Machine Learning 1
Exercise 4

Group: BSSBCH

November 13, 2017

Matthias Bigalke, 339547, maku@win.tu-berlin.de
Tolga Buz, 346836, buz_tolga@yahoo.de
Alejandro Hernandez, 395678, alejandrohernandezmunuera@gmail.com
Aitor Palacios Cuesta, 396276, aitor.palacioscuesta@campus.tu-berlin.de
Christof Schubert, 344450, christof.schubert@campus.tu-berlin.de
Daniel Steinhaus, 342563, dany.steinhaus@googlemail.com

1 Lagrange Multipliers

(a)

Find the parameter θ that minimizes $J(\theta)$ subject to the constraint $\theta^T b = 0$.

We have

$$\mathcal{L} = \sum_{k=1}^{n} ||\boldsymbol{\theta} - \boldsymbol{x}_k||^2 + \lambda \boldsymbol{\theta}^T \boldsymbol{b}$$

From this we get

$$0 = \frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{L} \tag{1}$$

$$\Leftrightarrow \quad 0 = \sum_{k=1}^{n} 2(\boldsymbol{\theta} - \boldsymbol{x}_k) + \lambda \boldsymbol{b} \tag{2}$$

$$\Leftrightarrow \quad 0 = \boldsymbol{\theta} - \overline{\boldsymbol{x}} + \frac{\lambda}{2n} \boldsymbol{b} \tag{3}$$

$$\Leftrightarrow \quad \boldsymbol{\theta} = \overline{\boldsymbol{x}} - \frac{\lambda}{2n} \boldsymbol{b} \tag{4}$$

and

$$0 = \frac{\partial}{\partial \boldsymbol{\lambda}} \mathcal{L} \tag{5}$$

$$\Leftrightarrow \quad 0 = \boldsymbol{\theta}^T \boldsymbol{b} \tag{6}$$

$$\overset{(4)}{\Leftrightarrow} \quad 0 = (\overline{\boldsymbol{x}} - \frac{\lambda}{2n} \boldsymbol{b})^T \boldsymbol{b} \tag{7}$$

$$\Leftrightarrow \quad 0 = \overline{\boldsymbol{x}}^T \boldsymbol{b} - \frac{\lambda}{2n} \boldsymbol{b}^T \boldsymbol{b} \tag{8}$$

$$\overset{\boldsymbol{b} \neq \boldsymbol{0}}{\Leftrightarrow} \quad \lambda = 2n \frac{\overline{\boldsymbol{x}}^T \boldsymbol{b}}{\boldsymbol{b}^T \boldsymbol{b}} \tag{9}$$

$$\tag{10}$$

With (4) and (10) we finally get

$$\boldsymbol{\theta} = \bar{\boldsymbol{x}} - \frac{\bar{\boldsymbol{x}}^T \boldsymbol{b}}{\boldsymbol{b}^T \boldsymbol{b}} \boldsymbol{b} = \bar{\boldsymbol{x}} - \frac{\bar{\boldsymbol{x}}^T \boldsymbol{b}}{||\boldsymbol{b}||} \cdot \frac{\boldsymbol{b}}{||\boldsymbol{b}||}$$

Geometrical interpretation

$\frac{\bar{\boldsymbol{x}}^T \boldsymbol{b}}{||\boldsymbol{b}||}$ discribes a projection from $\bar{\boldsymbol{x}}$ on $\boldsymbol{b}$. $\frac{\bar{\boldsymbol{x}}^T \boldsymbol{b}}{||\boldsymbol{b}||} \cdot \frac{\boldsymbol{b}}{||\boldsymbol{b}||}$ gives us the projection point. So this constraint gives us a minimum at the shifted emperical mean in opposite direction of $\boldsymbol{b}$ by the value of the projection.

(b)

Find the parameter θ that minimizes $J(\theta)$ subject to the constraint $||\boldsymbol{\theta} - \boldsymbol{c}||^2 = 1$.

We have

$$\mathcal{L} = \sum_{k=1}^{n} ||\boldsymbol{\theta} - \boldsymbol{x}_k||^2 + \lambda ||\boldsymbol{\theta} - \boldsymbol{c}||^2 - \lambda$$

From this we get

$$0 = \frac{\partial}{\partial \theta} \mathcal{L} \tag{11}$$

$$\Leftrightarrow \quad 0 = \sum_{k=1}^{n} 2\boldsymbol{\theta} - 2\boldsymbol{x}_k + 2\lambda(\boldsymbol{\theta} - \boldsymbol{c}) \tag{12}$$

$$\Leftrightarrow \quad 0 = n\boldsymbol{\theta} - \sum_{k=1}^{n} \boldsymbol{x}_k + \lambda\boldsymbol{\theta} - \lambda\boldsymbol{c} \tag{13}$$

$$\Leftrightarrow \quad 0 = \boldsymbol{\theta} - \bar{\boldsymbol{x}} + \frac{\lambda}{n}\boldsymbol{\theta} - \frac{\lambda}{n}\boldsymbol{c} \tag{14}$$

$$\Leftrightarrow \quad \frac{n+\lambda}{n}\boldsymbol{\theta} = \bar{\boldsymbol{x}} + \frac{\lambda}{n}\boldsymbol{c} \tag{15}$$

$$\Leftrightarrow \quad \boldsymbol{\theta} = \frac{n}{n+\lambda}\bar{\boldsymbol{x}} + \frac{\lambda}{n+\lambda}\boldsymbol{c} \tag{16}$$

and

$$0 = \frac{\partial}{\partial \lambda} \mathcal{L} \tag{17}$$

$$\Leftrightarrow \quad 0 = ||\boldsymbol{\theta} - \boldsymbol{c}||^2 - 1 \tag{18}$$

$$\Leftrightarrow \quad 0 = ||\boldsymbol{\theta} - \boldsymbol{c}||^2 - 1 \qquad \text{eq (16)} \tag{19}$$

$$\Leftrightarrow \quad 0 = ||\frac{n}{n+\lambda}\bar{\boldsymbol{x}} + \frac{\lambda}{n+\lambda}\boldsymbol{c} - \boldsymbol{c}||^2 - 1 \tag{20}$$

$$\Leftrightarrow \quad 0 = ||\frac{n}{n+\lambda}\bar{\boldsymbol{x}} - \frac{n}{n+\lambda}\boldsymbol{c}||^2 - 1 \tag{21}$$

$$\Leftrightarrow \quad 1 = \frac{n^2}{(n+\lambda})^2||\bar{\boldsymbol{x}} - \boldsymbol{c}||^2 \tag{22}$$

$$\Leftrightarrow \quad (n+\lambda)^2 = n^2||\bar{\boldsymbol{x}} - \boldsymbol{c}||^2 \tag{23}$$

$$\Rightarrow \quad \lambda = n \cdot (||\bar{\boldsymbol{x}} - \boldsymbol{c}|| - 1) \tag{24}$$

With Eq (16) and Eq (24) we get

$$\begin{aligned}\boldsymbol{\theta} &= \frac{n}{n + n||\overline{\boldsymbol{x}} - \boldsymbol{c}|| - n}\overline{\boldsymbol{x}} + \frac{n||\overline{\boldsymbol{x}} - \boldsymbol{c}|| - n}{n + n||\overline{\boldsymbol{x}} - \boldsymbol{c}|| - n}\boldsymbol{c} \\ &= \frac{1}{||\overline{\boldsymbol{x}} - \boldsymbol{c}||}\overline{\boldsymbol{x}} + \frac{||\overline{\boldsymbol{x}} - \boldsymbol{c}|| - 1}{||\overline{\boldsymbol{x}} - \boldsymbol{c}||}\boldsymbol{c} \\ &= \frac{1}{||\overline{\boldsymbol{x}} - \boldsymbol{c}||}\overline{\boldsymbol{x}} - \frac{1}{||\overline{\boldsymbol{x}} - \boldsymbol{c}||}\boldsymbol{c} + c \\ &= \frac{1}{||\overline{\boldsymbol{x}} - \boldsymbol{c}||}(\overline{\boldsymbol{x}} - \boldsymbol{c}) + c\end{aligned}$$

Geometrical interpretation

2 Bounds on Eigenvalues

(a)

Show that $\sum_{i=1}^{d} S_{ii} \geq \lambda_1$.

Assumptions

Since S is a scatter matrix we can infer:

(I) S is symmetrical

(II) S is positive semi definite

Proof

We have

$$\begin{aligned} & \lambda_1 \leq \sum_{i=1}^{d} S_{ii} = Tr(S) \overset{(I)}{=} \sum_{i=1}^{d} \lambda_i \\ \Leftrightarrow \quad & 0 \leq \sum_{i=2}^{d} \lambda_i \end{aligned}$$

which is fullfilled because with (II) we get $\forall i \in \{1, ..., d\} : \lambda_i \geq 0$.

(b)

(c)

(d)

3 Iterative PCA

(a)

(b)

4 Principal Component Analysis

4.1 Introduction

In this exercise, you will experiment with two different techniques to compute the principal components of a dataset:

- **Basic PCA**: The standard technique based on singular value decomposition.
- **Iterative PCA**: A technique that progressively optimizes the PCA objective function.

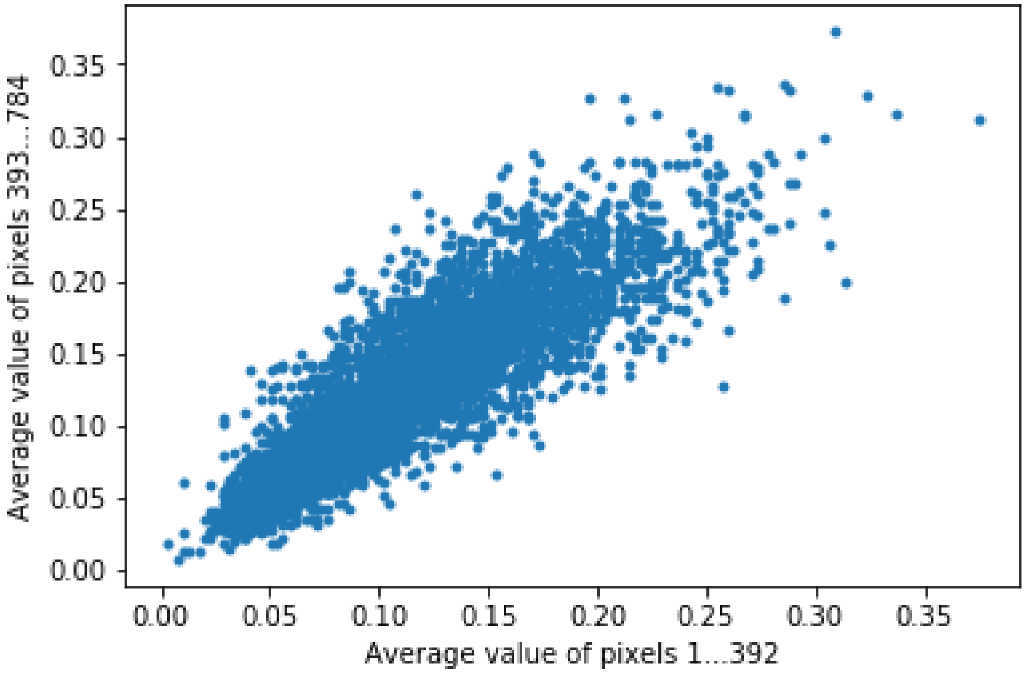
Principal component analysis is applied here to modeling handwritten characters data (characters "O" and "I") using the dataset introduced in the paper "L.J.P. van der Maaten. 2009. A New Benchmark Dataset for Handwritten Character Recognition". The dataset consists of black and white images of 28 × 28 pixels, each representing a handwritten character. For the purpose of the PCA analysis, these images are interpreted as 784-dimensional vectors with values between 0 and 1. Three methods are provided for your convenience and are available in the module `utils` that is included in the zip archive. The methods are the following:

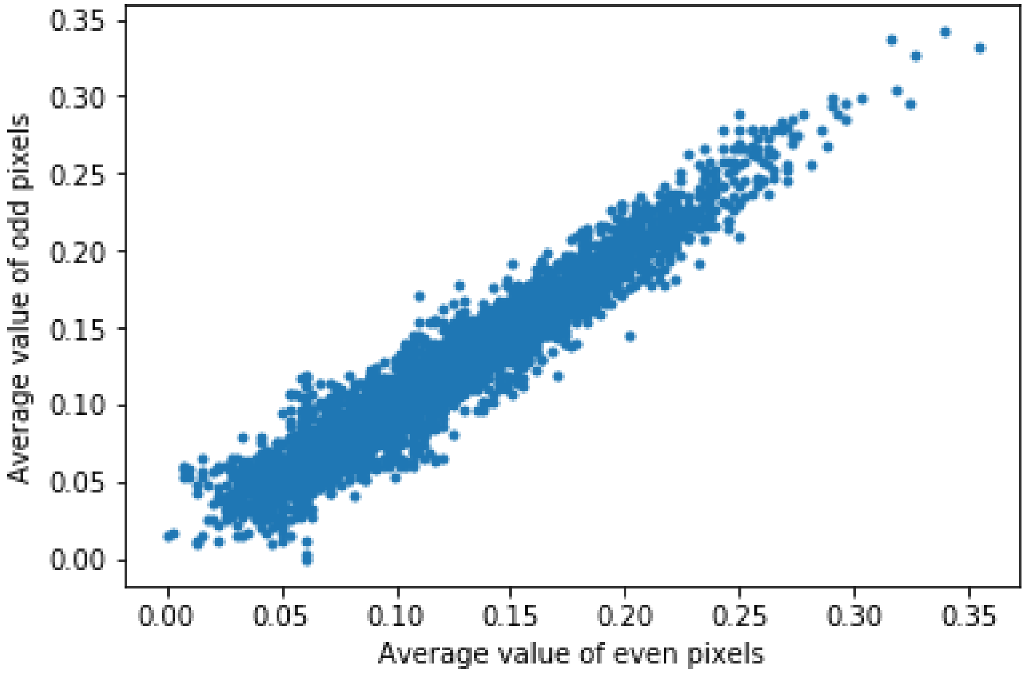
- `utils.load()` load data from the file `characters.csv` and stores them in a data matrix of size 4631 × 784. (The data is a subset of the original dataset available here: http://lvdmaaten.github.io/publications/misc/characters.zip)
- `utils.scatterplot(...)` produces a scatter plot from a two-dimensional data set. Each point in the scatter plot represents one handwritten character. This method provides a convenient way to produce two-dimensional PCA plots.
- `utils.render(...)` takes a matrix of size $n \times 784$ as input, interprets it as n images of size 28 × 28, and renders these images in the IPython notebook.

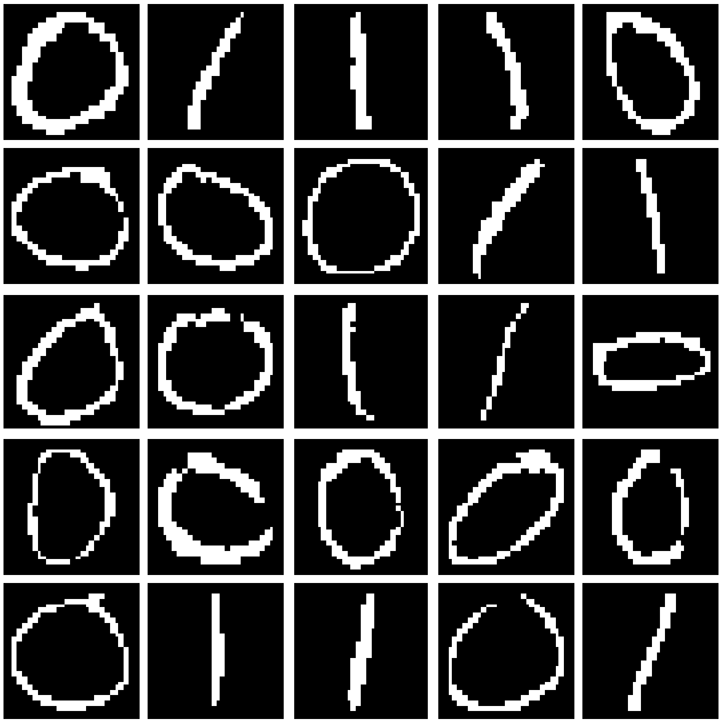
A demo code that makes use of these methods is given below. It performs basic data analysis, for example, plotting simple statistics for each data point in the dataset, or rendering a few examples randomly selected from the dataset.

```
In [1]: import utils,numpy
        %matplotlib inline

        # Load the characters "O" and "I" from the handwritten characters dataset
        X = utils.load()

        print('dataset size: %s'%str(X.shape))

        # Plot some statistics of the data using the scatterplot function
        utils.scatterplot(X[:,:392].mean(axis=1),X[:,392:].mean(axis=1),
                          xlabel='Average value of pixels 1...392',
                          ylabel='Average value of pixels 393...784')
        utils.scatterplot(X[:,::2].mean(axis=1),X[:,1::2].mean(axis=1),
                          xlabel='Average value of even pixels',
                          ylabel='Average value of odd pixels')

        # Render some randomly selected examples
        R=numpy.random.randint(0,len(X),[25])
        utils.render(X[R])
```

```
dataset size: (4631, 784)
```


0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Average value of pixels 393...784
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Average value of pixels 1...392


0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Average value of odd pixels
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Average value of even pixels



The preliminary data analysis above does not reveal particularly interesting structure in the data. For example scatter plots fail to let appear the two types of characters present in the dataset ("O" and "I"). Therefore, we would like to gain more insight on the dataset by performing a more sophisticated analysis based on PCA.

4.2 PCA with Singular Value Decomposition (15 P)

As shown during the lecture, principal components can be found by solving the eigenvalue problem

$$S\boldsymbol{w} = \lambda \boldsymbol{w}.$$

While we could eigendecompose the scatter matrix to find the desired eigenvalues and eigenvectors (for example, by using the function `numpy.linalg.eigh`), we usually prefer to recover principal components directly from singular value decomposition

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top,$$

where the principal components and projection of data onto these components can also be retrieved from the matrices $\boldsymbol{U}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{V}$.
Tasks:

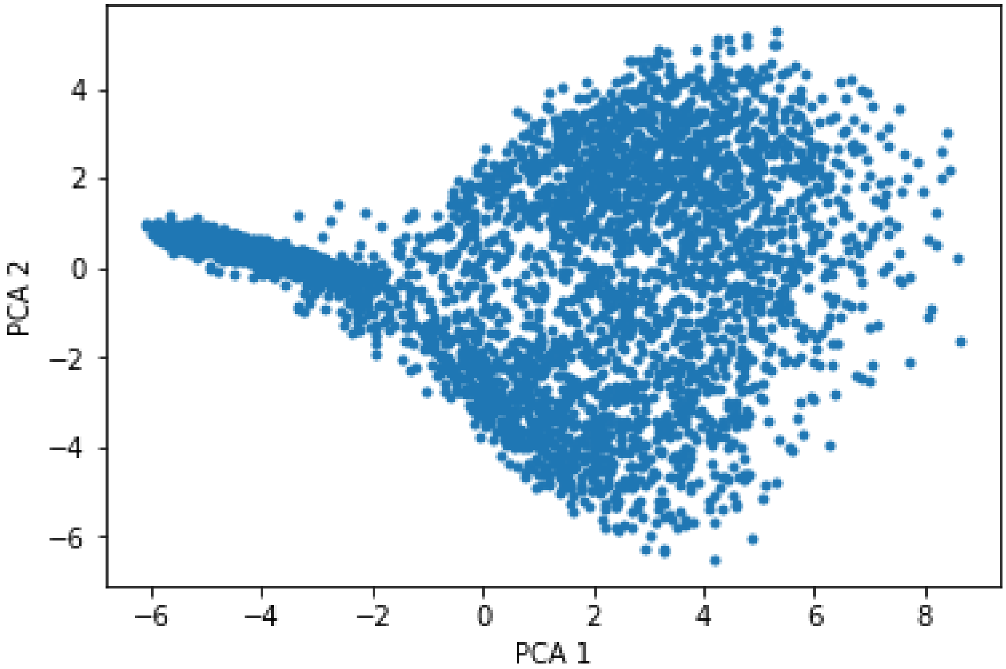
- **Compute the principal components of the data using the function** `numpy.linalg.svd`**.**
- **Measure the computational time required to find the principal components. Use the function** `time.time()` **for that purpose. Do *not* include in your estimate the computation overhead caused by loading the data, plotting and rendering.**
- **Plot the projection of the dataset on the first two principal components using the function** `utils.scatterplot`**.**

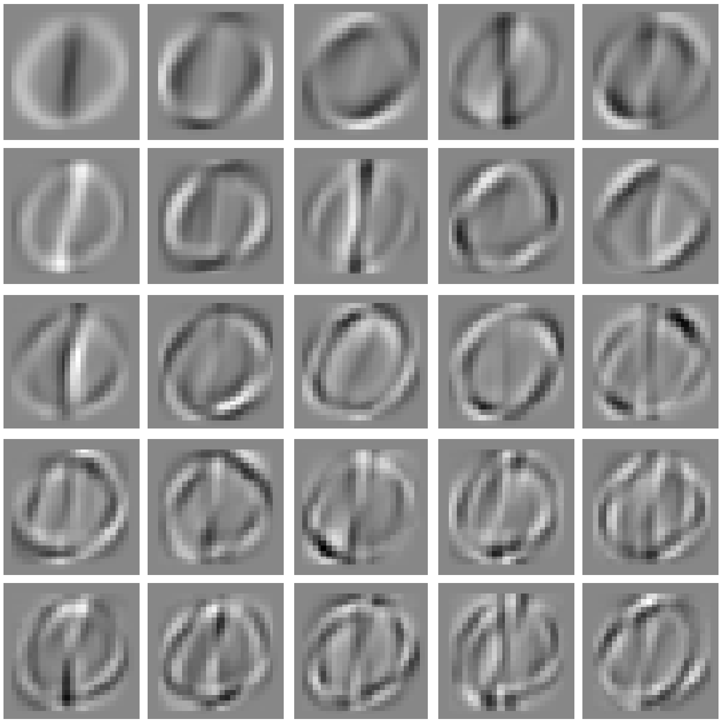
- **Visualize the 25 leading principal components using the function `utils.render`.**

Note that if the algorithm runs for more than 1 minute, you might be doing something wrong.

```
In [2]: import time

        t1 = time.time()
        X_c = X - X.mean(axis=0, dtype=numpy.float64)

        u,s,v = numpy.linalg.svd(X_c)
        W = v[numpy.argsort(s)[-2:],:]
        W = W[::-1,:]
        PCA = W.dot(X_c.T)
        print "Time: " + str(time.time() - t1) + " seconds."

        utils.scatterplot(PCA[0,:],PCA[1,:], xlabel='PCA 1', ylabel='PCA 2')

        eig = v[numpy.argsort(s)[-25:],:]
        eig = eig[::-1,:]
        utils.render(eig)
```

```
Time: 9.75734400749 seconds.
```





4.3 Iterative PCA (15 P)

The objective that PCA optimizes is given by

$$J(\boldsymbol{w}) = \boldsymbol{w}^\top S \boldsymbol{w}$$

subject to

$$\boldsymbol{w}^\top \boldsymbol{w} = 1.$$

The power iteration algorithm maximizes this objective using an iterative procedure. It starts with an initial weight vector $\boldsymbol{w}$, and iteratively applies the update rule

$$\boldsymbol{w} \leftarrow \frac{S\boldsymbol{w}}{\|S\boldsymbol{w}\|}$$

Tasks:

- **Implement the iterative procedure. Use as a stopping criterion the value of $J(\boldsymbol{w})$ between two iterations increasing by less than 0.01.**
- **Print the value of the objective function $J(\boldsymbol{w})$ at each iteration.**
- **Measure the time taken to find the principal component.**
- **Visualize the the eigenvector $\boldsymbol{w}$ obtained after convergence using the function `utils.render`.**

Note that if the algorithm runs for more than 1 minute, you might be doing something wrong.

```
In [3]: t1 = time.time()
        N = X.shape[0]
        X_c = X - X.mean(axis=0, dtype=numpy.float64)
        S = 1.0/N * X_c.T.dot(X_c)

        w = numpy.random.normal(0,1,size=(S.shape[0],1))
        w = 1/numpy.linalg.norm(w) * w

        j_last = -1
        j_current = -1
        i = 0
        while True:
            w = S.dot(w)
            w = 1/numpy.linalg.norm(w) * w
            j_last = j_current
            j_current = w.T.dot(S.dot(w))
            print "iteration " + str(i) + " J(w) = " + str(j_current[0,0])
            if (j_last > 0) and numpy.abs(j_last - j_current) < 0.01:
                print "Stopping creterion satisfied."
                break
            i += 1

        print "Time: " + str(time.time() - t1) + " seconds."
        utils.render(w.T)
```

```
iteration 0 J(w) = 2.9689384223
iteration 1 J(w) = 3.82131949849
iteration 2 J(w) = 4.84240465711
iteration 3 J(w) = 8.66375108562
iteration 4 J(w) = 12.2808867695
iteration 5 J(w) = 12.9987068232
iteration 6 J(w) = 13.0808942813
iteration 7 J(w) = 13.0896578376
Stopping creterion satisfied.
Time: 0.211755037308 seconds.
```