Multiple choice test
Jan. 29th, 2019

Exam duration: 2 hours.

- *Document allowed: 1 sheet A4 format (both sides). No calculators, no laptops, no tablets, no mobile phone...*
- *The test contains 7 exercises with 24 questions. Exercises 1-2-3 refer to the "R for data science" lecture and exercises 4-5-6-7 to the "Introduction to machine learning" lecture.*
- *Questions using the sign ♣ may have one or several correct answers. Other questions have a single correct answer.*
- *Only the last sheet (answer sheet page 9) is to be returned. You can keep all the other pages.*
- *Squares corresponding to good answers have to be* **colored with a black pen**. *Cross or circle marks are not sufficient! It is not possible to correct (once a square has been colored).*
- *The scoring process is as follows:*
  - *No answer to one question $\Longrightarrow$ 0 point for the question.*
  - *Questions with a single correct answer: positive score for a good answer, negative score for a bad answer.*
  - *Questions with several correct answers (sign ♣): positive score for each good answer, negative score for each bad answer.*

Exercise 1. This exercise deals with data importation and data merging.

Question 1 File data1.txt (saved in the working directory of R) contains the following dataset

```
name;size;age
John;174;32
Peter;?;28
Mary;165.5;NA
Steve;173;?
```

Among the following commands, which one provides the following output.

```
> data1
  name size age
1 John 174.0 32
2 Peter NA 28
3 Mary 165.5 NA
4 Steve 173.0 NA
```

A `data1 <- read.table("data1.txt",sep=";",header=TRUE,na.strings=c(" "))`

B `data1 <- read.table("data1.txt",sep=";",header=FALSE,na.strings=c("?","NA"))`

C `data1 <- read.table("data1.txt")`

D `data1 <- read.table("data1.txt",sep=";",header=TRUE,na.strings=c("?","NA"))`

E `data1 <- read.table("data1.txt",sep=";",header=1,na.strings=c("?"),row.names = 1)`

F None of these answers are correct.

We assume that the following dataset has also been imported.

```
> data2
  name1 weight
1 John   75
2 Mary   68
3 Fred   42
```

So, From now on, we have two datasets in this exercise: `data1` and `data2`.

Question 2 Among the following commands, which one provides the following output.

```
  name size age weight
1 John 174.0 32 75
2 Mary 165.5 NA 68
```

A `inner_join(data1,data2,by=c("name"="name1"))`

B `inner_join(data1,data2,by=c("name1"="name"))`

C `full_join(data1,data2,by=c("name"="name1"))`

D `full_join(data1,data2,by=c("name1"="name"))`

E None of these answers are correct.

Question 3 Among the following commands, which one provides the following output.

```
  name size age weight
1 John 174.0 32 75
2 Peter NA 28 NA
3 Mary 165.5 NA 68
4 Steve 173.0 NA NA
5 Fred NA NA 42
```

A `full_join(data1,data2,by=c("name1"="name"))`

B `inner_join(data1,data2,by=c("name"="name1"))`

C `inner_join(data1,data2,by=c("name1"="name"))`

D `full_join(data1,data2,by=c("name"="name1"))`

E None of these answers are correct.

Exercise 2. This exercise is about `dplyr` package.

Question 4 ♣ Among the following sentences, which ones are correct?

A `mutate` verb allows to filter individuals in a dataframe.

B `mutate` verb allows to create variables in a dataframe.

C `arrange` verb allows to reorder individuals according to the values of a variable.

D `select` verb allows to filter individuals in a dataframe.

E `arrange` verb allows to create variables in a dataframe.

F `group_by` verb allows to apply operations for group of individuals.

G *None of these answers are correct.*

We now consider the `iris` dataset presented in the lecture. For simplicity, we will only consider the 5 individuals presented below.

```
> iris1
# A tibble: 5 x 5
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
*        <dbl>       <dbl>        <dbl>       <dbl> <fct>
1          4.7         3.2          1.3         0.2 setosa
2          4.9         2.4          3.3         1   versicolor
3          5.7         2.5          5           2   virginica
4          4.6         3.1          1.5         0.2 setosa
5          5.1         2.5          3           1.1 versicolor
```

Question 5 When running the following code

```
iris1 %>% summarize(PL=mean(Petal.Length))
```

the output is

A 5 B 2.74 C 2.82 D 0.9
E None of these answers are correct.

Question 6 When running the following code

```
iris1 %>% group_by(Species) %>% summarize(M=mean(Petal.Width)) %>% arrange(desc(M))
```

the output is

A 2.82

B
```
Species M
<fct> <dbl>
1 setosa 0.2
2 versicolor 1.05
3 virginica 2
```

C
```
Species M
<fct> <dbl>
1 virginica 2
2 versicolor 1.05
3 setosa 0.2
```

D
```
Species M
<fct> <dbl>
1 virginica 5
2 versicolor 3.15
3 setosa 1.4
```

E
```
Species M
<fct> <dbl>
1 setosa 1.4
2 versicolor 3.15
3 virginica 5
```

F None of these answers are correct.

Question 7 When running the following code

```
iris1 %>% summarize(M=mean(Petal.Width)) %>% group_by(Species)
```

the output is

A 2.82

B
```
Species M
<fct> <dbl>
1 virginica 2
2 versicolor 1.05
3 setosa 0.2
```

C
```
Species M
<fct> <dbl>
1 setosa 0.2
2 versicolor 1.05
3 virginica 2
```

D
```
Species M
<fct> <dbl>
1 virginica 5
2 versicolor 3.15
3 setosa 1.4
```

E
```
Species M
<fct> <dbl>
1 setosa 1.4
2 versicolor 3.15
3 virginica 5
```

F None of these answers are correct.

Question 8 ♣ Among the following commands, which ones provide the following output.

```
Petal.Length Petal.Width
<dbl> <dbl>
1 1.3 0.2
2 3.3 1
3 5 2
4 1.5 0.2
5 3 1.1
```

A `iris1 %>% select(3:4)`
B `iris1 %>% filter(3:4)`
C `iris1 %>% select(contains("Petal"))`
D `iris1 %>% mutate(contains("Petal"))`
E `iris1 %>% slice(Petal)`
F `iris1[,3:4]`
G *None of these answers are correct.*

Exercise 3. This exercise is about `ggplot2` package.

For these questions on `ggplot` we consider the full iris dataset available on `R`. For each question of this exercice, only one answer is correct.

Question 9 Among the following commands, which one provides Figure 1 page 5.

A `ggplot(iris)+aes(x=Petal.Length)+geom_bar(bins=10)`
B `ggplot(iris)+aes(x=Petal.Length)+geom_histogram(bins=10)`
C `ggplot(iris)+geom_histogram(x=Petal.length,bins=10)`
D `ggplot(iris)+geom_histogram(y=Petal.length,bins=10)`
E `ggplot(iris)+aes(x=Petal.Length,y=count)+geom_histogram(bins=10)`

Question 10 Among the following commands, which one provides Figure 2 page 5.

A `ggplot(iris)+geom_boxplot(x=Species,y=Petal.Length)`
B `ggplot(iris)+group_by(x=Species,y=Petal.Length)+geom_boxplot()`
C `ggplot(iris)+aes(x=Species,y=Petal.Length)+geom_bar()`
D `ggplot(iris)+aes(y=Petal.Length)+geom_boxplot(x=Species)`
E `ggplot(iris)+geom_boxplot()+aes(x=Species,y=Petal.Length)`

Question 11 Among the following commands, which one provides Figure 3 page 5.

A `ggplot(iris)+aes(x=Petal.Length,y=Petal.Width)+geom_point(shape=Species,size=3)`
B `ggplot(iris)+aes(x=Petal.Length,y=Petal.Width,color=Species)+geom_point(size=3)`
C `ggplot(iris)+aes(x=Petal.Length,y=Petal.Width)+geom_point(size=3)`
D `ggplot(iris)+aes(x=Petal.Length,y=Petal.Width)+group_by(Species)+geom_point(size=3)`
E `ggplot(iris)+aes(y=Petal.Width,x=Petal.Length,shape=Species)+geom_point(size=3)`

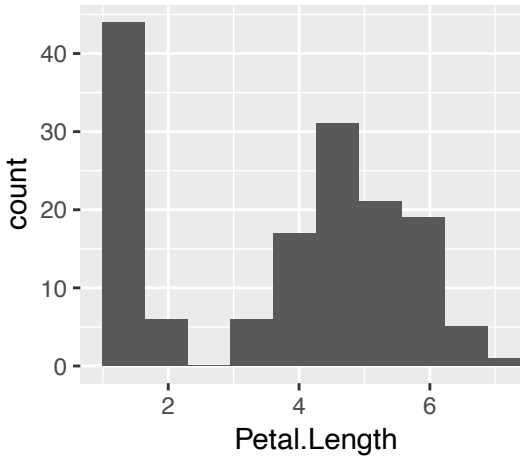

Figure 1: Question 9.

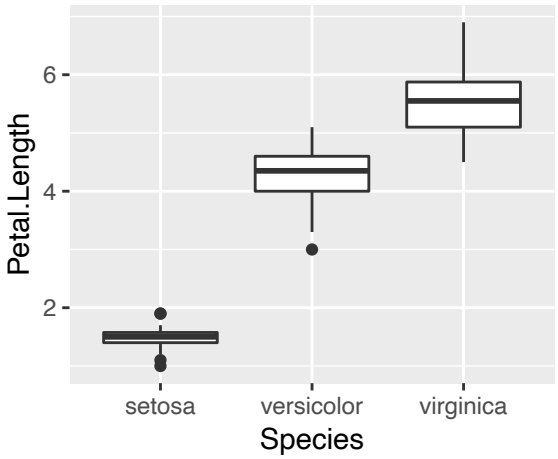

Figure 2: Question 10.

Figure 3: Question 11.

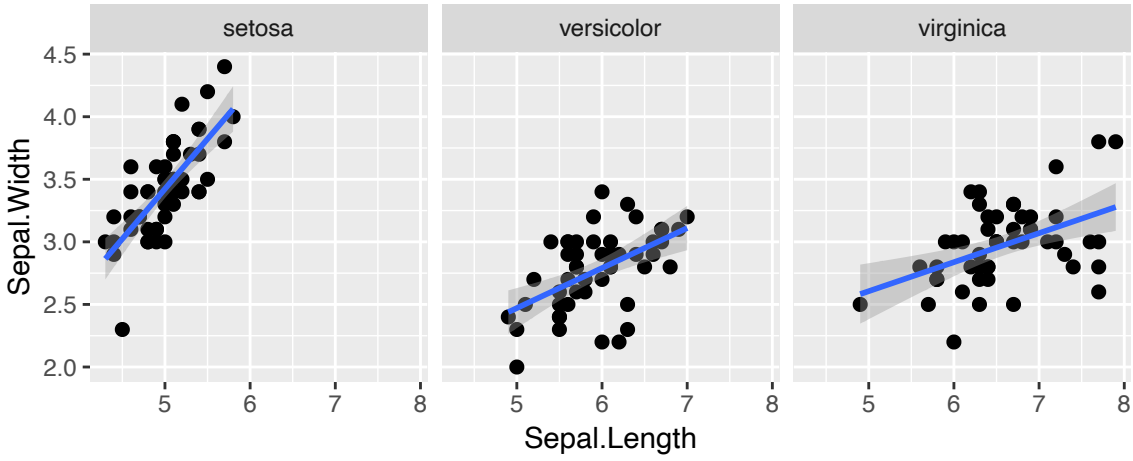

Figure 4: Question 12.

Question 12 Among the following commands, which one provides Figure 4 page 5.

A
```
ggplot(iris)+aes(x=Sepal.Length,y=Sepal.Width)+geom_point(size=2)+
geom_smooth(method="lm")+group_by(Species)
```

B
```
ggplot(iris)+aes(x=Sepal.Length,y=Sepal.Width,facet=Species)+geom_point(size=2)+
geom_smooth(method="lm")
```

C
```
ggplot(iris)+aes(x=Sepal.Length,y=Sepal.Width)+geom_point(size=2)+
geom_smooth(method="lm")+facet_wrap(~Species)
```

D
```
ggplot(iris)+aes(x=Sepal.Length,y=Sepal.Width,color=Species)+geom_point(size=2)
+geom_smooth(method="lm")
```

E
```
ggplot(iris) %>% group_by(Species)+aes(x=Sepal.Length,y=Sepal.Width)+
geom_point(size=2)+geom_smooth(method="lm")
```

Exercise 4. This exercise deals with general questions of machine learning.

Question 13 ♣ Among the following functions, which ones may be used as cost functions for a regression problem.

A $\ell(y, y') = (y - y')^2$.
B $\ell(y, y') = y - y'$.
C $\ell(y, y') = |y - y'|$.
D $\ell(y, y') = y + y'$.
E $\ell(y, y') = \frac{1}{2}(y - y')^2$.
F $\ell(y, y') = yy'$.
G *None of these answers are correct.*

Question 14 For a machine m we denote by $\mathcal{R}(m) = \mathbf{E}[(Y - m(X))^2]$ its quadratic risk. Let $m^\star$ denote the optimal machine for the quadratic risk. Among the following answers, which one is correct.

A $\forall m,\ \mathcal{R}(m^\star) > \mathcal{R}(m)$.
B $\forall m,\ \mathcal{R}(m^\star) = \mathcal{R}(m)$.
C $\forall m,\ \mathcal{R}(m^\star) \leq \mathcal{R}(m)$.
D $\forall m,\ \mathcal{R}(m^\star) > \mathcal{R}(m) + 1$.

Question 15 ♣ Let f_k and f_h denote the k-nearest neighbor and the kernel estimate with bandwidth h for a regression problem. Among the following answers, which ones are correct.

A f_k tends to overfit when k is too small.
B f_k tends to overfit when k is too large.
C f_h tends to overfit when h is too large.
D f_h tends to overfit when h is too small.
E *None of these answers are correct.*

Question 16 ♣ Let g be a classification rule which suffers from overfitting. Among the following choices, which ones are correct.

A Observations in a test dataset are very well predicted by g.
B Observations in the train data set are very well fitted by g.
C g has a large bias and a small variance.
D g has a small bias and a large variance.
E g has a small bias and a small variance.
F *None of these answers are correct.*

Exercise 5. This exercises is about penalized regressions.

In this exercise we consider a n i.i.d. sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ where $X_i = (X_{i1}, \ldots, X_{ip})$ takes values in $\mathbb{R}^p$ (with p large) and Y_i in $\mathbb{R}$. We consider the linear model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + \varepsilon_i.$$

Question 17 ♣ Among the following answers, which ones are correct.

- A Penalized regressions allow to reduce the variance of the least squares estimates.
- B Penalized algorithms allow to reduce the bias of the least squares estimates.
- C Ridge/lasso estimates always overperform least square estimates.
- D Ridge and lasso methods are widely used for large values of p.
- E Bias of ridge estimates is larger than the bias of the least squares estimate.
- F *None of these answers are correct.*

Question 18 Let $\lambda \geq 0$. Lasso estimates are calculated by minimizing the least square criterion penalized by

- A $\lambda \sum_{j=1}^{p} |\beta_j|$
- B $\lambda \sum_{j=1}^{p} \sqrt{\beta_j}$
- C $\lambda \sum_{j=1}^{p} \log(\beta_j^2)$
- D $\lambda \sum_{j=1}^{p} \beta_j^2$
- E $\lambda \sum_{j=1}^{p} \log(|\beta_j|)$
- F None of these answers are correct.

Question 19 ♣ We consider lasso estimates defined by the (correct) penalty term proposed at the previous question. Among the following choices, which ones are correct.

- A Lasso estimates are closed to 0 for very small values of λ.
- B Lasso estimates are closed to least squares estimates for very large values of λ.
- C Lasso estimates are closed to 0 for large values of λ.
- D Lasso estimates are closed to least squares estimates for very small values of λ.
- E We always have to choose λ as large as possible.
- F *None of these answers are correct.*

Question 20 ♣ Among the following answers, which ones are correct.

- A Function **glmnet** allows to compute lasso estimates.
- B Function **glmnet** allows to select the λ parameter in lasso regression.
- C Function **glmnet** allows to fit regression trees.
- D Function **cv.glmnet** allows to select the λ parameter in lasso regression.
- E Option lambda in **glmnet** function allows to specify if we want to make ridge or lasso regression.
- F Option alpha in **glmnet** function allows to specify if we want to make ridge or lasso regression.
- G *None of these answers are correct.*

Exercise 6. This exercises is about trees.

`rpart` function has been used to fit a sequence of tree for a binary classification problem. This sequence is assigned in the R object `tree`. `plotcp` function provides the following output:

```
> printcp(tree)$cptable
      CP nsplit rel error xerror xstd
1  0.2941176 0  1.000000 1.00000 0.053870
2  0.1225490 1  0.705882 0.71569 0.049838
3  0.0931373 3  0.460784 0.49020 0.043844
4  0.0637255 4  0.367647 0.43627 0.041928
5  0.0122549 5  0.303922 0.34314 0.038034
6  0.0098039 7  0.279412 0.35532 0.038034
7  0.0049020 9  0.259804 0.36275 0.038923
8  0.0040107 25 0.181373 0.38804 0.038260
9  0.0036765 41 0.112745 0.39216 0.040184
10 0.0032680 49 0.083333 0.40196 0.040586
11 0.0024510 52 0.073529 0.41176 0.040980
12 0.0001000 82 0.000000 0.43137 0.041742
```

We consider 2 trees defined by:

```
> tree1 <- prune(tree,cp=0.1225490)
> tree2 <- prune(tree,cp=0.0001000)
```

Question 21 ♣ Among the following choices, which ones are correct.

- A `tree1` overfits.
- B `tree2` has more cuts than `tree1`.
- C `tree1` has a large variance.
- D `tree2` overfits.
- E `tree2` has a large bias.
- F *None of these answers are correct.*

Question 22 Among the following trees, which one is the best (according to the pruning strategy proposed in the course).

- A `prune(tree,cp=0.2941176)`
- B `prune(tree,cp=0.0024510)`
- C `prune(tree,cp=0.34314)`
- D `prune(tree,cp=0.303922)`
- E `prune(tree,cp=0.0122549)`
- F `prune(tree,cp=0.0001000)`
- G None of these answers are correct.

Exercise 7.

In this exercise we consider a n i.i.d. sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ with the same distribution as (X, Y). We assume that X takes values in $\mathbb{R}$ and Y in $\{0, 1\}$. Let g be a classification rule and $L(g) = \mathbf{P}(g(X) \neq Y)$ be its error probability . We also denote by $g^\star$ the Bayes rule.

Question 23 ♣ Let $x \in \mathbb{R}$. Among the following choices, which ones are correct.

- A $L(g^\star) \leq L(g)$
- B $g^\star(x) = 1$ si $\mathbf{P}(Y = 1|X = x) = 0.75$
- C $L(g^\star) > L(g)$
- D $g : \mathbb{R} \to \{0, 1\}$
- E $g^\star(x) = 1$ if $\mathbf{P}(Y = 1|X = x) < 0.5$
- F $g : \{0, 1\} \to \mathbb{R}$
- G $g^\star(x) = 1$ si $\mathbf{P}(Y = 1|X = x) \geq 0.51$
- H $g^\star(x) = 1$ if $\mathbf{P}(Y = 1|X = x) = 0.15$
- I *None of these answers are correct.*

Question 24 ♣ Moreover, we assume that X has standard normal distribution $\mathcal{N}(0, 1)$ and that, for $x \in \mathbb{R}$, the conditional distribution of $Y|X = x$ is

- a Bernoulli distribution $\mathcal{B}(0.75)$ if $x \geq 0$;
- a Bernoulli distribution $\mathcal{B}(0.20)$ if $x < 0$;

Among the following choices, which ones are correct.

- A $g^\star(x) = 1$ if $x \geq 0$
- B $g^\star(1) = 0$
- C $g^\star(x) = 0$ si $x \geq 0$
- D $g^\star(-1) = 0$
- E $L(g^\star) = 0$
- F $L(g^\star) = 1$
- G $L(g^\star) = 9/40$
- H $L(g^\star) = 9/80$
- I $L(g^\star) = 31/80$
- J $L(g^\star) = 0.25$
- K *None of these answers are correct.*

Answer sheet:

Firstname and lastname:

.....

Answers must be given exclusively on this sheet: answers given on the other sheets will be ignored.

Question 1: A B C D E F

Question 2: A B C D E

Question 3: A B C D E

Question 4: A B C D E F G

Question 5: A B C D E

Question 6: A B C D E F

Question 7: A B C D E F

Question 8: A B C D E F G

Question 9: A B C D E

Question 10: A B C D E

Question 11: A B C D E

Question 12: A B C D E

Question 13: A B C D E F G

Question 14: A B C D

Question 15: A B C D E

Question 16: A B C D E F

Question 17: A B C D E F

Question 18: A B C D E F

Question 19: A B C D E F

Question 20: A B C D E F G

Question 21: A B C D E F

Question 22: A B C D E F G

Question 23: A B C D E F G H I

Question 24: A B C D E F G H I J K