Course: Numerical Analysis for Machine Learning

Prof. E. Miglio - June 25th 2021

Duration of the exam: 2.5 hours.

Exercise 1

Describe the steps required for performing PCA. How SVD can be used within this process ?

Suppose you perform PCA on a 2D dataset and it yields 2 eigenvalues which are equal. What does it mean concerning the importance of the dimension ? Would pursuing a dimensionality reduction be a good choice ? Why ? Draw a dataset with two eigenvalues of the same size.

Suppose now you have performed a PCA on a 2D dataset and you get the eigenvalues 6 and 2. Draw a distribution of sample points that may give rise to this results. Also draw the eigenvectors.

Consider the following 3 data points in the 2D space: $(-1, 1), (0, 0)$ and $(1, 1)$. What's the first principal component of the given dataset ?

If you project the original data points onto the 1D subspace spanned by the principal component, what are their coordinates in this subspace ? What is the variance of the projected data ?

If you represent the projected data in the original 2D space and consider them as the reconstruction of the original data points, what is the reconstruction error ? Compute the reconstruction of the points.

Exercise 2

Consider the following data

$$X = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{bmatrix}, \quad y = \begin{bmatrix} 11 \\ 10 \\ 8 \end{bmatrix}, \tag{1}$$

where X is the data matrix and y contains the labels.

We want to find the parameter vector $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2]^T$ that minimizes the loss over all instances $\mathbf{x}_i$ (the i-th row of the matrix X):

$$\mathcal{L}(X, \boldsymbol{\beta}, y) = \sum_{i=1}^{3} (\boldsymbol{\beta}^T x_i - y_i)^2 \tag{2}$$

- Explain the differences between the classical gradient method and the stochastic gradient method (SGD).
- Perform two epochs using SGD with a step size $\eta = 0.1$ and report the errors and the total loss after each epoch; use the initial guess $\boldsymbol{\beta} = [1, 1, 1]^T$. (Run through the instances in order instead of performing a random selection.).
- Describe the differences between SGD and ADAGRAD. Do you think ADAGRAD would help in this case ?

Exercise 3

Consider a sigmoid neuron with 1D input x, weight w, bias b and output $y = \sigma(wx + b)$. The target is the 1D variable z. Consider the cost function $J(w, b) = \frac{1}{2}(y - z)^2$.

- Find $\nabla J(w, b)$ and show that $\|\nabla J\| < \frac{1}{4}\sqrt{1 + x^2}(1 + |z|)$;
- write the gradient descent iteration for the sequence (w_n, b_n).