Assignment 3

March 13, 2017

1 Foundations of Data Mining: Assignment 3

Please complete all assignments in this notebook. You should submit this notebook, as well as a PDF version (See File > Download as).

```
In [1]: %matplotlib inline
        from preamble import *
        plt.rcParams['savefig.dpi'] = 100 # This controls the size of your figures
        # Comment out and restart notebook if you only want the last output of each cell.
        InteractiveShell.ast_node_interactivity = "all"
```

1.1 Random Projections with 1-NN (6 points, 3+3)

Implement random projections for dimensionality reduction as follows. Randomly generate a $k \times d$ matrix $\mathbf{R}$ by choosing its coefficients

$$r_{i,j} = \begin{cases} +\frac{1}{\sqrt{d}} & \text{with probability} \quad \frac{1}{2} \\ -\frac{1}{\sqrt{d}} & \text{with probability} \quad \frac{1}{2} \end{cases}$$

Let $f : \mathbb{R}^d \to \mathbb{R}^k$ denote the linear mapping function that multiplies a d-dimensional vector with this matrix $f(p) = \mathbf{R}p$. For the following exercises use the same data set as was used for Assignment 1 (MNIST). Use the following values of $k = 45, 90, 400$ in your experiments.
You should *not* use sklearn.random_projection for this assignment.

1.1.1 Study the effect on pairwise distances

Evaluate how well the Euclidean distance is preserved by plotting a histogram of the values $\phi(p,q) = \frac{\|f(p)-f(q)\|}{\|p-q\|}$ for all pairs of the first 500 images of the MNIST data set. These values should be concentrated around a certain value for fixed k. What is this value expressed in terms of k and d? Explain your answer.

```
In [2]: # This is a temporary read-only OpenML key. Replace with your own key later.
        oml.config.apikey = '11e82c8d91c5abece86f424369c71590'
```

```
In [3]: mnist_data = oml.datasets.get_dataset(554) # Download MNIST data
        # Get the predictors X and the labels y
        X, y = mnist_data.get_data(target=mnist_data.default_target_attribute);
```

```
In [4]: # Randomly sample with probability 1/2
        np.random.randint(0,2)

Out[4]: 1

In [5]: import matplotlib.pyplot as plt
        import numpy as np
```

1.1.2 Study the effect on classification

Compare the performance of a 1-NN classifier with and without random projection. Report multi-class confusion matrix, precision and recall for each class with and without projection and for each value of k. Evaluate your findings with respect to the use of random projections and classification.

```
In [ ]:
```

1.2 PCA of a handwritten digits (7 points, 3+2+2)

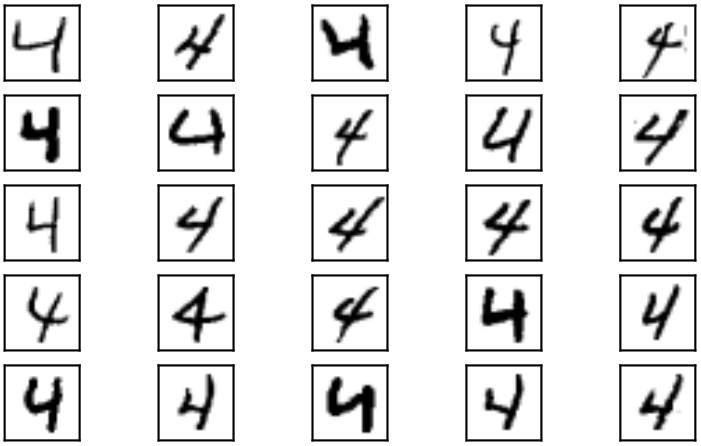
Analyze the first two principal components of the class with label 4 of the MNIST data set (those are images that each depict a handwritten "4"). Perfom the steps (a), (b), (c) described below. Note that these steps are similar to the analysis given in the lecture. Include all images and plots in your report. You may use sklearn.decomposition.PCA for this assignment. Do not scale the data.

```
In [6]: # build a list of figures on a 5x5 grid for plotting
        def buildFigure5x5(fig, subfiglist):

            for i in range(0,25):
                pixels = np.array(subfiglist[i], dtype='float')
                pixels = pixels.reshape((28, 28))
                a=fig.add_subplot(5,5,i+1)
                imgplot =plt.imshow(pixels, cmap='gray_r')
                a.axes.get_xaxis().set_visible(False)
                a.axes.get_yaxis().set_visible(False)
            return

        # find the first 25 instances with label '4' and plot them
        imgs = np.empty([25, 28*28], dtype='float')
        j=0
        for i in range(0,len(X)):
            if(y[i] == 4) and j < 25:
                imgs[j] = np.array(X[i], dtype='float')
                j += 1

        buildFigure5x5(plt.figure(1),imgs)
        plt.show()
```



1.2.1 Step (a)

Generate a scatter plot of the data in the space spanned by the first two principal components of PCA. Reconstruct 25 points on a 5×5 grid in this space that cover the variation of the data. Render each point as an image. Arrange the images in a 5×5 grid.

In [ ]:

1.2.2 Step (b)

For each of the reconstructed points, find the original instance that is closest to it in the projection on the first two components (measured using Euclidean distance). Render the instances arranged in a 5×5 grid such that their position matches the rendering in (a).

In [ ]:

1.2.3 Step (c)

Render the mean and the first two principal components as images. What is your interpretation of the first two components, i.e., which aspect of the data do they capture? Justify your interpretation, also using your results of Steps (a) and (b).

In [ ]:

1.3 Projection onto a hyperplane (4 points)

Let F be a k-dimensional hyperplane given by the parametric representation

$$g(\lambda) = \mu + \mathbf{V}\lambda,$$

where $\mu \in \mathbb{R}^d$ and the columns of $\mathbf{V}$ are pairwise orthogonal and normal vectors $\mathbf{v_1}, \ldots, \mathbf{v_k} \in \mathbb{R}^{\mathbf{d}}$. Let $f : \mathbb{R}^d \to F$ be the projection that maps every point $\mathbf{p} \in \mathbb{R}^d$ to its nearest point on F (where distances are measured using the Euclidean distance).

Prove that for any $\mathbf{p}, \mathbf{q} \in \mathbb{R}^d$, it holds that

$$\|f(\mathbf{p}) - f(\mathbf{q})\| \leq \|\mathbf{p} - \mathbf{q}\|.$$

(Hint: Assume first that $\mu = 0$. Write $f(\mathbf{p})$ and $f(\mathbf{q})$ using a projection onto the subspace spanned by the columns of V. Rewrite this using a rotation followed by an orthogonal projection. What happens to the distance in each step? Generalize to arbitrary μ.)

1.4 Locality-sensitive hashing (3 points, 1+2)

H is a family of (d_1, d_2, p_1, p_2)-locality-sensitive hash functions if it holds that

$$\text{if} \quad d(\mathbf{p}, \mathbf{q}) \leq d_1 \quad \text{then} \qquad \Pr[h(\mathbf{p}) = h(\mathbf{q})] \geq p_1 \tag{1}$$

$$\text{if} \quad d(\mathbf{p}, \mathbf{q}) \geq d_2 \quad \text{then} \qquad \Pr[h(\mathbf{p}) = h(\mathbf{q})] \leq p_2 \tag{2}$$

1.4.1 Case: $p_2 = 0$

Assume that $p_2 = 0$ and assume we have a total numer of m hash functions from this family available. Which combination of AND-constructions and OR-constructions should we use to amplify the hash family?

1.4.2 Case: $p_2 = \frac{1}{n}$

Now assume that $p_2 = \frac{1}{n}$ and assume we have n data points $\mathbf{P}$ which are stored in a hash table using a randomly chosen function h from H. Given a query point $\mathbf{q}$, we retrieve the points in the hash bucket with index $h(\mathbf{q})$ to search for a point which has small distance to $\mathbf{q}$. Let X be a random variable that is equal to the size of the set

$$\{\mathbf{p} \in \mathbf{P} : h(\mathbf{p}) = h(\mathbf{q}) \wedge d(\mathbf{p}, \mathbf{q}) \geq d_2\} \tag{3}$$

which consists of the false positives of this query.
Derive an upper bound on the expected number of false-positives $\mathbb{E}[X]$ in terms of n.