Ensimag – Grenoble INP – UGA
Introduction to Statistical Learning and Applications
Pedro L. C. Rodrigues
Alexandre Wendling

Year 2023-2024

pedro.rodrigues@inria.fr
alexandre.wendling@univ-grenoble-alpes.fr

# TD 3: Some questions from previous exams

## ▶ Exercise 1 (credits to EPFL CS-433)

We consider a classification problem on linearly separable data. Our dataset had an outlier – a point that is very far from the other datapoints in distance. We trained the linear discriminant analysis (LDA), logistic regression and 1-nearest-neighbour classifiers on this dataset. We tested the trained models on a test set that comes from the same distribution as the training set, but doesn't have any outlier points. After that, we removed the outlier and retrained our models.

After retraining, which classifier will **not change** its decision boundary around the test points.

(A) Logistic regression.

(B) 1-nearest-neighbors classifier.

(C) LDA.

(D) None of them.

## ▶ Exercise 2 (credits to EPFL CS-433)

In principal component analysis, the left singular vectors $\mathbf{U}$ of a data matrix $\mathbf{X} \in \mathbf{R}^{N \times p}$ are used to create a new data matrix $\mathbf{X}' = \mathbf{U}^\top \mathbf{X}$, where $N$ is the number of data points and $p$ is the number of features. Which property always holds for the matrix $\mathbf{X}'$?

(A) $\mathbf{X}'$ is a square matrix.

(B) The mean of any row $\mathbf{X}'_i$ is 0

(C) $\mathbf{X}'$ has only positive values.

(D) For any two rows $i, j$ with $i \neq j$ from $\mathbf{X}'$, the dot product between the rows $\mathbf{X}'_i$ and $\mathbf{X}'_j$ is 0.

## ▶ Exercise 3 (credits to EPFL CS-433)

Consider the logistic regression loss $\mathcal{L} : \mathbf{R}^p \to \mathbf{R}$ for a binary classification task with data $(\mathbf{x}_i, y_i) \in \mathbf{R}^p \times \{0, 1\}$:

$$\mathcal{L}(\boldsymbol{\beta}) = \frac{1}{N} \sum_{i=1}^{N} \left( \log\left(1 + e^{\mathbf{x}_i^\top \boldsymbol{\beta}}\right) - y_i \mathbf{x}_i^\top \boldsymbol{\beta} \right)$$

Which of the following is a gradient of the loss $\mathcal{L}$?

(A) $\nabla \mathcal{L}(\boldsymbol{\beta}) = \dfrac{1}{N} \displaystyle\sum_{i=1}^{N} \left( \mathbf{x}_i \frac{e^{\mathbf{x}_i^\top \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^\top \boldsymbol{\beta}}} - y_i \mathbf{x}_i^\top \boldsymbol{\beta} \right)$

(B) $\nabla \mathcal{L}(\boldsymbol{\beta}) = \dfrac{1}{N} \displaystyle\sum_{i=1}^{N} \mathbf{x}_i \left( y_i - \frac{e^{\mathbf{x}_i^\top \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^\top \boldsymbol{\beta}}} \right)$

(C) $\nabla \mathcal{L}(\boldsymbol{\beta}) = \dfrac{1}{N} \displaystyle\sum_{i=1}^{N} \mathbf{x}_i \left( \frac{1}{1 + e^{-\mathbf{x}_i^\top \boldsymbol{\beta}}} - y_i \right)$

(D) None of the above.

## ▶ Exercise 4

In this exercise you will perform multiple linear regression on simulated data under different conditions. To ensure reproducibility on your results, set the seed of the random with `set.seed(0)`.

(a) Simulate a dataset of size $N = 1000$ of the following generating model:

$$\begin{aligned} X_{1,i} &= \varepsilon_{1,i} \\ X_{2,i} &= 3X_{1,i} + \varepsilon_{2,i} \\ Y_i &= X_{2,i} + X_{1,i} + 2 + \varepsilon_{3,i} \end{aligned}$$

where $i \in \{1, \ldots, N\}$ and the $\varepsilon_{ij}$ are independent $\mathcal{N}(0,1)$ random variables. For a given $i$, what is the distribution of $(X_{1,i}, X_{2,i})$? Plot the clouds of points of the simulated values of $(X_{1,i}, X_{2,i})_{i=1,\ldots,n}$. What is its shape? Can you write an analytical formula for it?

(b) Let us consider the following two regression models:

$$\begin{aligned} \text{Model A:} \quad Y_i &= \alpha_1 X_{1,i} + \alpha_0 + \tilde{\varepsilon}_{A,i} \\ \text{Model B:} \quad Y_i &= \beta_2 X_{2,i} + \beta_0 + \tilde{\varepsilon}_{B,i} \end{aligned}$$

where $\tilde{\varepsilon}_{A,i} \sim \mathcal{N}(0, \sigma_A^2)$ and $\tilde{\varepsilon}_{B,i} \sim \mathcal{N}(0, \sigma_B^2)$. What should be the values of $\hat{\alpha}_0, \hat{\alpha}_1, \hat{\sigma}_A^2, \hat{\beta}_0, \hat{\beta}_2, \hat{\sigma}_B^2$ when $N \to \infty$? Consider $N = 1000$ and check whether the estimates of the parameters are close to the true values that you've calculated. Now do `set.seed(3)` and simulate again a dataset $X_{1,i}, X_{2,i}, Y_i$ for $n = 10$. Estimate the parameters. What happens?

(c) Let us now consider the full model

$$Y_i = \gamma_2 X_{2,i} + \gamma_1 X_{1,i} + \gamma_0 + \varepsilon_i$$

where $i \in \{1, \ldots, n\}$ and the $\varepsilon_i$ are independent $\mathcal{N}(0, \sigma^2)$ random variables. For the previously simulated data with $n = 10$, estimate $\hat{\gamma}_0, \hat{\gamma}_1, \hat{\gamma}_2, \hat{\sigma}^2$ and compare them with the parameters obtained in item (b). What can you say about the effects of $X_1$ and $X_2$ on $Y$? And about their correlation?

## ▶ Exercise 5

We consider the dataset `cars04`, which describes several properties of different car models in the market in 2004. Each observation (i.e. car) is described by 11 features (i.e. properties) listed in Table 1.

| Variable | Meaning |
|---|---|
| `Retail` | Builder recommended price(US$) |
| `Dealer` | Seller price (US$) |
| `Engine` | Motor capacity (liters) |
| `Cylinders` | Number of cylinders in the motor |
| `Horsepower` | Engine power |
| `CityMPG` | Consumption in city (Miles or gallon; proportional to km/liter) |
| `HighwayMPG` | Consumption on roadway (Miles or gallon) |
| `Weight` | Weight (pounds) |
| `Wheelbase` | Distance between front and rear wheels (inches) |
| `Length` | Length (inches) |
| `Width` | Width (inches) |

Table 1: Variable list for `cars04`

The aim of this exercise is to summarize and to interpret the data `cars04` using PCA. Using `R` we run the following instruction:

```
cars04.pca <- prcomp(cars04, scale=TRUE)
summary(cars04.pca)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.6655 1.3726 0.92181 0.59751 0.52482 0.44491 0.37486
## Proportion of Variance 0.6459 0.1713 0.07725 0.03246 0.02504 0.01799 0.01277
## Cumulative Proportion  0.6459 0.8171 0.89439 0.92685 0.95189 0.96988 0.98266
##                            PC8     PC9    PC10    PC11
## Standard deviation     0.29434 0.25766 0.19229 0.02811
## Proportion of Variance 0.00788 0.00604 0.00336 0.00007
## Cumulative Proportion  0.99053 0.99657 0.99993 1.00000
```

(a) What is the effect of the argument `scale=TRUE` in the result of the PCA?

(b) Are the first two principal components enough to summarize most of the information (i.e. variance) of the dataset? Justify in terms of the proportion of the total variance that they represent.

Principal components are linear combinations of the 11 variables. The coefficients of the first 2 principal components on the 11 feature are

```
cars04.pca$rotation[,1:2]
```

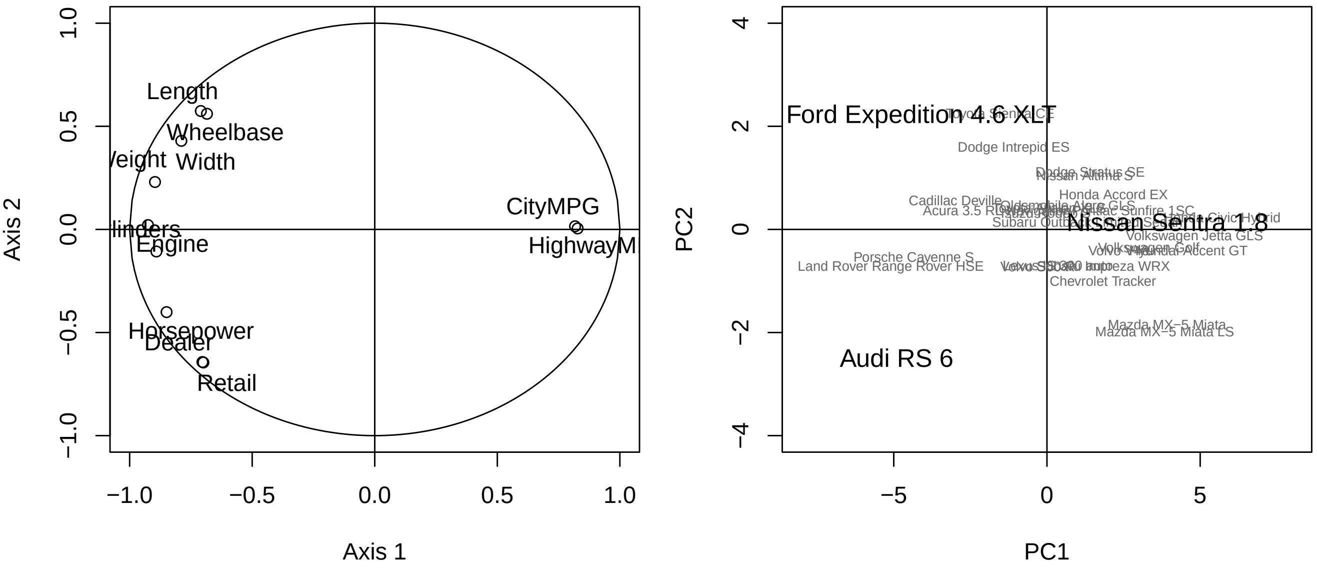
```
##                   PC1          PC2
## Retail     -0.2637504 -0.468508698
## Dealer     -0.2623186 -0.470146585
## Engine     -0.3470805  0.015347186
## Cylinders  -0.3341888 -0.078032011
## Horsepower -0.3186023 -0.292213476
## CityMPG     0.3104817  0.003365936
## HighwayMPG  0.3065886  0.010964460
## Weight     -0.3363294  0.167463572
## Wheelbase  -0.2662100  0.418177107
## Length     -0.2567902  0.408411381
## Width      -0.2960546  0.312891350
```

(c) What would be a good interpretation for these new variables in terms of the initial features of the dataset?

Figure 1 shows the projection of the dataset on its first two principal components.

(d) Interpret each quadrant of the figure.

(e) Can you describe which kind of car Audi RS 6, Ford Expedition 4.6 XLT and Nissan Sentra 1.8 are?



(a) Variable space

(b) Individual space

Figure 1: Principal component representation in the first plane of the variable and of the sample spaces.