CIS 520, Machine Learning, Fall 2020
Homework 5
Due: Monday, October 12th, 11:59pm
Submit to Gradescope

Chun Chang

1 Perceptron vs. Winnow

1.1 Sparse target vector, dense feature vectors.

$$
\begin{aligned}
M_p &= \frac{R_2^2 u_2^2}{\gamma^2} \\
x_{t2} &= \sqrt{d} \\
u_2 &= \sqrt{k} \ll \sqrt{d} \\
M_p &= \frac{dk}{\gamma^2}
\end{aligned}
$$

$$
\begin{aligned}
M_w &= 2\frac{R_\infty^2 u_1^2}{\gamma^2} ln(d) \\
x_{t\infty} &\leq 1 \\
u_1 &= d \\
M_w &= \frac{1k^2}{\gamma^2} ln(d) \\
&\because k \ll d \\
&\therefore k^2 ln(d) < dk
\end{aligned}
$$

Winnow gives lower number of mistakes

1.2 Dense target vector, sparse feature vectors.

$$
\begin{aligned}
M_p &= \frac{R_2^2 u_2^2}{\gamma^2} \\
x_{t2} &= \sqrt{k} \\
u_2 &\leq 2\sqrt{d} \\
M_p &\leq \frac{k * 4d}{\gamma^2}
\end{aligned}
$$

$$
\begin{aligned}
M_w &= 2\frac{R_\infty^2 u_1^2}{\gamma^2} ln(d) \\
x_{t\infty} &= 1 \\
u_1 &\leq d \\
M_w &= \frac{1d^2}{\gamma^2} ln(d) \\
&\because k \ll d \\
&\therefore d^2 ln(d) > 4dk
\end{aligned}
$$

Perceptron gives lower number of mistakes

1.3 If your problem has non-negative feature vectors, is the Winnow algorithm a meaningful choice? Why or why not?

if the feature vector are non-negative

$$
\begin{aligned}
x_t &>= 0 \\
\because u_i &= \frac{u_{i-1} exp(\eta u^T x_t)}{Z_t} >= 0 \\
\therefore u^T x_t &>= 0 \\
\therefore sign(u^T x_t) &= +1
\end{aligned}
$$

The classifier would not work and we concluded that the using Winnow on non-negative features vectors is not a reasonable choice.

2 Singular Value Decomposition

2.1

$$\frac{\partial (y-Xw)^T(y-Xw)}{\partial w} = -2X^T(y-Xw)$$
$$\frac{\partial (y-Xw)^T(y-Xw)}{\partial w} = 0 \rightarrow argmin_w (y-Xw)^T(y-Xw) \text{ reaches minimum}$$
$$w = (X^TX)^{-1}X^Ty$$

$$\because X = U\Lambda V^T$$
$$w = ((U\Lambda V^T)^T U\Lambda V^T)^{-1}(U\Lambda V^T)^T y$$
$$= (V\Lambda^2 V^T)^{-1}(U\Lambda V^T)^T y$$
$$= (V\Lambda^{-2}V^{-1})(V\Lambda U^T)y$$
$$= (V\Lambda^{-1}U^T)y$$
$$\because dim(X) = nxp, p < n \text{ singular value of i row is } 0, i > p$$
$$w = \hat{w}$$
$$= V_k\Lambda_k^{-1}U_k^T y$$

2.2 eigenvectors

$$dim(XX^T) = n \text{ x } n$$
$$XX^T = U\Lambda_{left}U^T$$
$$X^TX = V\Lambda_{right}V^T$$
$$\because XX^Tu = \lambda u$$
$$(X^T)XX^Tu = (X^T)\lambda u$$
$$X^TX(X^Tu) = \lambda(X^Tu)$$
$$\therefore v_i = \lambda_i X^T u_i \text{ for } \lambda_i > 0. i < p$$

2.3 find largest eigenvectors

$\because X^TX$ and XX^T share the nonzero eigenvalue λ and the eigenvectors is the mapping of each other

if the dataset has more features than samples

$$n < p$$
$$dim(XX^T) = n \text{ x } n < p \text{ x } p$$

we could use the previous results

$$v_1 = \lambda_1 X^T u_1$$

if the dataset has more samples than feautres

$$n > p$$
$$dim(X^TX) = p \text{ x } p < n \text{ x } n$$

we could directly use SVD on X^TX

$$X^TX = V\Lambda_{right}V^T$$
$$v_1 = \text{first column of } V$$

3 Principal Component Analysis

3.1 Comparing Principal Components

1. eigenvalues and eigenvectors

$$\Lambda = [1.28, 0.59]$$
$$Q = \begin{Bmatrix} 0.70710678 & -0.70710678 \\ 0.70710678 & 0.70710678 \end{Bmatrix}$$

2. Express mathematically, explain why the first PC is the eigenvector associated with the largest eigenvalue?

   From the distortion formula

$$\sum \sum_{j=K+1}^{m} (z_j^i)^2$$

   And we plug in the function of $z_j^i = (x_j^i - \bar{x})^T u_i$
   We get

$$\sum \sum_{j=K+1}^{m} (z_j^i)^2 = n \sum_{j=K+1}^{m} u_j^T \Sigma u_j^T$$

$$\Sigma = Q\Lambda Q^T = \sum \lambda v v^T$$

   Then we use Largrangian multiplier to minimize the distortion

$$minimize(\sum \sum_{j=K+1}^{m} u_j^T \Sigma u_j^T + \sum_{j=K+1}^{m} \lambda_j u_{j_2}^2)$$

And we have

$$argmin \sum_{j=K+1}^{m} \lambda_j$$

From the formula, the eigenvectors of largest sigular value could preserve the original data X the most. So, to minimize the distorsion, we pick PCA from large eigenvalue to small one.

3. What can you say about the relationship between the first principal component and the second?

   Every principle component is orthogonal to each other, and norm of each is 1.

3.2 Plotting Principal Components in Original Space

1. Please describe how the principal components relate to the points.

   The principle components are the vectors pointing to the first two largest variance directions of the data

2. Paste the graph here

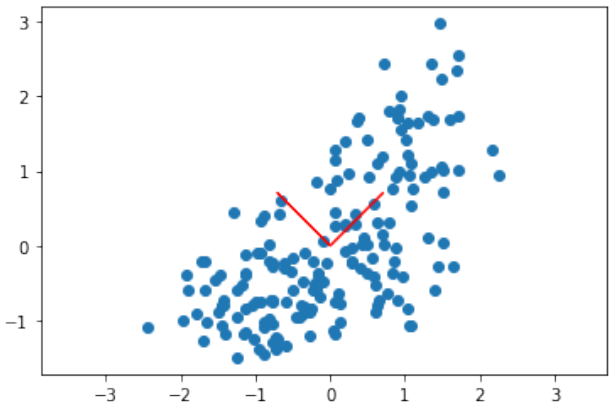

Figure 1: eigen vectors

3.3 Plotting Data Projected onto Component Space

1. Explain how the graph of points on principal component space relates to the graph of points on original space above

   Since we used full features, the projected data was the rotation of the original data.

   The data are centered at the origin of the PCs because we were trying to find out the directions of largest variance.

2. Explain the difference in distribution of points projected on the first component vs. projected on the second.

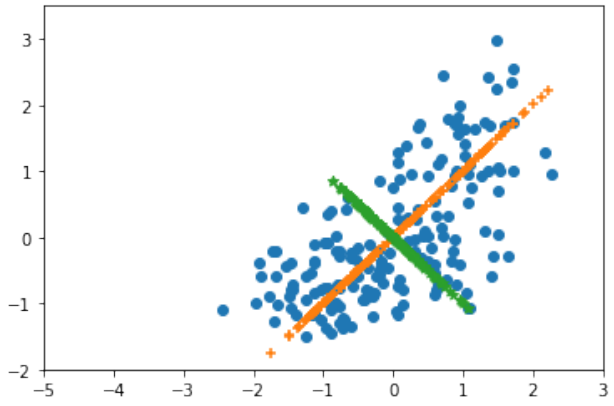

Figure 2: eigen vectors

Origin dots are the data projected onto the first PC, green dots are the data projected onto the second. The projected data on first PC has larger variance.

3. Paste the plot of the given points

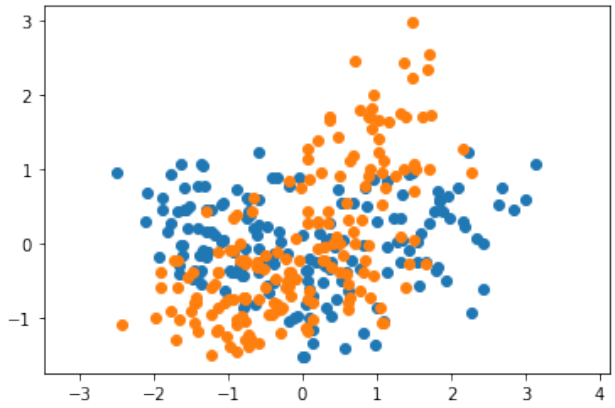

Figure 3: eigen vectors

3.4 PCA and Reconstruction Error

1. what is the reconstruction error using the first and second principal components.

$$err_{i1,2} = [148279353, 152309802]$$
$$err_{12} = [130114349]$$

2. Mathematically express how you come up with the answer, and explain how PCA is minimizing the reconstruction error..

From the distortion formula

$$\sum \sum_{j=K+1}^{m} (z_j^i)^2$$

And we plug in the function of $z_j^i = (x_j^i - \bar{x})^T u_i$
We get

$$\sum \sum_{j=K+1}^{m} (z_j^i)^2 = n \sum_{j=K+1}^{m} u_j^T \Sigma u_j^T$$

$$\Sigma = Q \Lambda Q^T = \sum \lambda v v^T$$

Then we use Largrangian multiplier to minimize the distortion

$$minimize(\sum \sum_{j=K+1}^{m} u_j^T \Sigma u_j^T + \sum_{j=K+1}^{m} \lambda_j u_{j_2}^2)$$

And we have

$$argmin \sum_{j=K+1}^{m} \lambda_j$$

In conclusion, we do not choose the eigenvectors corresponding to small eigenvalues as the PCA to minimize the distortion(reconstruction error).

4 Principal Component Analysis on Faces

4.1 PCA with SVD and Resulting Eigenfaces

1. report the singular values here.

$$\begin{bmatrix} 86.7 & 66.4 & 50.1 & 39.7 & 33.7 & 31.5 & 27.6 & 25.3 & 24.8 & 22.9 \\ 22.4 & 21.2 & 19.8 & 19.0 & 18.3 & 17.5 & 17.0 & 16.0 & 15.4 & 15.3 \\ 14.8 & 13.9 & 13.5 & 13.4 & 13.1 & 12.9 & 12.7 & 12.5 & 12.0 & 11.8 \\ 11.2 & 11.0 & 10.6 & 10.2 & 10.0 & 9.9 & 9.8 & 9.7 & 9.4 & 9.3 \\ 9.0 & 8.9 & 8.7 & 8.7 & 8.5 & 8.4 & 8.3 & 8.3 & 8.1 & 8.0 \end{bmatrix}$$

2. eigenfaces look like.
SVD projected the all the photos into subspace, and eigenvector preserves the eigen features of all the photo. The eigenface with largest eigenvalue preserves the most common features of all the photos

3. insert your eigenfaces output here.

4.2 PCA with SVD and Resulting Eigenfaces

1. How are they similar and how are they different.

The reconstructed image was mostly similar to the original image because the eigenvectors preserved the important features. Some features were lost because PCA discarded some "noisy" data, and it caused the reconstructed image blurry with less information.

2. What do you expect to see from the reconstructed images as the number of principal components chosen for PCA increases?

When we increase the number of features for reconstructing the image, we shall see the reconstructed image more similar to the original image as the distortion gets down until it reaches 0. Which means we use all features to reconstruct the image, it's obvious we shall get the original image back.

4.3 Variance Explanation

1. the plot of variance explained as the number of components to relate to the eigenvalues of the corresponding components?

$$EV = \frac{\lambda_i}{\sum \lambda_i} 100\%$$

So, if we take eigenvectors in descending order, we shall expect to see the curve of EV gradually cumulates until 100%

2. What is the relation between reconstruction error and the variance explained?

The reconstructed error goes down as we use more eigenvectors for PCA. So as the number of features goes up, we shall see the error moving towards to zero.

To summarize the 1. and 2., the plot of explained variance and the reconstructed error shall be complementary.

3. Insert the three line plots of explanation vs. number of components, descending eigenvalues vs. number of components, and reconstruction error vs. number of components here

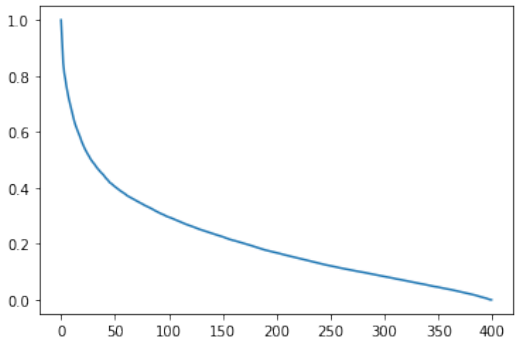

Figure 4: reconstruct error

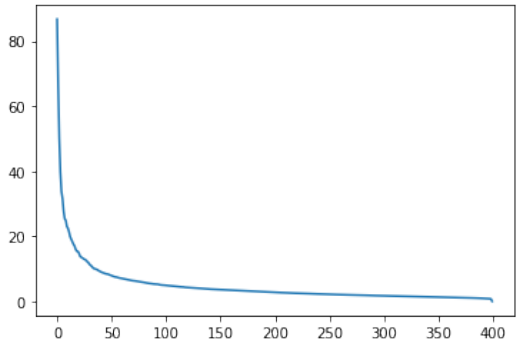

Figure 5: eigen value

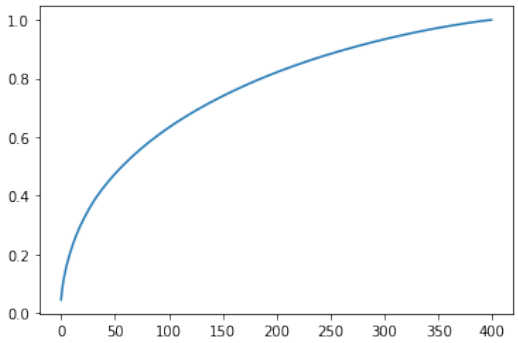

Figure 6: cumulative eigenvalues