Machine Learning Exercise Sheet 10

Dimensionality Reduction & Matrix Factorization, Part 1

In-class Exercises

Problem 1: In this exercise, we use proof by induction to show that the linear projection onto an M-dimensional subspace that maximizes the variance of the projected data is defined by the M eigenvectors of the data covariance matrix $\boldsymbol{S}$, given by

$$\boldsymbol{S} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n - \bar{\boldsymbol{x}})(\boldsymbol{x}_n - \bar{\boldsymbol{x}})^T \qquad \bar{\boldsymbol{x}} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n$$

corresponding to the M largest eigenvalues. In Section 12.1 in Bishop this result was proven for the case of $M = 1$. Now suppose the result holds for some general value of M and show that it consequently holds for dimensionality $M + 1$.

Suppose that the result holds for projection spaces of dimensionality M. The $M + 1$ dimensional principal subspace will be defined by the M principal eigenvectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_M$ together with an additional direction vector $\boldsymbol{u}$ whose value we wish to determine.

Following the argument given in section 12.1.1 for $\boldsymbol{u}_1$, we see that the variance of the data projected onto the subspace spanned by $\{\boldsymbol{u}_1, ..., \boldsymbol{u}_M, \boldsymbol{u}\}$ is given by $\boldsymbol{u}^T\boldsymbol{S}\boldsymbol{u} + \sum_{i=1}^{M}\boldsymbol{u}_i^T\boldsymbol{S}\boldsymbol{u}_i$. Since $\{\boldsymbol{u}_1, ..., \boldsymbol{u}_M\}$ are already chosen and fixed, our goal is to maximize $\boldsymbol{u}^T\boldsymbol{S}\boldsymbol{u}$.

We must constrain $\boldsymbol{u}$ such that it cannot be linearly related to $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_M$ (otherwise it will lie in the M-dimensional projection space instead of defining an $M + 1$ independent direction). For this, we introduce M constraints that require $\boldsymbol{u}$ to be orthogonal to $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_M$. These constraints are enforced using Lagrange multipliers $\eta_1, ..., \eta_M$.

Finally, we introduce a constraint with Lagrange multiplier λ that ensures that $\boldsymbol{u}^T\boldsymbol{u} = 1$.

Putting everything together, we obtain the Lagrangian

$$L(\boldsymbol{u}, \lambda, \eta_1, ..., \eta_M) = \boldsymbol{u}^T\boldsymbol{S}\boldsymbol{u} + \lambda(1 - \boldsymbol{u}^T\boldsymbol{u}) + \sum_{i=1}^{M}\eta_i\boldsymbol{u}_i^T\boldsymbol{u}$$

The solution to a constrained optimization problem is found at the points $(\boldsymbol{u}, \lambda, \eta_1, ..., \eta_M)$, where the gradient of the Lagrangian w.r.t. $\boldsymbol{u}$ is zero and all other constraints are fulfilled, that is

$$\begin{cases} \nabla_{\boldsymbol{u}} L(\boldsymbol{u}, \lambda, \eta_1, ..., \eta_M) = 2\boldsymbol{S}\boldsymbol{u} - 2\lambda\boldsymbol{u} + \sum_{i=1}^{M}\eta_i\boldsymbol{u}_i = \boldsymbol{0} \\ \boldsymbol{u}^T\boldsymbol{u} = 1 \\ \boldsymbol{u}^T\boldsymbol{u}_i = 0 \text{ for } i = 1, ..., M \end{cases}$$

Upload a single PDF file with your homework solution to Moodle by 12.01.2022, 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.

By left-multiplying the first equality with $\boldsymbol{u}_j^T$, we can conclude that it must hold $\eta_j = 0$ for $j = 1, ..., M$. This means we can simplify the problem to

$$\ldots \lambda \boldsymbol{u}$$
$$\ldots 1$$
$$\ldots = 0 \text{ for } i = 1, ..., M$$

This set of equalities ... eigenvalue-eigvector pair $(\lambda_i, \boldsymbol{u}_i)$ of $\boldsymbol{S}$, except the first M eigenvalues/eigenvecto... orthogonal to $\boldsymbol{u}_1, ..., \boldsymbol{u}_M$.

Now we just need to s... solutions $\{\boldsymbol{u}_{M+1}, ..., \boldsymbol{u}_D\}$ that maximizes the variance:

$$\underset{i \in \{M+1, ..., D\}}{\arg\max} \quad \boldsymbol{u}_i^T \boldsymbol{S} \boldsymbol{u}_i$$

This happens to be the $M+1$'s eigenvector $\boldsymbol{u}_{M+1}$ that corresponds to variance λ_{M+1}, since we assumed w.l.o.g. that the eigenvalues are sorted in decreasing order.

To summarize, the variance in the projected space is maximized by choosing $\boldsymbol{u}$ to be the eigenvector having the largest eigenvalue amongst those not previously selected. Thus the result holds also for projection spaces of dimensionality $M+1$, which completes the inductive step. Since we have already shown this result explicitly for $M = 1$ if follows that the result must hold for any $M \leq D$.

Problem 2: Proof that minimizing the error is equivalent to maximizing the variance.

See Bishop chapter 12.1.2

Homework

PCA

Problem 3: Let the ma[illegible]sent N data points of dimension $D = 10$ (samples stored as rows). We applied PCA [illegible] $K = 5$ top principal components, we transformed/projected $\boldsymbol{X}$ into $\tilde{\boldsymbol{X}} \in \mathbb{R}^{N \times K}$. We [illegible]eserves 70% of the variance of the original data $\boldsymbol{X}$.

Suppose now we apply P[illegible] matrices:

a) $\boldsymbol{Y}_1 = \boldsymbol{X}\boldsymbol{S}$ [illegible] $\boldsymbol{S} = \lambda \boldsymbol{I}$, 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 $\boldsymbol{P} = \operatorname{diag}(+5, -5, \dots, +5, -5)$ is a $D \times D$ diagonal matrix

d) $\boldsymbol{Y}_4 = \boldsymbol{X}\boldsymbol{Q}$ where $\boldsymbol{Q} = \operatorname{diag}(1, 2, 3, \dots, D-1, D)$ 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 $\operatorname{rank}(\boldsymbol{A}) = 5$

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 $\operatorname{rank}(\boldsymbol{A}) = 5$, $\operatorname{rank}(\boldsymbol{Y}_6) \leq 5$ 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 t... $\bar{\boldsymbol{x}} = [2, 1, 1]$, thus we have

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

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 $\boldsymbol{\Sigma}_{X_c} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^T$. The principal components are the canonical basis vectors.

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 $\mathbf{\Sigma}_{X_c}$ 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.

SVD

Problem 5: Use the S[illegible]uppose 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

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

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 $\hat{\boldsymbol{U}}, \hat{\mathbf{\Sigma}}, \hat{\boldsymbol{V}}^T$ are the full singular values/vectors of $\boldsymbol{M}$ (obtained by performing full SVD on $\boldsymbol{M}$) and $\boldsymbol{U}, \mathbf{\Sigma}, \boldsymbol{V}^T$ 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: $\boldsymbol{P} = \boldsymbol{U} \cdot \mathbf{\Sigma}$ 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 6: You want to perform linear regression on a data set with features $\boldsymbol{X} \in \mathbb{R}^{N \times D}$ and targets $\boldsymbol{y} \in \mathbb{R}^N$. Assume that you have already computed the SVD of the feature matrix $\boldsymbol{X} = \boldsymbol{U\Sigma V}^T$. Additionally, assume that $\boldsymbol{X}$ has full rank.

Show how we can compu[illegible] regression weights $\boldsymbol{w}^\star$ in $\mathcal{O}(ND)$ operations by using the result of the SVD.

Hint: Matrix operations [illegible]*ymptotic complexity*

- *Matrix multiplicati*[illegible] $\boldsymbol{A} \in \mathbb{R}^{P \times Q}$ *and* $\boldsymbol{B} \in \mathbb{R}^{Q \times R}$ *takes* $\mathcal{O}(PQR)$
- *Matrix multiplicati*[illegible]*ary* $\boldsymbol{A} \in \mathbb{R}^{P \times Q}$ *and a diagonal* $\boldsymbol{D} \in \mathbb{R}^{Q \times Q}$ *takes* $\mathcal{O}(PQ)$
- *Matrix inversion* $\boldsymbol{C}^{-1}$ *for an arbitrary matrix* $\boldsymbol{C} \in \mathbb{R}^{M \times M}$ *takes* $\mathcal{O}(M^3)$
- *Matrix inversion* $\boldsymbol{D}^{-1}$ *for a diagonal matrix* $\boldsymbol{D} \in \mathbb{R}^{M \times M}$ *takes* $\mathcal{O}(M)$

$$
\begin{aligned}
\boldsymbol{w}^* &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \\
&= ((\boldsymbol{U\Sigma V}^T)^T(\boldsymbol{U\Sigma V}^T))^{-1}(\boldsymbol{U\Sigma V}^T)^T\boldsymbol{y} \\
&= (\boldsymbol{V\Sigma U}^T\boldsymbol{U\Sigma V}^T)^{-1}\boldsymbol{V\Sigma U}^T\boldsymbol{y} \\
&= (\boldsymbol{V\Sigma}^2\boldsymbol{V}^T)^{-1}\boldsymbol{V\Sigma U}^T\boldsymbol{y} \\
&= (\boldsymbol{V}^T)^{-1}(\boldsymbol{\Sigma}^2)^{-1}\boldsymbol{V}^{-1}\boldsymbol{V\Sigma U}^T\boldsymbol{y} \\
&= \boldsymbol{V\Sigma}^{-2}\boldsymbol{V}^T\boldsymbol{V\Sigma U}^T\boldsymbol{y} \\
&= \boldsymbol{V\Sigma}^{-2}\boldsymbol{V}^T\boldsymbol{V\Sigma U}^T\boldsymbol{y} \\
&= \boldsymbol{V\Sigma}^{-1}\boldsymbol{U}^T\boldsymbol{y}
\end{aligned}
$$

Multiplication $\boldsymbol{a} = \boldsymbol{U}^T\boldsymbol{y}$ takes $\mathcal{O}(N \cdot D \cdot 1)$

Multiplication $\boldsymbol{b} = \boldsymbol{\Sigma}^{-1}\boldsymbol{a}$ takes $\mathcal{O}(D)$

Multiplication $\boldsymbol{w} = \boldsymbol{V}\boldsymbol{b}$ takes $\mathcal{O}(D^2)$

In total, $\mathcal{O}(ND + D + D^2) = \mathcal{O}(ND)$ if $N > D$.

Coding

Problem 7: Download the notebook `exercise_10_notebook.ipynb` from Moodle. 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 Moodle.