## Streaming Data Analytics

# Preview of the part of the exam about Streaming Data Science

Emanuele Della Valle
Politecnico di Milano

POLITECNICO
MILANO 1863

# Exam content

- **Questions** on all lectures about Streaming Data Science **to test**
  - **The breadth** of your knowledge
  - **The depth** of your knowledge

(6 points)

- **Exercises**
  - Given a python **code snippet** that addresses a streaming binary classification problem or a time series forecasting problem **comment** what the code tries to achieve and **fill in missing parts** explaining **why** you did so
  - Given a particular **problem/task/situation**, explain **how** you would solve it (methods, metrics, evaluation used) and comment **why**

(9 points)

# Questions to test the breadth of your TSA knowledge

- What is the i.i.d. assumption? How do SML and TSA allow the loosening of part of it?
- Which characteristics can a non-stationary time series have? Illustrate it with an example.
- Why is the white noise a predictable time series? What is its forecast? What is its error? Explain it w.r.t. the possible components of a time series.
- Which are the methods to test for stationarity? Explain one of them in detail
- Which are the typical time series components? Illustrate your explanation with an example
- Which are the methods to detrend a time series? Explain one of them in detail
- Which are the methods to identify seasonality in a time series? Explain one of them in detail
- Which are the methods to forecast a time series? Compare two methods of your choice (excluding the basic ones)
- What are the components of a SARIMAX model?

# Questions to test the depth of your TSA knowledge

- What's the difference between an additive and a multiplicative model for time series decomposition? Illustrate your explanation with an example
- Why is exponential smoothing named in this way? Discuss the formula of at least one of the three methods presented in the course.
- What's the difference between simple, double, and triple exponential smoothing? Illustrate your explanation with an example
- What's the difference between the meaning of moving average in time series decomposition and the MA component in ARMA models? Illustrate your explanation with an example
- What's the definition of Autocorrelation? How does it differ from the definition of correlation? Illustrate your explanation with an example
- What's the difference between the AR and the MA part of an ARMA model? What is their relation to Autocorrelation and Partial Autocorrelation? Illustrate your explanation with an example.
- How does the Box-Jenkins Methodology for ARMA models allow us to estimate the orders of the model? Illustrate your explanation with an example
- What are the exogenous variables in TSA? How do they contribute to the forecasting? What's their impact on the confidence interval? Illustrate your explanation with an example

# **Questions** to test the **breadth** of your SML knowledge

- What are the differences between batch-oriented Machine Learning and Streaming Machine Learning?
- What are the benefits and the challenges of Streaming Machine Learning?
- What is a concept drift? Which are the types of concept drift? Why is it so important to detect it? Illustrate the difference using the Bayes Theorem and with an example
- How can you classify a concept drift with respect to the speed of change? Illustrate the difference with several examples
- Which are the typical Streaming Machine Learning algorithms for classification? Illustrate one of them in detail.
- What are the components of an SML Ensemble Classification model?
- How does SML regression compare to time series forecasting w.r.t. types of input features, training, forecasting horizon, and adaptability?

# Questions to test the depth of your SML knowledge

- Which are the concept drift detectors that monitor the error rate? Illustrate how one of them works.
- Which are the data drift detectors? Illustrate how one of them works.
- How does ADWIN detect a concept drift? Illustrate it with an example.
- How does the SML version of KNN work? Illustrate it with an example.
- What is Hoeffding Bound? How and why is it used in SML to build a decision tree?
- Why is the Poisson distribution used in SML ensemble methods? What's the impact of the value of lambda? Illustrate how lambda is used in the SML methods.

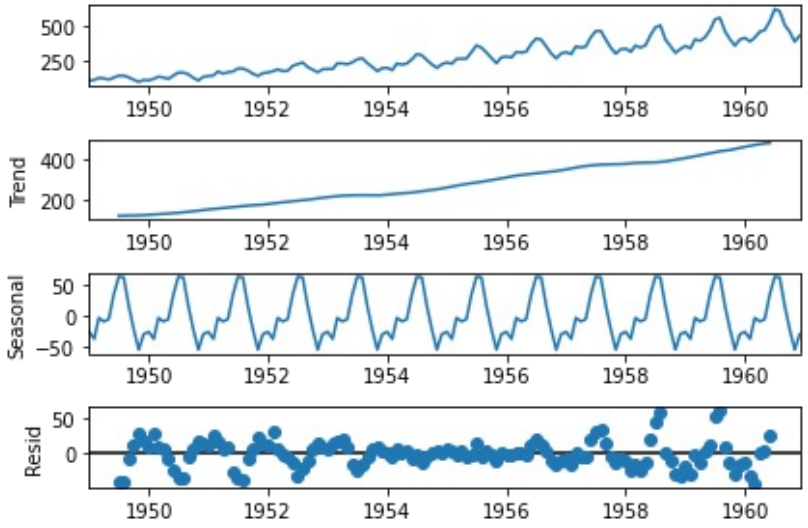
# Questions to test the breadth & depth of your CL knowledge

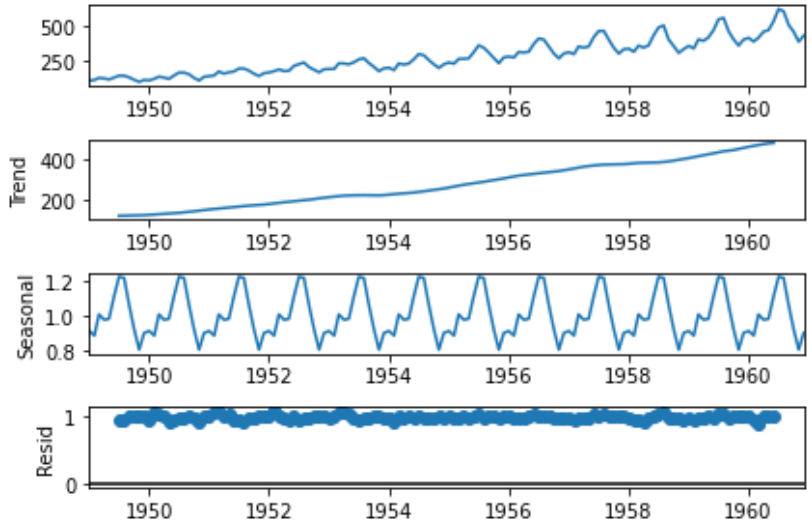
- What's the stability-plasticity dilemma? Illustrate your explanation with an example and discuss the different learning abilities.
- What are the main differences between Streaming Machine Learning and Continual Learning paradigms? Discuss it w.r.t to the objective of the methods and the type of data they process
- What are the three main categories of Continual Learning strategies? Explain one in detail.
- Which are the primary evaluation metrics used in Continual Learning? Illustrate what they measure and why they are all useful to assess the properties of a method.

# Exercises on TSA, 1st Type

Given an additive decomposition plot and a multiplicative decomposition plots:

- **discuss which decomposition is the most suitable;**
- **discuss which of the TSA model(s) seen in lectures you would apply to this time series to make forecasts.**



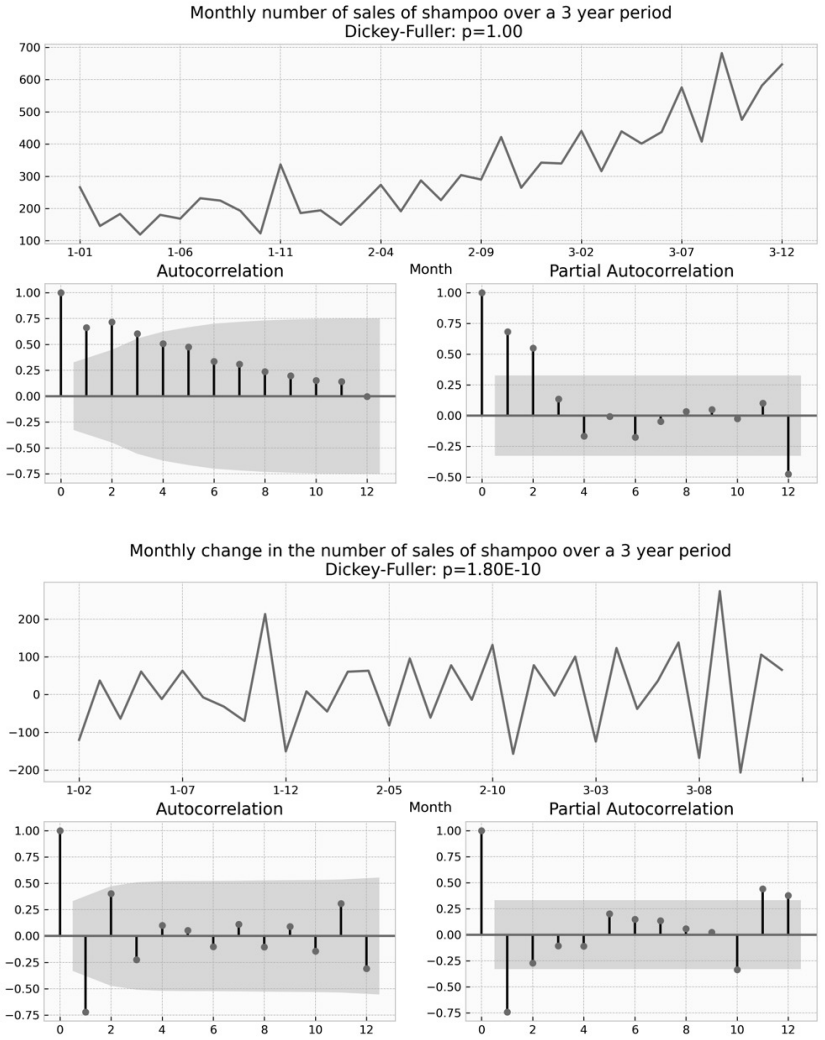



# Exercises on TSA, 2nd Type part 1

Given a a code snippet that computes ACF and PACF plots illustrated:

- **Fill in the blank spaces**

```
1.  data_set = pd.read_csv(  'shampoo.csv',
2.                           infer_datetime_format=True,
3.                           parse_dates=["Month"],
4.                           index_col=["Month"])
    # ACF and PACF plots of monthly number
5.  ……(data_set['Sales'], lags= ……, ax=acf_ax)
6.  ……(data_set['Sales'], lags= ……, ax=pacf_ax)
    # ACF and PACF plots of monthly changes (month_i – month_i-1)
7.  y = ……
8.  ……(y, lags= ……, ax=acf_ax)
9.  ……(y, lags= ……, ax=pacf_ax)
10. plot.show()
```



# Exercises on TSA, 2nd Type part 2

- **Comment each line of the code-snippet.**

| | |
|---|---|
| 1. | |
| 2. | |
| 3. | |
| 4. | |
| 5. | |
| 6. | |
| 7. | |
| 8. | |
| 9. | |
| 10. | |

- **Discuss a possible order of an ARIMA model using the output of the plots.**

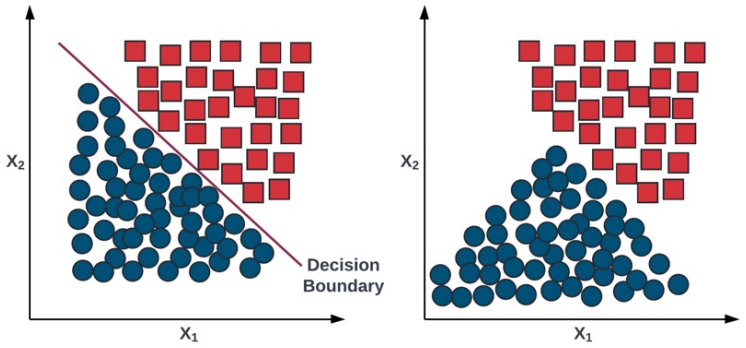
# Exercises on SML

Given a data distribution plot, a code-snippet representing the algorithm training and testing phases, and the resulting error rate plot over time:

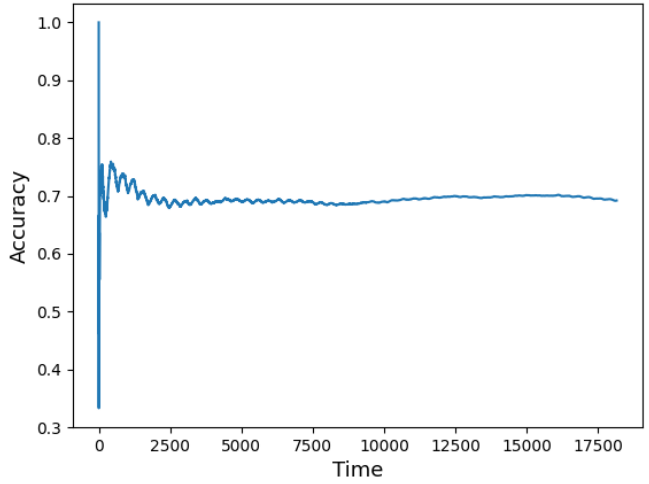
- Comment each line of the code-snippet;
- Discuss the ability of the proposed code to address the data distribution and the error rate illustrated in the two plots;
- Tell if a concept drift occurred, its type, and if it necessary to change the decision boundary;
- How would you modify the code to obtain … (a different situation w.r.t. the illustrated one).

# Solution 1/2

Given a data distribution plot, a code-snippet representing the algorithm training and testing phases, and the resulting error rate plot over time:




```
1. data  = pd.read_csv("stream.csv")
2. features = data.columns[:-1]
3. label = data.columns[-1]
4. stream = iter_pandas(X=data[features], y=data[label])
5. model = GaussianNB()
6. metric = Accuracy()
7. progressive_val_score(dataset=stream,
                         model=model,
                         metric=metrics,
                         print_every=1000)
```



# Solution 2/2

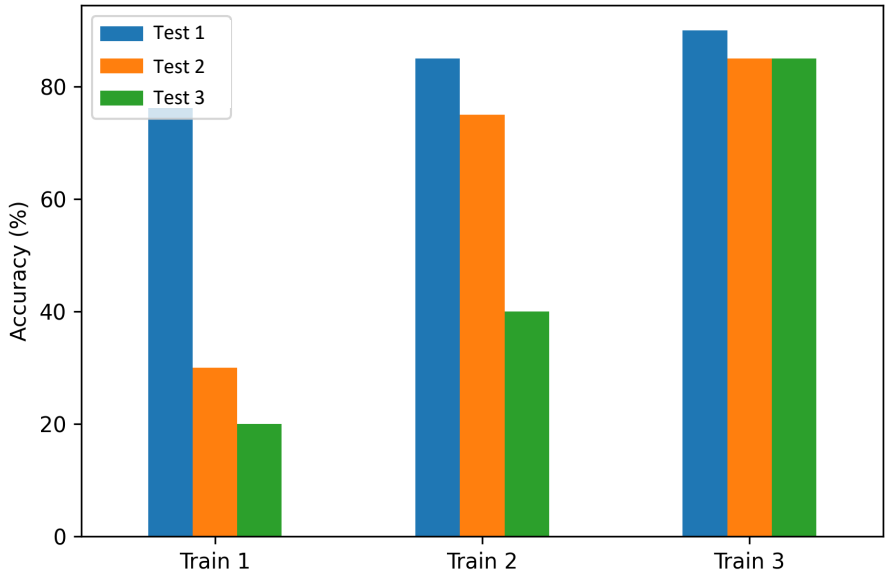
- **Comment each line of the code-snippet**

| | |
|---|---|
| 1. | |
| 2. | |
| 3. | |
| 4. | |
| 5. | |
| 6. | |
| 7. | |

- **Discuss the ability of the proposed code to address the data distribution and the error rate illustrated in the two plots**
- **Tell if a concept drift occurred, its type, and if it necessary to change the decision boundary;**
- **How would you modify the code to obtain a model able to adapt to a concept drift?**

# Exercises on CL

Given the following plot showing the performance of a Continual Learning strategy, describe the learning abilities of the model. Each experience contains 2 classes (consider ½ as the accuracy of the random model).




For the solution see slides 11-12 of lectures/13_CL/13_01_Continual_Learning_Introduction.pdf and codes/CL/01_intro_to_CL.ipynb

## Streaming Data Analytics

# Preview of the part of the exam about Streaming Data Science

Emanuele Della Valle
Politecnico di Milano

POLITECNICO
MILANO 1863