Tutorial - Exercise sheet 3

Pattern and Speech Recognition

Introduction

In this exercise you will experiment with dimensionality reduction.

Background

Consider the following equation:

$$XW = T$$

Where X is your original $(m \times n)$ data with m samples and n features. W is a $(n \times n)$ matrix, which we use to transform your data to T $(m \times n)$. Now, if we suppose that W is an orthogonal matrix, then W^{-1} is the same as W^T, so multiplying the equation from the right by W^T we get:

$$X = TW^T$$

This means that we can reconstruct our original matrix X by multiplying the transformed data by the transpose of W. So far the transformed data has the same dimensions as the original data $(m \times n)$, but as we decrease the number of columns of W so will shrink the dimensions of T.

$$\hat{X}_{m \times n} = T_{m \times k} W^T_{k \times n}$$

By reducing the dimensions of T we obviously lose some information, but we will aim for finding such a transformation matrix W which allows the reconstructed data $\hat{X}$ to be as close to the original data X as possible.

$$\|X - \hat{X}\|_2^2 = \|X - T_{m \times k} W^T_{k \times n}\|_2^2$$

So we will try to minimize the distance between the original and the reconstructed matrix (also called as reconstruction error).

1. (1 point) Load the Iris dataset. Keep all instances of Versicolor and Setosa. Use only sepal length and petal length (columns 1 and 3). Perform linear classification, plot your data, the decision boundary and report the error. You can use your code from the first exercise.

2. (1 points) Now we will try to find W, such that:

$$XW = T$$

If we have m samples and n features and we want to reduce the number of features to 1 what are the dimensions of W?

3. (3 points) Now implement a grid search to find the values of W. Choose those parameters that minimize reconstruction error. You can use the interval [-1, 1] and stepsize 0.01 for all your parameters.

4. (1 point) Plot the line that goes through the origin and the point ($W(1)$, $W(2)$) together with your data. This is the line to which you are going to project your data. What can you tell about this line? In which direction does it lie with respect to the data points?

5. (2 points) Project your data: multiply X by W. Plot your transformed data so that each class has a different color. Is your data still linearly separable? What is your new optimal decision boundary and error?

6. (2 points) Now try the same with sepal length and sepal width (1 and 2 columns). What is the error on the projected data? What do you think went wrong? Plot your original data and the projection line and explain why our method resulted in what you see.

Bonus (3 points)

Now load all 4 dimensions of your data. Center your data around zero (subtract the mean of each dimension from the the values of the corresponding dimension). Compute the covariance matrix. Use the following transformation matrix W :

0.323657	0.656822	0.673198	-0.103125
-0.168100	0.748155	-0.631676	0.113979
0.869291	-0.091511	-0.266429	0.406171
0.333649	-0.021792	-0.277136	-0.900777

The values of this matrix have been chosen in such a way that if you use the first 3 column to project to 3D the reconstruction error will be minimized, if use the first two to project to 2D the reconstruction error will be minimized, if you use the first column to project to 1D the reconstruction error will be minimized. Now compute the covariance matrix for your transformed data. Compare the two covariance matrices! Pay attention to how the values on the diagonal (variance) and off the diagonal (covariance) have changed. What has our method achieved?

Submission architecture

You have to generate a **single ZIP file** respecting the following architecture:

```
tutorial1_<matriculation_nb1>_<matriculation_nb2>_<matriculation_nb3>
|
+--- source
|      |
|      +------- file 1
|      +------- file 2
|      +------- ...
+--- rapport.pdf
+--- README.txt
```

where

- **source** contains the source code of your project,
- **rapport.pdf** is the report where you present your solution with **the explanations (!)** and the plots,
- **README** which contains group member informations (name, matriculation numbers and emails) and a **clear** explanation about how to compile and run your source code

The ZIP filename has to be :

```
tutorial1_<matriculation_nb1>_<matriculation_nb2>_<matriculation_nb3>.zip
```

You have to choose between the following languages **python** or **matlab**. Other languages won't be accepted.

Some hints

We advice you to follow the following guidelines in order to avoid problems :

- Avoid building complex systems. The exercises are simple enough.
- Do not include any executables in your submission, as this will cause the e-mail server to reject it.

Grading

Send your assignment to the tutor who is responsible of your group:

- Gerrit Großmann gerritgr@gmail.com
- Sébastien Le Maguer slemaguer@coli.uni-saarland.de
- Kata Naszádi b.naszadi@gmail.com

The email subject should start with `[PSR TUTORIAL 3]`