Computational Statistics and Data Analysis (MVComp2)

Exercise 12

Lecturer Tristan Bereau

Semester Wi23/24
Due Feb. 1, 2024, 23:59

1 Second principal component (5 points)

Recall the reconstruction loss for the first principal component

$$J(\boldsymbol{v}_1, \boldsymbol{z}_1) = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - z_{i1}\boldsymbol{v}_1)^\top(\boldsymbol{x}_i - z_{i1}\boldsymbol{v}_1),$$

which can be used to optimally solve for the principal component, $\boldsymbol{z}_1$.

(a) Extend the loss to the *second* principal component, $J(\boldsymbol{v}_2, \boldsymbol{z}_2; \boldsymbol{v}_1, \boldsymbol{z}_1)$, and show that the solution yields $z_{i2} = \boldsymbol{v}_2^\top \boldsymbol{x}_i$.

(b) Show that the value of $\boldsymbol{v}_2$ that minimizes

$$\tilde{J}(\boldsymbol{v}_2) = -\boldsymbol{v}_2^\top \mathbf{C}\boldsymbol{v}_2 + \lambda_2(\boldsymbol{v}_2^\top\boldsymbol{v}_2 - 1) + \lambda_{12}(\boldsymbol{v}_2^\top\boldsymbol{v}_1 - 0)$$

is given by the eigenvector of $\mathbf{C}$ with the second largest eigenvalue.

Hint: recall that $\mathbf{C}\boldsymbol{v}_1 = \lambda_1\boldsymbol{v}_1$ and $\frac{\partial \boldsymbol{x}^\top \mathbf{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = (\mathbf{A} + \mathbf{A}^\top)\boldsymbol{x}$.

2 Unsupervised exploration of nutrient database (5 points)

Download the following dataset from the USDA national nutrient database: nutrient.csv. Each record is for 100 grams. The are many features, most of which have a naming convention of the type `name_unit`, where unit may be `kcal` for an energy, (`g`, `mg`, `mcg`) for a weight in (gram, milligram, microgram), respectively.

(a) Do you find any redundant features? If so, remove them. Either way, justify your answer.

(b) Scale all your features to 0 mean and unit norm. Run principal component analysis (PCA) on your data. Plot the cumulative explained variance. Use it to identify how many of the first principal components (PCs) you need to explain $1 - \frac{1}{\mathrm{e}}$ of the data.

(c) Check that the first few PCs you've identified in (b) are all orthogonal to each other.

(d) Let's interpret the PCs. Analyze the components of your first three PCs to extract from each one their nutritional composition. Provide the top and bottom three nutrients to help you describe each PC.

(e) For the top three PCs, we wish to identify the food groups that best represent high values of each PC. In each case, focus on the top 500 entries with highest PC. What are the five food groups most represented?

Hint: In Python, you may find the following classes and functions useful:

- `sklearn.preprocessing.StandardScaler`
- `sklearn.decomposition.PCA`