Course: Numerical Analysis for Machine Learning
Prof. E. Miglio - July 5th 2024
Duration of the exam: 2.5 hours.

Exercise 1

Consider the Breast Cancer Wisconsin dataset. It contains features computed from digitized images of fine needle aspirate (FNA) of breast masses. These features are used to classify the tumors into malignant (cancerous) or benign (non-cancerous). You can load the dataset as follows:

```
from sklearn.datasets import load_breast_cancer
data = load_breast_cancer()
X = data.data
X = np.transpose(X)
Y = data.target
```

Here's an overview of the dataset:

- Number of Instances: 569,
- Number of Features: 30 numeric (real-valued features),
- Target Variable: Binary (0 for malignant, 1 for benign).

1. Implement a function that, provided the dataset and the number of components to keep, exploits the covariance matrix and its eigenvalues to compute the principal components of the dataset. (*Hint*: use `np.cov(X)` to compute the covariance matrix).
2. Compute the first two principal components of the dataset by using the function implemented at the previous point.
3. Compare and comment the results of the previous point, in terms of principal components, with the ones obtained by means of the exact SVD of rank $r = 2$.
4. Plot the relations between the first 5 features in the original dataset.
5. Make a scatterplot of the first two principal components of the patients. Compare the result with the ones of the previous point.
6. Compute and plot the explained variance ratio, defined as the variance of a principle component divided by the total variance. (*Hint*: it can be computed by exploiting the eigenvalues of the principle components).
7. What is the percentage of the information lost by selecting 2 components ?

Exercise 2

Load the data contained in file `Concrete_Data.csv` as follows:

```
import numpy as np
data = np.loadtxt("Concrete_Data.csv",delimiter=",")
A = data[:,:-1]
b = data[:,-1]
mean_A = np.mean(A,axis = 0)
A = A - mean_A
std_A = np.std(A, axis = 0)
A = A / std_A
```

The dataset contains the characteristics of $n = 1030$ samples of concrete; each sample i is characterized by a feature vector $\mathbf{x}_i$ (with 8 features) and by the corresponding target value y_i which is the concrete compressive strength (in MPa).

Consider the Least Square method where the objective function is given by

$$J(\mathbf{w}) = \frac{1}{2n}\|X\mathbf{w} - \mathbf{y}\|^2, \tag{1}$$

where X is a $n \times 8$ matrix, $\mathbf{y}$ is a n-dimensional vector and $\mathbf{w}$ is a vector with 8 components.

- Proof that the vector $\mathbf{w}^*$ that minimize $J(\mathbf{w})$ is given by $\mathbf{w}^* = (X^T X)^{-1} X^T \mathbf{y}$.
- Show how the SVD of X can be used to simplify the ccomputation of $\mathbf{w}^*$. Compute $\mathbf{w}^*$ (using the SVD) and the corresponding value $J(\mathbf{w}^*)$.
- Implement the Stochastic Gradient Descent method for minimizing (1).
- Run your function for computing $\mathbf{w}^*$. Perform 10000 iterations with different values of the learning rate η; use as initial guess a vector of all zeros.
- Modify your implementation in order to have a decreasing learning rate given by

$$\eta_k = \frac{\eta_0}{\sigma_{min}} \frac{1}{k+1}, \tag{2}$$

  where $\eta_0 = 0.01$, σ_{min} is the smallest singular value of A and k is the iteration index.
- Plot, on the same graphic, the value of the quantity $J(\mathbf{w}_k) - J(\mathbf{w}^*)$ as a function of k for the case with fixed learning rate and for the case with a decreasing learning rate. What do you observe ?

Exercise 3

Consider a perceptron that accepts complex inputs x_1 and x_2. The weights w_1 and w_2 are also complex numbers, and the threshold is zero. The perceptron fires if the condition $\Re(x_1 w_1 + x_2 w_2) \geq \Im(x_1 w_1 + x_2 w_2)$ is satisfied. The binary input 0 is coded as the complex number $(1, 0)$ and the binary input 1 as the number $(0, 1)$. How many of the logical functions of two binary arguments can be computed with this system? Can XOR be computed?