PRS L6

Principal Component Analysis – PCA

Principal Component Analysis – PCA

- Objective: dimensionality reduction, compression
- Given a set of data points lying in a high dimensional space (ND), our goal is to reduce the dimensionality (KD) of the data points while preserving as much information as possible. (e.g. 7D points (7 features) will be reduced to 2D (2 features))
- The data will be projected in a smaller K dimensional subspace
- Idea: Compute the top K “major axis of variation” (the directions on which the data approximately lies) – such that when the data is projected, the variance of the projected data is maximized

Principal Component Analysis – PCA

- https://www.youtube.com/playlist?list=PLl9TabrCI60Oa5hSWnYmaMuBexbUO95Ps
- http://cs229.stanford.edu/notes2020spring/cs229-notes10.pdf

Principal Component Analysis – PCA

- Ex 1-8

1. **Read the data** from pca2d.txt and initialize a Mat
   - `Mat X(n,d,CV_64FC1);` (e.g X has size [1000 x 7])
   - `X.at<double>(i, j)`

2. **Data normalization** Calculate the mean vector and subtract it from the data points. e.g. mean vector for X is an array of 7 values.

Principal Component Analysis – PCA

3. Compute the **covariance matrix**

   `Mat C = X.t()*X/(n-1);` (e.g C [7x7])

4. Perform the **eigenvalue decomp**osition on the covariance matrix

   ```
   Mat Lambda, Q;
   eigen(C, Lambda, Q);
   Q = Q.t();
   ```
   (e.g. Lambda [7x1], Q [7x7])

5. Print the eigenvalues. (Lambda)
   - For pca2d: First eigenvalue is 8090.21
   - For pca3d: First eigenvalue is 5462.33

Principal Component Analysis – PCA

6. Calculate the **PCA coefficients**

```
Mat Xcoeff = X * Q;
```
(e.g Xcoeff [1000 x 7])

- Compute Xk for the input data using the first kth cols from Q

```
Rect submat(0, 0, k, Q.rows);
Mat Qk = Q(submat);
```
(e.g. Qk [7x1])

```
Mat Xk = X * Qk * Qk.t();
```
(Xk has the same size as X e.g [1000 x 7])

X * Qk is the projection of X in a k-dimensional subspace

```
X * Qk * Qk.t() is the reconstruction of X from the projection
```

7. Evaluate the mean absolute difference between the original points (X) and their approximation using k principal components (Xk). When computing the mean use all the values from X and Xk.

For pca2d: Mean absolute difference using only one dimension: 22.43

For pca3d: Mean absolute difference using only one dimension: 14.50

8. For the input data from pca2d.txt select the first two columns from Xcoeff as x and y coordinates and plot the data as black 2D points on a white background (check with Fig. 2)