Calcolo Numerico 2025-26
Homework 3- Analisi dati con regressione lineare polinomiale

1. **Regressione lineare semplice polinomiale**

   Dati $(x_i, y_i), i = 1, \dots n$, approssimare i dati mediante un modello dato dalla funzione polinomiale:

   $$f(x; \alpha) = \sum_{k=0}^{d} a_k x^k,$$

   Scrivere un codice python che risolve il problema dei minimi quadrati lineari.

   $$min \|X\alpha - y\|_2^2$$

   secondo le notazioni utilizzate a lezione (X è la matrice di Vandermonde) con la fattorizzazione SVD.

   - Creare un problema test scegliendo un prefissato valore di d , un vettore α, n valori x_i equispaziati in $[0, 1]$, un vettore di numeri casuali e con la funzione: `np.random.normal(loc=0, scale=sigma, size=(1, ))` , con deviazione standard sigma = 0.1, generare i valori y_i come:

     $$y_i = f(x_i, \alpha) + e_i = \sum_{k=0}^{n} a_k x_i^k + e_i,$$

   - Testare il codice precedente sui dati generati al variare modificando il grado del polinomio approssimante, per verificare underfit e overfit (variare il numero n dei punti e il valore d almeno un paio di volte).
   - Scaricare da sito di Kaggle un data set adatto per la regressione lineare semplice (ricercare tramite parole chiave su Kaggle). Verificare che il dataset abbia solo due colonne. Applicare la regressione polinomiale al dataset variando il grado del polinomio approssimante. Plottare i risultati e commentarli.

2. **Regressione lineare multipla**

   Scrivere un codice python per realizzare una regressione lineare multipla utilizzando le funzioni della libreria Pandas e scikit-learn.

- Scaricare da Kaggle un data set adatto alla regressione lineare multipla e testare il codice su di esso, ripetendo la regressione più volte eliminando a turno alcune features per identificare quelle più importanti. Valutare i risultati tramite MSE e coefficiente R^2.