

Business Analytics

Exercise Sheet 8

Josif Grabocka (josif@ismll.de)
Information Systems and Machine Learning Lab (ISMLL)
Universität Hildesheim

24 June 2013
Submission Until 30 June 2013 23:59

Question 1: PCA - Stochastic Gradient Descent (5 points)

Principal Component Analysis (PCA) is a dimensionality reduction technique which aims at projecting a dataset $X \in \mathbb{R}^{N \times M}$ via latent principal components $V \in \mathbb{R}^{K \times M}$ for $K \ll M$. The procedure aims at learning both the latent components and a linear combinations of the components via weights $Z \in \mathbb{R}^{N \times K}$, such that the original data is approximated via the following loss L :

$$\operatorname{argmin}_{Z, V} L = \|X - Z \cdot V\|^2 = \sum_{i=1}^N \sum_{j=1}^M \left(X_{i,j} - \sum_{k=1}^K Z_{i,k} V_{k,j} \right)^2 \quad (1)$$

- (a) Explain the relation between the PCA definition above and the truncated SVD dimensionality reduction? (1 point)
- (b) One method used to compute the PCA of a dataset is called Stochastic Gradient Descent and is shown in Algorithm 1. The values of Z, V are updated to decrease the loss L in the negative direction of the gradient by step η . The procedure is conducted for every cell of X for a total of E many epochs. Each update rule learns the error with respect to one cell (i, j) at a time, hence via the gradient $L_{i,j}$ defined in line 3.

Algorithm 1 Compute PCA through Stochastic Gradient Descent

Require: Original Data $X \in \mathbb{R}^{N \times M}$, Number of latent dimensions K , Learning Rate η , Number of epochs E

Ensure: Low-rank data $Z \in \mathbb{R}^{N \times K}$, Principal components $V \in \mathbb{R}^{K \times M}$

```
1: for  $1, \dots, E$  do
2:   for  $i = 1, \dots, N$   $j = 1, \dots, M, k = 1, \dots, K$  do
3:      $L_{i,j} \leftarrow \left( X_{i,j} - \sum_{k=1}^K Z_{i,k} V_{k,j} \right)^2$ 
4:      $V_{k,j} \leftarrow V_{k,j} - \eta \frac{\partial L_{i,j}}{\partial V_{k,j}}$ 
5:      $Z_{i,k} \leftarrow Z_{i,k} - \eta \frac{\partial L_{i,j}}{\partial Z_{i,k}}$ 
6:   end for
7: end for
8: return  $Z, V$ 
```

Derive the update rule gradients $\frac{\partial L_{i,j}}{\partial V_{k,j}} = ?$, $\frac{\partial L_{i,j}}{\partial Z_{i,k}} = ?$ (3 points)

- (c) Argue how can we find a good value for E ? Does it depend on η ? (1 points)

Question 2: PCA - Gradient Descent (5 points)

- (a) What is the difference between a full gradient descent and stochastic gradient descent? (1 point)
- (b) Another method to compute the PCA of a dataset is through gradient descent, where the latent data Z and the principal components V are updated via computing the full gradient over L , as shown in Algorithm 2.

Algorithm 2 Compute PCA through Gradient Descent

Require: Original Data $X \in \mathbb{R}^{N \times M}$, Number of latent dimensions K , Learning Rate η , Number of epochs E

Ensure: Low-rank data $Z \in \mathbb{R}^{N \times K}$, Principal components $V \in \mathbb{R}^{K \times M}$

```
1: for  $1, \dots, E$  do
2:   for  $i = 1, \dots, N$   $j = 1, \dots, M, k = 1, \dots, K$  do
3:      $V_{k,j} \leftarrow V_{k,j} - \eta \frac{\partial L}{\partial V_{k,j}}$ 
4:      $Z_{i,k} \leftarrow Z_{i,k} - \eta \frac{\partial L}{\partial Z_{i,k}}$ 
5:   end for
6: end for
7: return  $Z, V$ 
```

Derive the update rule gradients $\frac{\partial L}{\partial V_{k,j}} = ?$, $\frac{\partial L}{\partial Z_{i,k}} = ?$ (3 points)

- (c) Argument on the advantages and disadvantages of both learning algorithms. (1 point)

Submission

- Electronically to josif@ismll.de
- Email title must be *BA2013-NAME-Tutorial-NO*, e.g.: *BA2013- JosifGrabocka-Tutorial-1*
- Report file must be a PDF with file name like the email title *BA2013-NAME-Tutorial-NO.pdf*
- Source codes and other materials must be a ZIP with file name like the email title.