Machine Learning Exercise Sheet 11

Dimensionality Reduction & Matrix Factorization

Homework

PCA & SVD

Problem 1: Use the SVD shown below. Suppose a new user Leslie assigns rating 3 to Alien and

	Matrix	Alien	Star Wars	Casablanca	Titanic
Joe	1	1	1	0	0
Jim	3	3	3	0	0
John	4	4	4	0	0
Jack	5	5	5	0	0
Jill	0	0	0	4	4
Jenny	0	0	0	5	5
Jane	0	0	0	2	2

Figure 11.6: Ratings of movies by users

$$\underbrace{\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 0 & 0 & 2 & 2 \end{bmatrix}}_{M} = \underbrace{\begin{bmatrix} .14 & 0 \\ .42 & 0 \\ .56 & 0 \\ .70 & 0 \\ 0 & .60 \\ 0 & .75 \\ 0 & .30 \end{bmatrix}}_{U} \underbrace{\begin{bmatrix} 12.4 & 0 \\ 0 & 9.5 \end{bmatrix}}_{\Sigma} \underbrace{\begin{bmatrix} .58 & .58 & .58 & 0 & 0 \\ 0 & 0 & 0 & .71 & .71 \end{bmatrix}}_{V^{\mathrm{T}}}$$

rating 4 to Titanic, giving us a representation of Leslie in the 'original space' of $[0, 3, 0, 0, 4]$. Find the representation of Leslie in concept space. What does that representation predict about how well Leslie would like the other movies appearing in our example data?

The projection is given by $\boldsymbol{P} = \boldsymbol{M} \cdot \boldsymbol{V}$, thus the representation of Leslie in concept space is given by $[0, 3, 0, 0, 4] \cdot \boldsymbol{V} = [1.74, 2.84]$. It seems that Leslie has a higher preference for "classic" movies (the score is 2.84) such as "Titanic" and "Casablanca" compared to the "sci-fi" movies (the score is 1.74). Thus, since she already saw "Titanic", "Casablanca" would be a reasonable recommendation.

In general, if $\hat{\boldsymbol{U}}, \hat{\boldsymbol{\Sigma}}, \hat{\boldsymbol{V}^T}$ are the full singular values/vectors of $\boldsymbol{M}$ (obtained by performing full SVD on $\boldsymbol{M}$) and $\boldsymbol{U}, \boldsymbol{\Sigma}, \boldsymbol{V}^T$ are the respective truncated versions (i.e. by taking only the top K singular

values/vectors) it holds that the projected data $\boldsymbol{P}$ can be obtained in two alternative and equivalent ways: $\boldsymbol{P} = \boldsymbol{U} \cdot \boldsymbol{\Sigma}$ or $\boldsymbol{P} = \boldsymbol{M} \cdot \boldsymbol{V}$. We usually prefer the second way since we only need to compute the top k singular vectors.

Problem 2: Consider the latent space distribution

$$p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{0}, \boldsymbol{I})$$

and a conditional distribution for the observed variable $\boldsymbol{x} \in \mathbb{R}^d$,

$$p(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{W}\boldsymbol{z} + \boldsymbol{\mu}, \boldsymbol{\Phi})$$

where $\boldsymbol{\Phi}$ is an arbitrary symmetric, positive-definite noise covariance variable. Now suppose that we make a nonsingular linear transformation of the data variables $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ where $\boldsymbol{A}$ is a non-singular $d \times d$ matrix. If $\boldsymbol{\mu}_{\text{MLE}}$, $\boldsymbol{W}_{\text{MLE}}$, and $\boldsymbol{\Phi}_{\text{MLE}}$ represent the maximum likelihood solution corresponding to the original untransformed data, show that $\boldsymbol{A}\boldsymbol{\mu}_{\text{MLE}}$, $\boldsymbol{A}\boldsymbol{W}_{\text{MLE}}$, and $\boldsymbol{A}\boldsymbol{\Phi}_{\text{MLE}}\boldsymbol{A}^T$ will represent the corresponding maximum likelihood solution for the transformed data set. Finally, show that the form of the model is preserved if $\boldsymbol{A}$ is orthogonal and $\boldsymbol{\Phi}$ is proportional to the unit matrix so $\boldsymbol{\Phi} = \sigma^2\boldsymbol{I}$ (i.e. probabilistic PCA). The transformed $\boldsymbol{\Phi}$ matrix remains proportional to the unit matrix, and hence probabilistic PCA is covariant under a rotation of the axes of data space, as is the case for conventional PCA.

The model for $\boldsymbol{y}$ is a *noiseless* linear transformation. Since the distribution of $\boldsymbol{x}$ is known we also know the distribution of $\boldsymbol{y}$. Because of the definitions for $\boldsymbol{z}$ and $\boldsymbol{x}|\boldsymbol{z}$ we know that $\boldsymbol{x}$ is a Gaussian with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{W}\boldsymbol{W}^T + \boldsymbol{\Phi}$. Thus, $\boldsymbol{y}$ is also Gaussian with mean $\boldsymbol{A}\boldsymbol{\mu}$ and covariance $\boldsymbol{A}\boldsymbol{W}\boldsymbol{W}^T\boldsymbol{A}^T + \boldsymbol{A}\boldsymbol{\Phi}\boldsymbol{A}^T$. Now, assuming that the maximum likelihood solutions for the conditional model for $\boldsymbol{x}$ are $\boldsymbol{\mu}_x$, $\boldsymbol{W}_x$ and $\boldsymbol{\Phi}_x$, we obtain the MLE solutions for $\boldsymbol{y}$ by simple *pattern matching* as $\boldsymbol{A}\boldsymbol{\mu}_x$, $\boldsymbol{A}\boldsymbol{W}_x$ and $\boldsymbol{A}\boldsymbol{\Phi}_x\boldsymbol{A}^T$.

Now, if $\boldsymbol{A}$ is orthogonal and $\boldsymbol{\Phi}$ a scaled identity matrix, the model characteristics are also preserved since $\boldsymbol{A}\boldsymbol{\Phi}_x\boldsymbol{A}^T = \sigma^2\boldsymbol{I}\boldsymbol{A}\boldsymbol{A}^T = \sigma^2\boldsymbol{I}^2 = \sigma^2\boldsymbol{I}$.

Problem 3: Let the matrix $\boldsymbol{X} \in \mathbb{R}^{N\times D}$ represent N data points of dimension $D = 10$ (samples stored as rows). We applied PCA to $\boldsymbol{X}$. By using the $K = 5$ top principal components, we transformed/projected $\boldsymbol{X}$ into $\tilde{\boldsymbol{X}} \in \mathbb{R}^{N\times K}$. We computed that $\tilde{\boldsymbol{X}}$ preserves 70% of the variance of the original data $\boldsymbol{X}$.

Suppose now we apply PCA on the following matrices:

a) $\boldsymbol{Y}_1 = \boldsymbol{X}\boldsymbol{S}$ where $\boldsymbol{S} = \lambda\boldsymbol{I}$, with $\lambda \in \mathbb{R}$ and $\boldsymbol{I} \in \mathbb{R}^{D\times D}$ is the identity matrix

b) $\boldsymbol{Y}_2 = \boldsymbol{X}\boldsymbol{R}$ where $\boldsymbol{R} \in \mathbb{R}^{D\times D}$ and $\boldsymbol{R}\boldsymbol{R}^T = \boldsymbol{I}$

c) $\boldsymbol{Y}_3 = \boldsymbol{X}\boldsymbol{P}$ where $\boldsymbol{P} = \text{diag}(+5, -5, \ldots, +5, -5)$ is a $D \times D$ diagonal matrix

d) $\boldsymbol{Y}_4 = \boldsymbol{X}\boldsymbol{Q}$ where $\boldsymbol{Q} = \text{diag}(1, 2, 3, \ldots, D-1, D)$ is a $D \times D$ diagonal matrix

e) $\boldsymbol{Y}_5 = \boldsymbol{X} + \boldsymbol{1}_N\boldsymbol{\mu}^T$ where $\boldsymbol{\mu} \in \mathbb{R}^D$ and $\boldsymbol{1}_N$ is an N-dimensional column vector of all ones

f) $\boldsymbol{Y}_6 = \boldsymbol{X}\boldsymbol{A}$ where $\boldsymbol{A} \in \mathbb{R}^{D\times D}$ and $\text{rank}(\boldsymbol{A}) = 5$

Upload a single PDF file with your homework solution to Moodle by 19.01.2020, 11:59pm CET. We recommend to typeset your solution (using LaTeX or Word), but handwritten solutions are also accepted. If your handwritten solution is illegible, it won't be graded and you waive your right to dispute that.

and obtain the projected data $\tilde{\boldsymbol{Y}}_1, \dots \tilde{\boldsymbol{Y}}_6 \in \mathbb{R}^{N \times K}$ using the principal components corresponding to the top $K = 5$ largest eigenvalues of the respective $\boldsymbol{Y}_i$.

What fraction of variance of each $\boldsymbol{Y}_i$ will be preserved by each respective $\tilde{\boldsymbol{Y}}_i$? *Justify your answer.*

The answer "cannot tell without additional information" is also valid if you provide a justification.

a) 70%. All eigenvalues are scaled by the same amount λ^2, so the fraction doesn't change.

b) 70%. $\boldsymbol{R}$ is a rotation/reflection/permutation matrix. The direction of the eigenvectors of the covariance matrix is changed, but the eigenvalues stay the same.

c) 70%. This is just combination of (a) and (b). All data points are scaled by 5 (i.e. eigenvalues of $\boldsymbol{X}^T\boldsymbol{X}$ are all scaled by 25), and some dimensions are reflected around origin, but the fraction of variance explained by the first K components stays the same.

d) We cannot tell without additional information. since each column (i.e. each dimension) is scaled by a different amount.

e) 70%. All data points are shifted by $\boldsymbol{\mu}$. But since we center the data as the first step of PCA, shifting has no effect.

f) 100%. Since $\text{rank}(\boldsymbol{A}) = 5$, $\text{rank}(\boldsymbol{Y}_6) \leq 5$ as well. This means that the data lies in a ≤ 5 dimensional subspace, and the first 5 principal components capture all the variance.

Problem 4: You are given $N = 4$ data points: $\{\boldsymbol{x}_i\}_{i=1}^4, \boldsymbol{x}_i \in \mathbb{R}^3$, represented with the matrix $\boldsymbol{X} \in \mathbb{R}^{4\times 3}$.

$$\boldsymbol{X} = \begin{bmatrix} 4 & 3 & 2 \\ 2 & 1 & -2 \\ 4 & -1 & 2 \\ -2 & 1 & 2 \end{bmatrix}$$

Hint: In this task the results of all (final and intermediate) computations happen to be integers.

a) Perform principal component analysis (PCA) of the data $\boldsymbol{X}$, i.e. find the principal components and their associated variances in the transformed coordinate system. Show your work.

First we center the data. The mean is $\bar{\boldsymbol{x}} = [2, 1, 1]$, thus we have

$$\boldsymbol{X}_c = \boldsymbol{X} - \bar{\boldsymbol{x}} = \begin{bmatrix} 2 & 2 & 1 \\ 0 & 0 & -3 \\ 2 & -2 & 1 \\ -4 & 0 & 1 \end{bmatrix}$$

Then we compute the covariance matrix.

$$\boldsymbol{\Sigma}_{X_c} = \frac{1}{N}\boldsymbol{X}_c^T\boldsymbol{X}_c = \begin{bmatrix} 6 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

Since $\boldsymbol{\Sigma}_{X_c}$ it is already in a diagonal form we can conclude that $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}_{X_c}$ and $\boldsymbol{\Gamma} = \boldsymbol{I_3}$, and that it holds $\boldsymbol{\Sigma}_{X_c} = \boldsymbol{\Gamma\Lambda\Gamma}^T$. The principal components are the canonical basis vectors.

Upload a single PDF file with your homework solution to Moodle by 19.01.2020, 11:59pm CET. We recommend to typeset your solution (using LaTeX or Word), but handwritten solutions are also accepted. If your handwritten solution is illegible, it won't be graded and you waive your right to dispute that.

b) Project the data to two dimensions, i.e. write down the transformed data matrix $\boldsymbol{Y} \in \mathbb{R}^{4\times 2}$ using the top-2 principal components you computed in (a). What fraction of variance of $\boldsymbol{X}$ is preserved by $\boldsymbol{Y}$?

The projection matrix is:

$$\boldsymbol{\Gamma}_{trunc} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix}$$

since we pick the first and the third principal vector corresponding to the two largest eigenvalues. Thus, we have

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\Gamma}_{trunc} = \begin{bmatrix} 2 & 1 \\ 0 & -3 \\ 2 & 1 \\ -4 & 1 \end{bmatrix}$$

We preserve $\frac{6+3}{6+2+3} = \frac{9}{11}$ of the variance.

c) Let $\boldsymbol{x}_5 \in \mathbb{R}^3$ be a new data point. Specify the vector $\boldsymbol{x}_5$ such that performing PCA on the data including the new data point $\{\boldsymbol{x}_i\}_{i=1}^5$ leads to exactly the same principal components as in (a).

Let $\boldsymbol{x}_5 = \bar{\boldsymbol{x}}$, i.e. the new data point equals the mean before including $\boldsymbol{x}_5$ to the dataset. Therefore, the new mean including $\boldsymbol{x}_5$ is equal to the old mean. We have:

$$\boldsymbol{X}_c = \boldsymbol{X} - \bar{\boldsymbol{x}} = \begin{bmatrix} 2 & 2 & 1 \\ 0 & 0 & -3 \\ 2 & -2 & 1 \\ -4 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

which leads to the same $\boldsymbol{\Sigma}_{X_c}$ as in (a) up to a difference in the multiplicative constant. In (a) we had $\frac{1}{4}\boldsymbol{X}_c^T\boldsymbol{X}_c$ and here we have $\frac{1}{5}\boldsymbol{X}_c^T\boldsymbol{X}_c$. While this difference leads to different eigenvalues, the eigenvectors and thus the principal components stay the same.

Problem 5: Download the notebook `exercise_11_notebook.ipynb` from Piazza. Fill in the missing code and run the notebook. Convert the evaluated notebook to pdf and add it to the printout of your homework.

The solution notebook is uploaded on Piazza.

In-class Exercises

Probabilistic PCA

Problem 6: For pPCA, we consider the latent space distribution

$$p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{0}, \boldsymbol{I})$$

and a conditional distribution for the observed variable $\boldsymbol{x} \in \mathbb{R}^d$,

$$p(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{W}\boldsymbol{z} + \boldsymbol{\mu}, \sigma^2\boldsymbol{I})$$

a) Verify that the covariance of the marginal distribution $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{C})$ is given by $\boldsymbol{C} = \boldsymbol{W}\boldsymbol{W}^T + \sigma^2\boldsymbol{I}$. What is the interpretation of this result?

b) Verify that the model is unidentifiable, i.e. that the matrix $\boldsymbol{W}$ is only defined up to a rotation $\boldsymbol{R}$. What is the interpretation of this result?

c) Derive an expression for the posterior of the latent variables $p(\boldsymbol{z}|\boldsymbol{x})$!

a) We know from probability theory that $p(\boldsymbol{x}) = \int \mathrm{d}\boldsymbol{z}\ p(\boldsymbol{x}, \boldsymbol{z}) = \int \mathrm{d}\boldsymbol{z}\ p(\boldsymbol{z})p(\boldsymbol{x}|\boldsymbol{z})$. From the expression for marginals of linear Gaussian systems (Murphy equation 4.126 or Bishop equation 2.115) we derive

$$p(\boldsymbol{x}) = \int \mathrm{d}\boldsymbol{z}\, \mathcal{N}(\boldsymbol{z}|\boldsymbol{0}, \boldsymbol{I})\mathcal{N}(\boldsymbol{x}|\boldsymbol{W}\boldsymbol{z} + \boldsymbol{\mu}, \sigma^2\boldsymbol{I}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{W}\boldsymbol{W}^T + \sigma^2\boldsymbol{I}).$$

A direct way to verify this result is given by observing that the distribution of $\boldsymbol{x} = \boldsymbol{W}\boldsymbol{z} + \boldsymbol{\mu} + \boldsymbol{\varepsilon}$ needs to be a Gaussian: $\boldsymbol{z}$ is Gaussian ($p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{0}, \boldsymbol{I})$), a linear transformation $\boldsymbol{W}\boldsymbol{z} + \boldsymbol{\mu}$ of a Gaussian RV is also Gaussian, $\boldsymbol{\varepsilon}$ is Gaussian ($p(\boldsymbol{\varepsilon}) = \mathcal{N}(\boldsymbol{\varepsilon}|\boldsymbol{0}, \sigma^2\boldsymbol{I})$), and the sum of two independent Gaussian RVs $(\boldsymbol{W}\boldsymbol{z} + \boldsymbol{\mu}) + \boldsymbol{\varepsilon}$ is also Gaussian.

Computing the mean and variance of the resulting Gaussian yields the same result as before:

$$\mathbb{E}[\boldsymbol{x}] = \mathbb{E}[\boldsymbol{W}\boldsymbol{z} + \boldsymbol{\mu} + \boldsymbol{\varepsilon}] = \boldsymbol{W}\mathbb{E}[\boldsymbol{z}] + \mathbb{E}[\boldsymbol{\mu}] + \mathbb{E}[\boldsymbol{\varepsilon}] = \boldsymbol{\mu}$$

because of the linearity of the Expectation and the definitions given above. For the Covariance we obtain

$$\begin{aligned}\mathrm{Cov}[\boldsymbol{x}] &= \mathbb{E}[(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])^T] = \mathbb{E}[(\boldsymbol{W}\boldsymbol{z} + \boldsymbol{\varepsilon})(\boldsymbol{W}\boldsymbol{z} + \boldsymbol{\varepsilon})^T] \\ &= \mathbb{E}[\boldsymbol{W}\boldsymbol{z}\boldsymbol{z}^T\boldsymbol{W}^T + \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T] = \boldsymbol{W}\mathbb{E}[\boldsymbol{z}\boldsymbol{z}^T]\boldsymbol{W}^T + \mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T] \\ &= \boldsymbol{W}\boldsymbol{I}\boldsymbol{W}^T + \sigma^2\boldsymbol{I} = \boldsymbol{W}\boldsymbol{W}^T + \sigma^2\boldsymbol{I}\end{aligned}$$

Interpretation: If the matrix $\boldsymbol{W}$ is a $D \times M$ matrix this means that we get a low-rank approximation of a full Covariance $\boldsymbol{\Sigma}$. The main idea of pPCA is to "force the explanation" of the connection between the components x_i into the latent $\boldsymbol{z}$ and the matrix $\boldsymbol{W}$ by using a diagonal covariance in the Gaussian $p(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{W}\boldsymbol{z} + \boldsymbol{\mu}, \sigma^2\boldsymbol{I})$, instead of learning a general covariance as e.g. in $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to explain the correlations between the x_i.

b) Setting $\widetilde{\boldsymbol{W}} = \boldsymbol{W}\boldsymbol{R}$ with $\boldsymbol{R}$ being an (orthonormal) rotation, we see that $\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{W}}^T = \boldsymbol{W}\boldsymbol{R}(\boldsymbol{W}\boldsymbol{R})^T = \boldsymbol{W}\boldsymbol{R}\boldsymbol{R}^T\boldsymbol{W}^T = \boldsymbol{W}\boldsymbol{I}\boldsymbol{W}^T = \boldsymbol{W}\boldsymbol{W}^T$. By that, the marginal distribution is unaltered:

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \widetilde{\boldsymbol{W}}, \sigma) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{W}}^T + \sigma^2\boldsymbol{I}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{W}\boldsymbol{W}^T + \sigma^2\boldsymbol{I}) = p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{W}, \sigma).$$

Interpretation: $\boldsymbol{R}$ is a rotation in the latent space (the space of $\boldsymbol{z}$). Since the "generative story" of pPCA is "sample $\boldsymbol{z}$ from $\mathcal{N}(\boldsymbol{z}|\boldsymbol{0}, \boldsymbol{I})$ and then sample $\boldsymbol{x}$ from $\mathcal{N}(\boldsymbol{x}|\boldsymbol{W}\boldsymbol{z} + \boldsymbol{\mu}, \sigma^2\boldsymbol{I})$", we see that rotations in $\boldsymbol{z}$ do not alter the model, because rotating $\boldsymbol{z}$ before applying $\boldsymbol{W}$ corresponds to having sampled a rotated $\boldsymbol{z}$ from $\mathcal{N}(\boldsymbol{z}|\boldsymbol{0}, \boldsymbol{I})$, which is equally likely because the prior Gaussian is isotropic (has variance $\mathrm{Cov}[z] \propto \boldsymbol{I}$, here even $\mathrm{Cov}[z] = \boldsymbol{I}$)).

c) Using Murphy equation 4.125 or Bishop equation 2.116 for the conditional distribution, we can directly read off $p(\boldsymbol{z}|\boldsymbol{x})$ as $p(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_{z|x}, \boldsymbol{\Sigma}_{z|x})$ where

$$\boldsymbol{\Sigma}_{z|x}^{-1} = \sigma^{-2}(\sigma^2\boldsymbol{I} + \boldsymbol{W}^T\boldsymbol{W})$$
$$\boldsymbol{\mu}_{z|x} = \boldsymbol{\Sigma}_{z|x}[\sigma^{-2}\boldsymbol{W}^T(\boldsymbol{x} - \boldsymbol{\mu})]$$

Using $\boldsymbol{M} = \sigma^2\boldsymbol{I} + \boldsymbol{W}^T\boldsymbol{W}$ we can simplify these equations to obtain

$$\boldsymbol{\Sigma}_{z|x} = \sigma^2\boldsymbol{M}^{-1}$$
$$\boldsymbol{\mu}_{z|x} = \boldsymbol{M}^{-1}\boldsymbol{W}^T(\boldsymbol{x} - \boldsymbol{\mu})$$