CMSC25400 Homework 2

University of Chicago, Autumn 2017 due Thursday, October 12 at 11:59pm

1. **(10 points)**

   (a) Let $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_d$ be d mutually orthogonal unit vectors in $\mathbb{R}^d$ (i.e., a basis for $\mathbb{R}^d$), let $\lambda_1, \ldots, \lambda_d$ be real numbers, and

   $$\boldsymbol{A} = \sum_{i=1}^{d} \lambda_i \mathbf{v}_i \mathbf{v}_i^\top . \qquad (1)$$

   Show that $\mathbf{v}_1, \ldots, \mathbf{v}_d$ are eigenvectors of $\boldsymbol{A}$ with corresponding eigenvalues $\lambda_1, \ldots, \lambda_d$.

   (b) Conversely, show that if $\mathbf{v}_1, \ldots, \mathbf{v}_d$ is an orthonormal system of eigenvectors and $\lambda_1, \ldots, \lambda_d$ are the corresponding eigenvalues of a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$, then $\boldsymbol{A}$ is of the form (1).

2. **(20 points)** Let $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ be a symmetric matrix, and for simplicity assume that all its eigenvalues are positive and distinct, $0 < \lambda_1 < \lambda_2 < \ldots < \lambda_d$. Let $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ be the corresponding (normalized) eigenvectors.

   (a) Prove that any two of the eigenvectors $\mathbf{v}_i$ and $\mathbf{v}_j$ (assuming $i \neq j$) are orthogonal (you may wish to compare $\mathbf{v}_i \boldsymbol{A} \mathbf{v}_j$ and $\mathbf{v}_j \boldsymbol{A} \mathbf{v}_i$).

   (b) Explain why this implies that $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_d$ form an orthonormal basis for $\mathbb{R}^d$.

   (c) The Rayleigh quotient of $\boldsymbol{A}$ is

   $$R(\mathbf{w}) = \frac{\mathbf{w}^\top \boldsymbol{A} \mathbf{w}}{\mathbf{w}^\top \mathbf{w}} \qquad \mathbf{w} \in \mathbb{R}^n .$$

   Prove that the maximum of $R(\mathbf{w})$ is λ_d, and that the maximum is attained at $\mathbf{w} = \mathbf{v}_d$.

3. **(20 points)** Recall that the empirical covariance matrix (sample covariance matrix) of a dataset $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ with $\mathbf{x}_i \in \mathbb{R}^d$ (assuming that $\frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i = 0$) is

   $$\widehat{\mathbf{\Sigma}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top .$$

   (a) Since $\widehat{\mathbf{\Sigma}}$ is symmetric, it has an orthonormal basis of eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_d$ with $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_d$, and it can be expressed as

   $$\widehat{\mathbf{\Sigma}} = \sum_{i=1}^{d} \lambda_i \mathbf{v}_i \mathbf{v}_i .$$

   Let $\widehat{\mathbf{\Sigma}}^{(1)}$ be the reduced empirical covariance matrix

   $$\widehat{\mathbf{\Sigma}}^{(1)} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i - (\mathbf{x}_i \cdot \mathbf{v}_d)\mathbf{v}_d)(\mathbf{x}_i - (\mathbf{x}_i \cdot \mathbf{v}_d)\mathbf{v}_d)^\top .$$

   Show that $\widehat{\mathbf{\Sigma}}^{(1)} \mathbf{v}_d = 0$, while $\widehat{\mathbf{\Sigma}}^{(1)} \mathbf{v}_i = \lambda_i \mathbf{v}_i$ for all $i < d$. What are the eigenvalues and eigenvectors of $\widehat{\mathbf{\Sigma}}^{(1)}$ then? Use this to show that the second principal component is $\mathbf{v}_{d-1}$.

   (b) Use induction to show that the k'th principal component of the data is $\mathbf{v}_{d-k+1}$.

4. **(25 points)** The file `3Ddata.txt` is a dataset of 500 points in $\mathbb{R}^3$ sampled from a manifold with some added noise. The last number in each line is just an index in $\{1, 2, 3, 4\}$ related to the position of the point on the manifold to make the visualization prettier (for example, you can plot those points with index 1 in green, index 2 in yellow, 3 in blue and 4 in red).

   Apply PCA and Isomap to map this data to $\mathbb{R}^2$. To construct the graph (mesh) for Isomap you can use $k = 10$ nearest neighbors. Plot the results and comment on the differences. For both methods you need to write your own code (it shouldn't be more than a few lines each) and submit it together with the write-up.

5. **(25 points)** The file `train35.digits` contains 2000 images of 3's and 5's from the famous MNIST database of handwritten digits in text format. The size of each image is 28×28 pixels. Each row of the file is a representation one image, with the 28×28 pixels flattened into a vector of size 784. A value of 1 for a pixel represents black, and value of 0 represents white. The corresponding row of `train35.labels` is the class label: $+1$ for the digit 3, or -1 for the digit 5. The file `test35.digits` contains 200 testing images in the same format as `train35.digits`.

   Implement the perceptron algorithm and use it to label each test image in `test35.digits`. Submit the predicted labels in a file named `test35.predictions`. In the lectures, the perceptron was presented as an online algorithm. To use the perceptron as a batch algorithm, train it by simply feeding it the training set M times. The value of M can be expected to be less than 10, and should be set by cross validation. Naturally, in this context, the "mistakes" made during training are not really errors. Nonetheless, it is intructive to see how the frequency of mistakes decreases as the hypothesis improves. Include in your write-up a plot of the cumulative number of "mistakes" as a function of the number of examples seen. You may find that it improves performance to normalize each example to unit norm.