**Parte 1 Data Visualization**

1) Lancia lo script dataVisualization. In quale riga e perché lo script va in errore?

2) Correggi lo script in modo che carichi i dati del contest correttamente e rilancia lo script. Se tutto è corretto si dovrebbe aprire il browser con i dati. Salva la pagina e condividila via e-mail

3) prova a fare zoom sui dati. Cliccando sulla leggenda puoi anche accendere/spegnere le visualizzazioni. Cosa noti sulla serie temporale di houseId e solar -panel-mq?

4) Clicca sul link in basso a destra export to plotly. Dopo che ti sarà aperta la nuova pagina puoi aggiungere nuove analisi. Calcola la correlazione tra wattComsumption e temperature. Che Valore ha?

5)Prova ad aggiungere una media mobile della variabile target wattComsumption e verifica se c’è qualche stagionalità

5) Lo script visualizza solo i dati relativi ad houseId=1. Modificalo e prova a visualizzare i dati relativi ad houseId diversi

**Parte 2 Linear Regression**

Lo script è diviso in 3 parti:

* Lettura e manipolazione dei dati.
* Regressione lineare
* Analisi dei risultati della regressione

La divisione è segnalata con un commento del tipo:

# Fine parte 1 ----------------------------------

**1 Data manipulation**

1) l’obbiettivo della parte di manipolazione dati è ottenere la seguente struttura dati partendo dai due csv. Quante righe e quante colonne e quante r hanno le varibili X\_ e Y\_ alla fine della parte 1? compila la tabella

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |
|  | Train.csv | **Datetime** | **Temp/Radiation/houseid…** | | | **WattCosumption** |  |
|  | 2016-01 | X\_train  #Colonne:  #righe: | | | Y\_train   #Colonne:  #righe: |  |
|  | 2016-02 |  |
|  | 2016-03 |  |
|  | 2016-04 |  |
|  | 2016-05 |  |
|  | 2016-06 |  |
|  | 2016-07 |  |
|  | 2016-08 |  |
|  | 2016-09 |  |
|  | 2016-10 |  |
|  | 2016-11 | X\_Validation  #Colonne:  #righe: | | | Y\_validation  #Colonne:  #righe: |  |
|  | 2016-12 |  |
|  |  |  |  |  |  |  |  |
|  | Test.csv | **Datetime** | **Temp/Radiation/houseid…** | | |  |  |
|  | 2017-01 | X\_test  #Colonne:  #righe: | | |  |  |
|  | 2017-02 |  |  |
|  | 2017-03 |  |  |
|  | 2017-04 |  |  |
|  | 2017-05 |  |  |
|  | 2017-06 |  |  |
|  | 2017-07 |  |  |
|  | 2017-08 |  |  |
|  | 2017-09 |  |  |
|  | 2017-10 |  |  |
|  | 2017-11 |  |  |
|  | 2017-12 |  |  |
|  |  |  |  |  |  |  |  |

3) Quale potrebbe essere il senso di dividere il train set in validation e train-set?

**2 Linear regression**

La regressione lineare è un modello molto semplice, che richiede poca computazione ed è facile da interpretare.

L’idea è abbastanza semplice. Supponiamo che X ha 3 colonne. Mentre Y ha sempre una sola colonna. La regressione lineare cerca e trova 3+1 parametri Theta che minimizzano l’errore sommato su tutte le righe :

In sklearn la regressione lineare è fatta con tre comandi:

* **LinearRegression(…)** l’inizializzazione, in cui definisco i parametri del modello
* **.fit** in cui il modello impara dai dati
* **.predict** in cui il modello applica quello che ha appreso su nuovi dati

Questo schema si applica a tutti i modelli di supervised learning presenti nella libreria scikit-learn.

Qui una lista di tutti i modelli che potete provare (basta modificare la chiamata di inizializzazione):

http://scikit-learn.org/stable/supervised\_learning.html

1. Esegui lo script e sottometti il risultato su kaggle:

<https://www.kaggle.com/c/energy-consumption/submit>

1. Prova a mettere il parametro di inizializzazione fit\_intecept=True e risottometti il risultato. Lo score migliora o peggiora?
2. Prova a generare nuove colonne moltiplicando e/o dividendo tra loro le colonne esistenti.

Sottometti il risultato e verifica se la precisione delle tue previsioni migliora o peggiora.

1. Prova ad estrarre informazione dalle features che non sono misure (datetime e idHouse).

Come potresti generare nuove features partendo da queste colonne? prova a migliorare il tuo score includendo queste informazioni

1. (Opzionale) Genera un grafico in cui viene comparato il consumo di tutte le abitazioni del dataset