CS-E3210- Machine Learning Basic Principles
Home Assignment 6 - "Feature Learning"

Your solutions to the following problems should be submitted as one single pdf which does not contain any personal information (student ID or name). The only rule for the layout of your submission is that for each problem there has to be exactly one separate page containing the answer to the problem. You are welcome to use the LaTeX-file underlying this pdf, available under `https://version.aalto.fi/gitlab/junga1/MLBP2017Public`, and fill in your solutions there.

Problem 1: The Principal Component

Consider $N = 20$ snapshots, available at `https://version.aalto.fi/gitlab/junga1/MLBP2017Public/tree/master/Clustering/images`, which are named according to the season when they have been taken, i.e., either "winter??.jpeg" or "summer??.jpeg". We represent the ith snapshot, with $i = 1, \ldots, N$, by the feature vector $\mathbf{x}^{(i)} \in \mathbb{R}^d$ with entries representing the greyscale values of the image pixels belonging to the lower left square of size 40×40 pixels (this results in a feature length $d = 40^2 = 1600$).

In order to speed up subsequent computations, we transform the original feature vector $\mathbf{x}^{(i)}$ into one single number $z^{(i)} = \mathbf{w}^T \mathbf{x}^{(i)}$ using a normalized vector $\mathbf{w} \in \mathcal{S}^d$, with the unit sphere $\mathcal{S}^d = \{\mathbf{u} \in \mathbb{R}^d : \|\mathbf{u}\|_2^2 = 1\}$ (which is the set of all unit-norm vectors). The vector $\mathbf{w}$ should be chosen such that we can accurately reconstruct the original feature vector using $\mathbf{v}z^{(i)}$ with some normalized vector $\mathbf{v} \in \mathcal{S}^d$ (which might be different from $\mathbf{w}$). Let us measure the reconstruction error, when reconstructing $\mathbf{x}^{(i)}$ from $z^{(i)}$, as

$$\mathcal{E}(\mathbf{v}, \mathbf{w}|\mathbb{X}) := (1/N) \sum_{i=1}^{N} \|\mathbf{x}^{(i)} - \mathbf{v}z^{(i)}\|_2^2 = (1/N) \sum_{i=1}^{N} \|\mathbf{x}^{(i)} - \mathbf{v}\mathbf{w}^T \mathbf{x}^{(i)}\|_2^2. \tag{1}$$

We are interested in finding the vectors $\hat{\mathbf{v}}, \hat{\mathbf{w}} \in \mathcal{S}^d$ which minimize the reconstruction error, i.e.,

$$\mathcal{E}(\hat{\mathbf{v}}, \hat{\mathbf{w}}|\mathbb{X}) = \min_{\mathbf{v}, \mathbf{w} \in \mathcal{S}^d} \mathcal{E}(\mathbf{v}, \mathbf{w}|\mathbb{X}). \tag{2}$$

What is the relation of the vectors $\hat{\mathbf{v}}, \hat{\mathbf{w}} \in \mathcal{S}^d$, which satisfy (2), to the eigenvectors of the matrix $\mathbf{\Sigma} = (1/N)\mathbf{X}^T\mathbf{X}$ with $\mathbf{X} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)})^T \in \mathbb{R}^{N \times d}$? Using this relation, compute the optimal vectors $\hat{\mathbf{v}}, \hat{\mathbf{w}} \in \mathcal{S}^d$ and the associated minimum reconstruction error $\mathcal{E}(\hat{\mathbf{v}}, \hat{\mathbf{w}}|\mathbb{X})$ for the given dataset $\mathbb{X} = \{\mathbf{x}^{(i)}\}_{i=1}^N$. Illustrate the vectors $\hat{\mathbf{v}}, \hat{\mathbf{w}} \in \mathcal{S}^d$ using grayscale plots
(cf. `https://se.mathworks.com/help/images/ref/mat2gray.html`).

Answer. Let us introduce the shorthand $f(\mathbf{v}, \mathbf{w}) := \mathcal{E}(\hat{\mathbf{v}}, \hat{\mathbf{w}}|\mathbb{X})$. First, we note the identity

$$f(\mathbf{v}, \mathbf{w}) = (1/N) \sum_{i=1}^{N} \|\mathbf{x}^{(i)}\|_2^2 - 2\mathbf{v}^T \mathbf{\Sigma} \mathbf{w} + \mathbf{w}^T \mathbf{\Sigma} \mathbf{w} \tag{3}$$
$$\stackrel{(c)}{=} (1/N) \sum_{i=1}^{N} \|\mathbf{x}^{(i)}\|_2^2 + (\mathbf{w} - \mathbf{v})^T \mathbf{\Sigma} (\mathbf{w} - \mathbf{v}) - \mathbf{v}^T \mathbf{\Sigma} \mathbf{v}.$$

The last step (c) makes use of the simple but helpful algebraic identity $a^2 + 2ba = (a + b)^2 - b^2$, which also known as "completing the squares". For a fixed choice $\mathbf{v}$ let us denote a vector which minimizes the objective function $g^{(\mathbf{v})}(\mathbf{w}) = f(\mathbf{v}, \mathbf{w})$ over $\mathbf{w} \in \mathcal{S}^d$ as $\mathbf{w}^{(\mathbf{v})}$. It can be verified easily from (3) that $\mathbf{w}^{(\mathbf{v})} = \mathbf{v}$. Note that, rather trivially, if $\mathbf{v} \in \mathcal{S}^d$ also $\mathbf{w}(\mathbf{v}) \in \mathcal{S}^d$. Thus,

$$\min_{\mathbf{v},\mathbf{w} \in \mathcal{S}^d} f(\mathbf{v}, \mathbf{w}) = \min_{\mathbf{v},\mathbf{w} \in \mathcal{S}^d} g^{(\mathbf{v})}(\mathbf{w}) = \min_{\mathbf{v} \in \mathcal{S}^d} g^{(\mathbf{v})}(\mathbf{w}^{(\mathbf{v})}) = \min_{\mathbf{v} \in \mathcal{S}^d} (1/N) \sum_{i=1}^{N} \|\mathbf{x}^{(i)}\|_2^2 - 2\mathbf{v}^T \mathbf{\Sigma} \mathbf{v} + \mathbf{v}^T \mathbf{\Sigma} \mathbf{v}$$
$$= (1/N) \sum_{i=1}^{N} \|\mathbf{x}^{(i)}\|_2^2 + \min_{\mathbf{v} \in \mathcal{S}^d} -\mathbf{v}^T \mathbf{\Sigma} \mathbf{v} = (1/N) \sum_{i=1}^{N} \|\mathbf{x}^{(i)}\|_2^2 - \max_{\mathbf{v} \in \mathcal{S}^d} \mathbf{v}^T \mathbf{\Sigma} \mathbf{v}. \tag{4}$$

The matrix $\mathbf{\Sigma}$ is positive semidefinite (psd) and therefore $\max_{\mathbf{v} \in \mathcal{S}^d} \mathbf{v}^T \mathbf{\Sigma} \mathbf{v}$ is solved by choosing $\mathbf{v}$ to be an (not necessarily unique) eigenvector of $\mathbf{\Sigma}$ corresponding to the largest eigenvalue of $\mathbf{\Sigma}$. For background on psd matrices see

p. 647 in `https://web.stanford.edu/~boyd/cvxbook/bv_cvxbook.pdf`.

Thus, the optimal vectors $\hat{\mathbf{v}} = \hat{\mathbf{w}}$ can be found by an eigenvalue decomposition of $\mathbf{\Sigma}$ and choosing an eigenvector corresponding to the largest eigenvalue of $\mathbf{\Sigma}$.

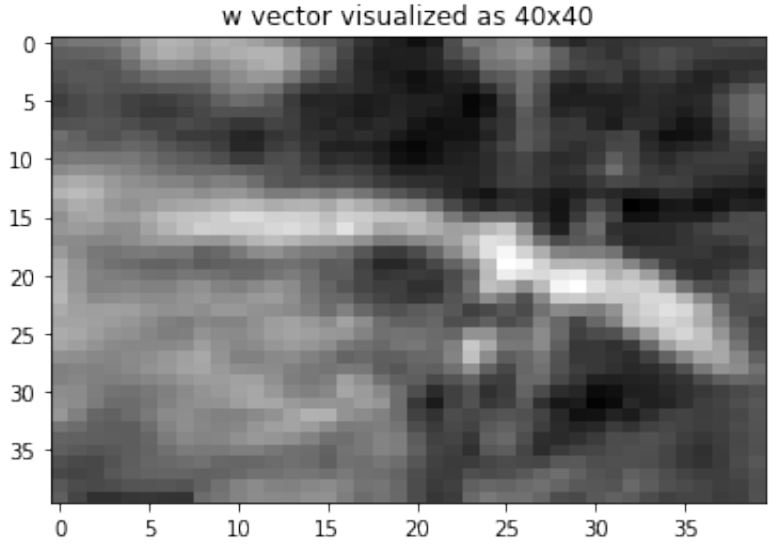

Figure 1: Greyscale plot of the vector $\hat{\mathbf{w}}$ (figure provided by S. Fadnis).

For a related exercise on principal component analysis, see Section 2.12 of the course book `http://www.deeplearningbook.org/contents/linear_algebra.html`.