally the number of cigarettes smoked per day, here dichotomized in Non-smoker/Smoker), has been collected, together with their gender (`sex`) and age (`age`). The response variable, `cancer`, indicates who experienced oral cancer (i.e., `cancer = 1` for those who experienced it).

a

Let us start with logistic regression. Fitting the full model on the training data, one obtains the following results,

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.2422	0.9407	-2.38	0.0171
drinks	0.0260	0.0061	4.27	0.0000
age	0.0143	0.0146	0.98	0.3267
cigsSmoker	0.6337	0.3898	1.63	0.1040
sexMale	0.5170	0.3543	1.46	0.1446

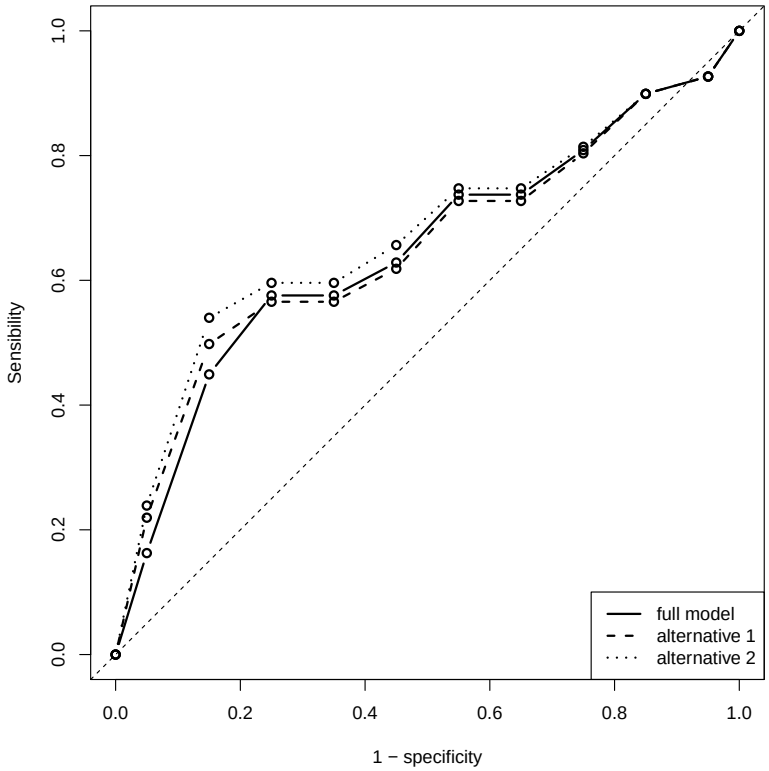
Is there any protecting factor (i.e., a variable that reduces the chance of experiencing oral cancer) among those analysed in the study?

How can one interpret the estimate of the regression coefficient for the intercept? Does it have a "physical" meaning, considering the variable `age`? What can be done in order to solve the issue related to `age`?

b

Consider two alternative models, `alternative 1` and `alternative 2`, in which some variables have been excluded. When evaluated on the test set, they produce the following ROC curves,

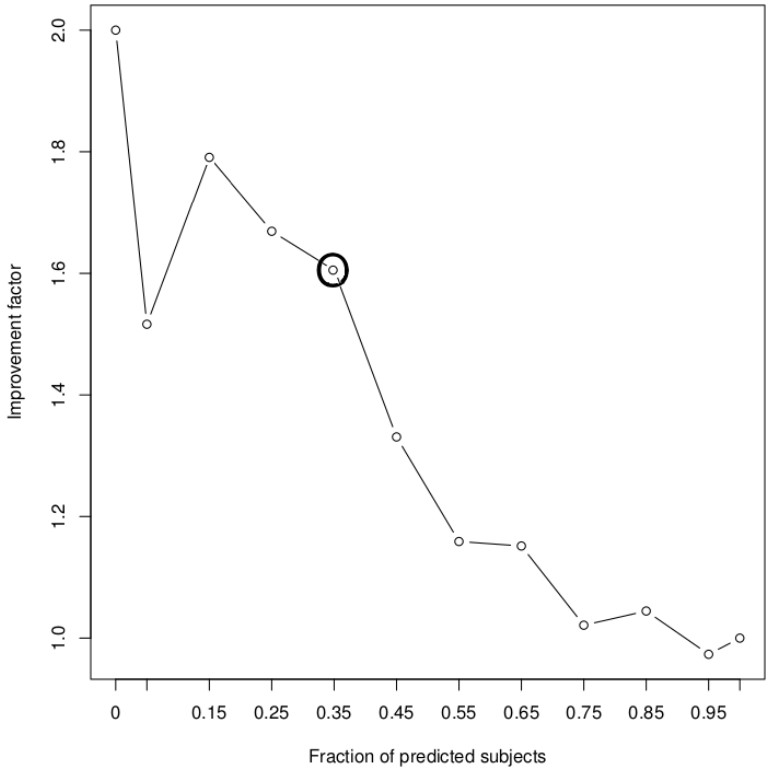
(Continued on page 2.)



Which model would you use for prediction? Why? Explain what is a ROC curve, defining accurately the concepts of sensitivity and specificity.

c

The best of the three models above gives the following lift curve,



(Continued on page 3.)

Explain how to interpret this plot and, in particular, what the point highlighted with a bold circle tells.

d

Consider the following logistic regression model, in which only the variable `cigs` is included,

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.96141	0.3261	-2.95	0.0032
cigsSmoker	1.11964	0.3644	3.07	0.0021

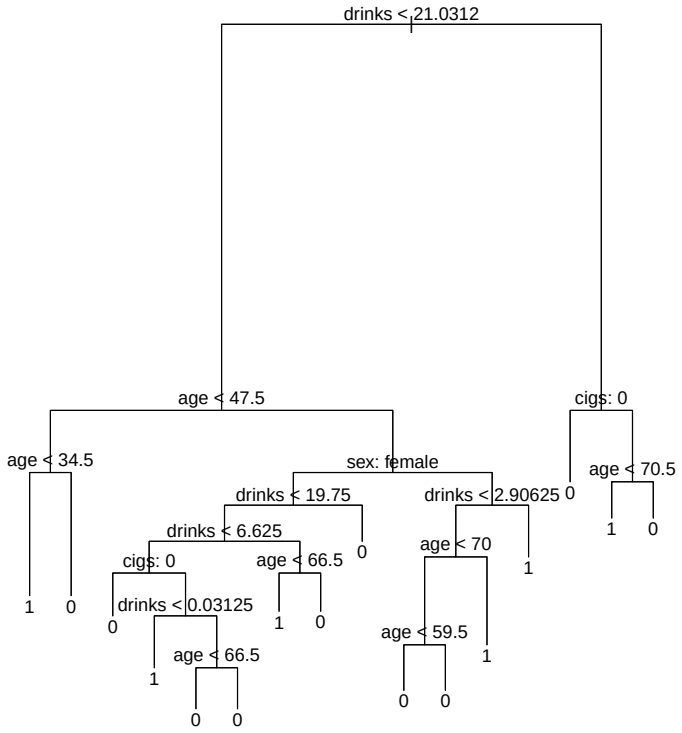
.

Use the estimated values of the regression coefficients to assign the right values to a_{11}, a_{12}, a_{21} and a_{22} in the following 2×2 table,

		`cigs`		
		Non-smoker	Smoker	Total
`cancer`	no	a_{11}	a_{12}	tot_{no}
	yes	a_{21}	a_{22}	tot_{yes}
	Total	47	152	199

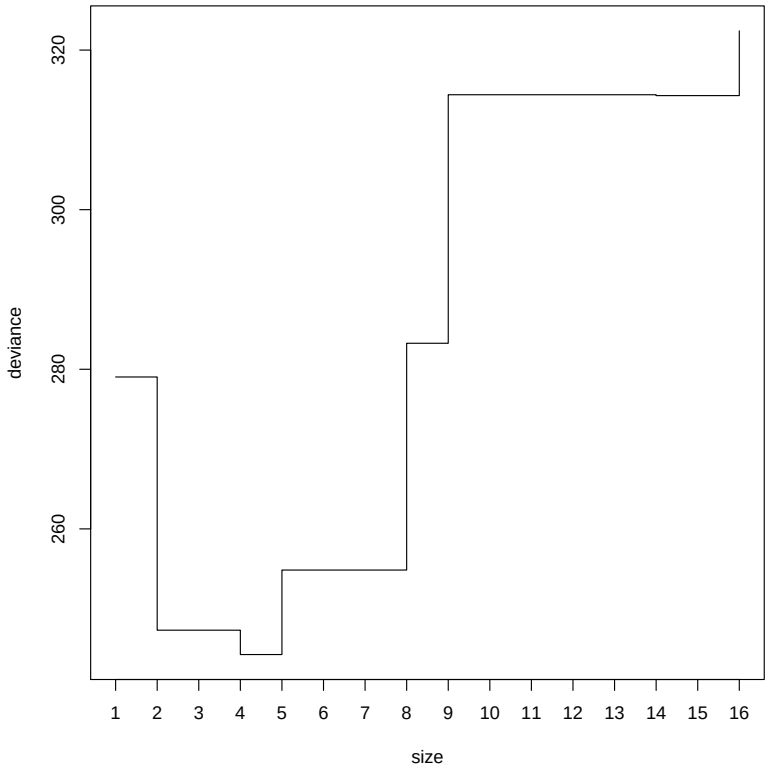
e

Consider the following plot, which visualizes the model obtained by fitting a classification tree on the training data,



Which is the predicted response value for a 40-year old man who smokes and drinks, on average, 16 ounces of alcoholic drinks per week?

(Continued on page 4.)



In order to prune the tree above, a 10-fold cross-validation procedure has been implemented. The results are reported in the following figure,

Provide a description of cross-validation, and explain why it is used here to select the number of leaves (in the plot, denoted by `size`). Finally, draw a pruned tree (from the first one) that takes into account the result of the cross-validation procedure and provides the same prediction for the man described in the previous sentence.

Problem 2 Clustering

a

Many clustering methods are built upon the concept of dissimilarity. Consider the following decomposition of the *total dissimilarity*,

$$\sum_{ij} d(i,i') = \sum_{k=1}^{K} \sum_{G(i)=k} \sum_{G(i')=k} d(i,i') + \sum_{k=1}^{K} \sum_{G(i)=k} \sum_{G(i')\neq k} d(i,i')$$

where $G(i)$ indicates the cluster that the i-th observation belongs to, K is the total number of clusters and $d(i,i') = \sum_i d_j(x_{ij}, x_{i'j}))$ is the dissimilarity between observation i and i'. Here x_{ij} denotes the j-th component of the observation i.

Explain what the two terms on the right-hand side of the formula represent and discuss why minimizing one of them corresponds to maximizing the other. Moreover, explain why, in most of the cases, it is important to introduce some form of normalization for the variables X_j (note that x_{ij} is the value of the variable X_j for the observation i).

(Continued on page 5.)

b

Now consider the Euclidean distance to construct the dissimilarities. Then,

$$\sum_{k=1}^{K} \sum_{G(i)=k} \sum_{G(i')=k} d(i,i') = 2 \sum_{k=1}^{K} \sum_{G(i)=k} ||x_i - m_k||^2,$$

where x_i, $i = 1, \ldots, n$, denotes the i-th observation and m_k, $k = 1, \ldots, K$, the arithmetic mean vector of the observations belonging to the group k.

In this case, the following algorithm can be used,

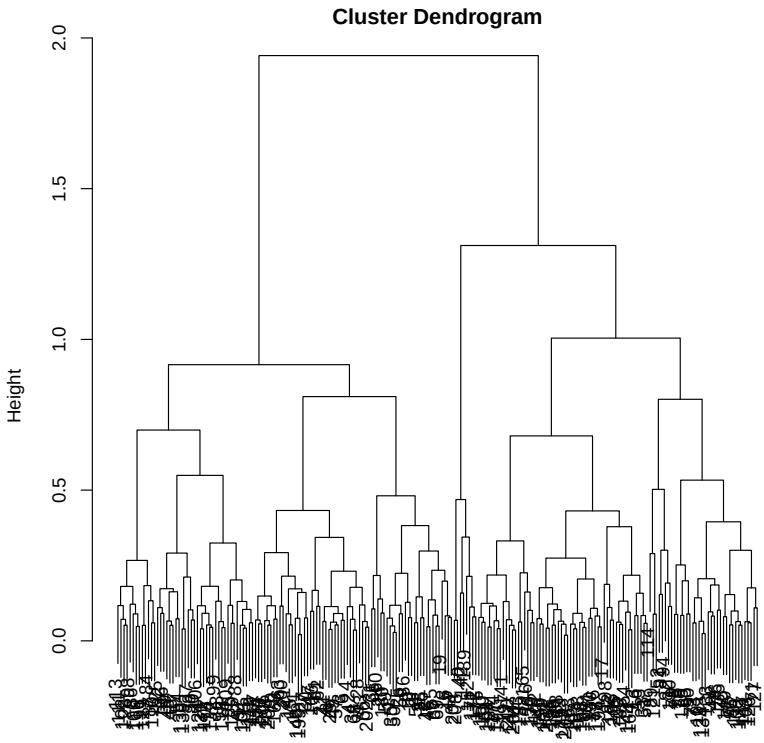
Algorithm 1
1. Choose K and initialize $m_1, \ldots, m_K$.
2. Cycle for $r = 1, 2, \ldots$
(a) for $i = 1, \ldots, n$, assign x_i to group k, so that $\|\|x_i - m_k\|\|^2$ is minimum;
(b) for $k = 1, \ldots, K$, let m_k be equal to the arithmetic mean of the subjects belonging to group k;
until $m_1, \ldots, m_K$ stabilize.

Which method does Algorithm 1 implement?

Discuss the role of points 2.(a) and 2.(b) in the algorithm and the limitations of the method.

c

Consider the following dendrogram, obtained when using the *complete link* to measure the dissimilarity among clusters,



(Continued on page 6.)

Explain the difference between *single link* and *complete link* and how the choice between the two may influence the clustering algorithm.

Based on the dendrogram, provide a reasonable value for K, the number of clusters, explaining the reason for your choice. Moreover, discuss why in this case a cross-validation procedure cannot be profitably used to find K.

Problem 3 Non-parametric estimation

a

In contrast to *k-nearest-neighbor*, *local regression methods* weight the influence of an observation to the estimation of the response based on its distance from the point of interest. The weights usually take the form

$$w_i = \frac{1}{h} w\left(\frac{x_i - x_0}{h}\right)$$

for a certain kernel $w(\cdot)$.

Discuss the role of h and relate it to the concept of bias-variance trade-off. Briefly explain why the concept of bias-variance trade-off is important for prediction.

b

It is well known that non-parametric methods are particularly affected by the curse of dimensionality. Discuss the problem and provide a possible way to overcome it.

THE END