Exam of Statistical Learning

2025-11-11

Exercises

In this exam, you are asked to analyse a dataset from a simulated study on **hypertension management**. The dataset represents information collected from adult patients enrolled in a 6-month intervention programme aimed at reducing systolic blood pressure. Your task is to explore and model the relationship between the **reduction in systolic blood pressure (in mmHg)** and a set of clinical, behavioural, and treatment-related covariates.

Variable	Type	Description
`y`	Numeric	Reduction in systolic blood pressure after 6 months (mmHg).
`age`	Numeric	Patient age (30–80 years).
`sex`	Factor (Male/Female)	Biological sex.
`baseline_sbp`	Numeric	Baseline systolic blood pressure (mmHg).
`BMI`	Numeric	Body Mass Index (kg/m^2).
`smoker`	Factor (No/Yes)	Smoking status.
`exercise_level`	Factor (Low/Moderate/High)	Level of physical activity.
`salt_intake`	Numeric	Daily salt intake (grams/day).
`treatment`	Factor (Control/DrugA/DrugB)	Assigned treatment arm.
`adherence`	Numeric (0–1)	Fraction of days the patient adhered to therapy.
`diabetes`	Factor (No/Yes)	Diabetes status.

In **all exercises involving supervised methods**, you should use a training set containing 70% of observations, whose raw indexes are selected with the following code, and a random seed equal to 1120.

```
n = dim(data.all)[1]
set.seed(seed)
select.train = sample(1:n,n*7/10)
train = data.all[select.train,]
test = data.all[-select.train,]
```

Answer to all following questions in the markdown file. Pay attention to motivate (with code or text) **all** your answers.

Exercise 1 (points: 1+3+3)

Run a k-means clustering on the data set, setting $k = 4$. Use ten random initializations of the algorithm and set the seed to 1120.

a. Write down the estimate of the model parameters: the centers of the clusters $\boldsymbol{\mu}_k$ and the assignments r_{ik} to points to clusters. Also, obtain the value of the loss $J = \sum_{i=1}^{n} \sum_{k=1}^{K} r_{ik} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2$.

b. Now retry the clustering with a number of clusters k going from 1 to 10. Remember to set the seed before each clustering run. Plot J as a function of k. **How many clusters would you suggest to use?**

c. Now cluster the data with the suggested number of clusters. Plot the data colored according to the cluster labels. **Give a comment on the R script on the obtained plot, and try to give an interpretation to the obtained clusters.**

Exercise 2 (points: 1+3+3)

Create a categorical response with two levels: 1 (Y higher than thres=17) and 0 (Y lower than 17), using the following R code:

```r
Y_cat = factor(ifelse(data.all$Y < thres,0,1))
data.all.cat = data.all
data.all.cat$Y = Y_cat
train.cat = data.all.cat[select.train,]
test.cat = data.all.cat[-select.train,]
```

Using all covariates, run a Support Vector Machine based on a radial kernel, using a cost equal to 10 and gamma equal to 1.

a. Compute the classification accuracy on the training and test set.

b. Now, tune the cost and gamma parameters, exploring the values of 0.1, 1, 10 for the cost and 1, 2 for gamma. Compute the confusion matrix and classification accuracy for the best model and **write a comment on the result**.

c. Plot the roc curve of the classification obtained at points a. and b. **Write a comment on the comparison between the two curves.**

Exercise 3 (points: 1+3+3)

Run a GAM the data set, using smoothing splines with 3 degrees of freedom for all numerical covariates.

a. Compute the training and test set MSE of the GAM model.
b. Perform a suitable p-value adjustment based on FWER control on the p-values of the summary table. **Write a comment on the significance of nonlinear effects.**
c. Reduce the model eliminating the non significant covariate (using either the summary table, the plot of effects, or a combination of both). **Write a comment on the effect of the remaining covariates on the response.**

Exercise 4 (points: 1+3+3)

Run a boosting model on the data set, setting the following parameters:

- number of trees: $B = 5000$
- shrinkage: $\lambda = 0.01$
- interaction depth $d = 4$

Remember to set the seed to 1120 before fitting the boosting model. Then, answer to the following questions.

a. Compute the train and test set MSE of the boosting model.
b. Now, recompute the MSE for d spanning from 1 to 7. **Which value would you suggest for the parameter d? What is the interpretation of such value?**
c. **What can you say about the most informative covariates? Write a comment on the markdown file.**

Exercise 5 (4 points)

Fit a linear model on the dataset, and compute the test set MSE.

In the markdown file, write a comment on the model that in your opinion performs better on this data set among the ones that you fitted for explaining the relationship between the outcome and the covariates. Describe this model in terms of how the covariates influence the response. Finally, specify whether (and how) you would suggest to improve the model.