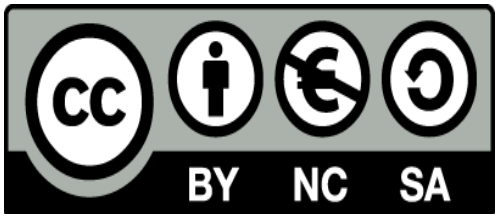
uc3m | Universidad Carlos III de Madrid



OPENCOURSEWARE
ADVANCED PROGRAMMING
STATISTICS FOR DATA SCIENCE

Ricardo Aler

MASTER IN STATISTICS FOR DATA SCIENCE. UC3M.

ADVANCED PROGRAMMING.

2 HOURS

---

1. What is “Rcpp sugar” and why it is useful?

2. Explain the difference between dynamic and static typing.

3. Explain what *data-leakage* is. Give an actual example of *data-leakage*.

4. What is the difference betwen a shallow copy and a deep copy? Explain with an example (code not required, you can use a picture if you prefer).

5. What is the main difference between an iterator and a list?

6. What is meant by “*broadcasting*” in *Numpy*? Write an example where *broadcasting* is used and explain why it is useful (correct syntax is not important).

7. What are universal functions? Is *np.sum()* a universal function (*np.sum*([1,2,3])=6)? Why?

8. Both Numpy 2-dimensional arrays and Pandas dataframes are 2-dimensional structures. What is the main difference between them?

9. If you want to do boolean selection in Pandas, what could be used, *.loc* or *.iloc*? Why?

10. If the following C++ function is compiled:

```
NumericVector f(NumericVector x)
{
  NumericVector out = x;
  out[0] = 3;
  return(out);
}
```

and then it is called from R, as follows, what would be printed by *print(x)* and *print(y)*? Why?

```
x <- c(1,2,3)
y = f(x)
print(x)
print(y)
```

11. Is the following code sensible? Why?

```
param_grid = {'max_depth': [1,2,3]}
rs = RandomizedSearchCV(DecisionTreeRegressor(),param_grid,n_iter=10)
rs.fit(X, y)
```

12. What would be the contents of variable *result* after executing this piece of code? Why?

```
import numpy as np
result = np.array(['1',2,3])
```

13. What would be the contents of variable *result* after executing this code? Why?

```
d = {'a': 3}      # Note: d is a standard dictionary
result = d['b']
```

14. What would be the contents of variable *result* after executing this code?

```
result = 'abcdef'[2::-1]
```

15. What would be the the main difference between code A and code B below, after executing them?

```
# CODE A
x = [1,2,3]
x+[4,5,6]
```

```
# CODE B
x = [1,2,3]
x.extend([4,5,6])
```

16. What would be the contents of variable *result* after executing this code?

```
result = [2*a for a in [b+5 for b in range(10) if b%2==0]  if a>=2 ]
```

17. Let's suppose that we want to use:
    a. 3-fold crossvalidation for model evaluation
    b. 5-fold crossvalidation for hyper-parameter tuning with grid-search. There are two hyper-parameters: *C* with possible values 0.01, 0.1, and 1; and *gamma* with possible values "linear" and "rbf".

    Then, how many models will be trained? Explain why.

18. Why is caching useful in a Pipeline? Explain an example of a pipeline where it is useful (with words, code is not necessary).

19. Why are priors useful in *Stan*?

20. Give an example of a dataframe in wide format, and its translation to long format