CMSC 25400 Machine Learning
Homework 2

Deqing Fu

October 25, 2017

Problem 1

(a)

Proof. $\forall \mathbf{v_j} \in \{\mathbf{v_1}, \mathbf{v_2}, \cdots, \mathbf{v_d}\}$,

$$
\begin{aligned}
A\mathbf{v_j} &= (\sum_{i=1}^{d} \lambda_i \mathbf{v_i}\mathbf{v_i}^T)\mathbf{v_j} \\
&= \sum_{i=1}^{d} \lambda_i \mathbf{v_i}\mathbf{v_i}^T\mathbf{v_j} \\
&= \sum_{i=1}^{d} \lambda_i \mathbf{v_i}(\mathbf{v_i}^T\mathbf{v_j}) \\
&= \sum_{i=1}^{d} \lambda_i \mathbf{v_i}\langle \mathbf{v_i}, \mathbf{v_j}\rangle \qquad (\langle\cdot,\cdot\rangle \text{ denotes inner product of two vectors})
\end{aligned}
$$

As $\{\mathbf{v_1}, \mathbf{v_2}, \cdots, \mathbf{v_d}\}$ are mutually orthogonal unit vectors, that is $\langle \mathbf{v_i}, \mathbf{v_j}\rangle = \begin{cases} 1, i = j \\ 0, i \neq j \end{cases}$. Hence $A\mathbf{v_j} = \sum_{i=1}^{d} \lambda_i \mathbf{v_i}\langle \mathbf{v_i}, \mathbf{v_j}\rangle = 0 + 0 + \cdots \lambda_j \mathbf{v_j} \cdot \langle \mathbf{v_j}, \mathbf{v_j}\rangle + 0 + 0 + \cdots + 0 = \lambda_j \mathbf{v_j}$. Hence, λ_j is an eigenvalue relative to $\mathbf{v_j}$. Thus, $\mathbf{v_1}, \mathbf{v_2}, \cdots, \mathbf{v_d}$ are eigenvectors with corresponding eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_d$. $\square$

(b)

Per eigenvalue decomposition, we know that $A = Q\Lambda Q^{-1}$, where $\Lambda = diag(\lambda_1, \lambda_2, \cdots, \lambda_d)$ and $Q = \begin{bmatrix}\mathbf{v_1} & \mathbf{v_2} \cdots \mathbf{v_d}\end{bmatrix}$. As $\{\mathbf{v_1}, \mathbf{v_2}, \cdots, \mathbf{v_d}\}$ is a orthonormal basis, then

$$
\begin{aligned}
Q \cdot Q^T &= \begin{bmatrix}\mathbf{v_1} & \mathbf{v_2} \cdots \mathbf{v_d}\end{bmatrix} \cdot \begin{bmatrix}\mathbf{v_1}^T \\ \mathbf{v_2}^T \\ \vdots \\ \mathbf{v_d}^T\end{bmatrix} \\
&= \sum_{k=1}^{d} \mathbf{v_k}\mathbf{v_k}^T \\
&= \begin{bmatrix}1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1\end{bmatrix} = I
\end{aligned}
$$

Hence, $Q^T = Q^{-1}$, then

$$
A = Q\Lambda Q^T = \begin{bmatrix}\mathbf{v_1} & \mathbf{v_2} \cdots \mathbf{v_d}\end{bmatrix} \cdot diag(\lambda_1, \lambda_2, \cdots, \lambda_d) \cdots \begin{bmatrix}\mathbf{v_1}^T \\ \mathbf{v_2}^T \\ \vdots \\ \mathbf{v_d}^T\end{bmatrix} = \sum_{i=1}^{d} \mathbf{v_i}\lambda_i\mathbf{v_i}^T = \sum_{i=1}^{d} \lambda_i\mathbf{v_i}\mathbf{v_i}^T
$$

Problem 2

(a)

Proof. Consider two eigenvectors, $\mathbf{v_i}, \mathbf{v_j} \in \mathbb{R}^d$ with corresponding eigenvalues λ_i, λ_j, where $\lambda_i \neq \lambda_j$. Then

$$
\lambda_i\langle \mathbf{v_i}, \mathbf{v_j}\rangle = \langle \lambda_i\mathbf{v_i}, \mathbf{v_j}\rangle = \langle A\mathbf{v_i}, \mathbf{v_j}\rangle = \langle \mathbf{v_i}, A\mathbf{v_j}\rangle = \langle \mathbf{v_i}, \lambda_j\mathbf{v_j}\rangle = \lambda_j\langle \mathbf{v_i}, \mathbf{v_j}\rangle
$$

As $\lambda_i \neq \lambda_j$, $\langle \mathbf{v_i}, \mathbf{v_j}\rangle = 0$, that is v_i and v_j are orthogonal. □

(b)

Proof. As $\{\mathbf{v_1}, \mathbf{v_2}, \cdots, \mathbf{v_d}\}$ are normalized. Then $\langle \mathbf{v_i}, \mathbf{v_i}\rangle = ||v_i||^2 = 1$. Hence, we have $\langle \mathbf{v_i}, \mathbf{v_j}\rangle = \begin{cases} 1, i = j \\ 0, i \neq j \end{cases}$. Then, by definition, $\{\mathbf{v_1}, \mathbf{v_2}, \cdots, \mathbf{v_d}\}$ is an orthonormal basis. □

(c)

Proof. Let $\mathbf{w} = \sum_{i=1}^{d} \alpha_i \mathbf{v_i}$ be a linear combination of the orthonormal basis. Then

$$\begin{aligned}
R(\mathbf{w}) &= \frac{\mathbf{w}^T A \mathbf{w}}{\mathbf{w}^T \mathbf{w}} \\
&= \frac{(\sum_{i=1}^{d} \alpha_i \mathbf{v_i})^T (\sum_{i=1}^{d} \alpha_i A \mathbf{v_i})}{\langle \sum_{i=1}^{d} \alpha_i \mathbf{v_i}, \sum_{i=1}^{d} \alpha_i \mathbf{v_i} \rangle} \\
&= \frac{(\sum_{i=1}^{d} \alpha_i \mathbf{v_i}^T)(\sum_{i=1}^{d} \alpha_i \lambda_i \mathbf{v_i})}{\sum_{i=1}^{d} \alpha_i^2} \\
&= \frac{\sum_{i=1}^{d} \alpha_i^2 \lambda_i ||\mathbf{v_i}||^2}{\sum_{i=1}^{d} \alpha_i^2} \\
&= \frac{\sum_{i=1}^{d} \alpha_i^2 \lambda_i}{\sum_{i=1}^{d} \alpha_i^2} \\
&\leq \frac{\sum_{i=1}^{d} \alpha_i^2 \lambda_d}{\sum_{i=1}^{d} \alpha_i^2} \qquad \text{Because } \forall i,\ \lambda_i \leq \lambda_d \\
&= \lambda_d
\end{aligned}$$

And the equal sign can be satisfied when $\forall i$, $\lambda_i = \lambda_d$ and at this time, $\mathbf{w} = \mathbf{v_d}$. Hence, when the maximum is reached, $R(\mathbf{w}) = \lambda_d$ and $\mathbf{w} = \mathbf{v_d}$. □

Problem 3

(a)

Proof.

$$\begin{aligned}
\hat{\Sigma}^{(1)} \mathbf{v_d} &= \Big(\frac{1}{n} \sum_{i=1}^{n} (\mathbf{x_i} - (\mathbf{x_i} \cdot \mathbf{v_d}) \mathbf{v_d})(\mathbf{x_i} - (\mathbf{x_i} \cdot \mathbf{v_d}) \mathbf{v_d})^T)\Big) \mathbf{v_d} \\
&= \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x_i} - (\mathbf{x_i} \cdot \mathbf{v_d}) \mathbf{v_d})((\mathbf{x_i} - (\mathbf{x_i} \cdot \mathbf{v_d}) \mathbf{v_d})^T \mathbf{v_d}) \\
&= \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x_i} - (\mathbf{x_i} \cdot \mathbf{v_d}) \mathbf{v_d}) \langle (\mathbf{x_i} - (\mathbf{x_i} \cdot \mathbf{v_d}) \mathbf{v_d}), \mathbf{v_d} \rangle \\
&= \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x_i} - (\mathbf{x_i} \cdot \mathbf{v_d}) \mathbf{v_d})(\langle \mathbf{x_i}, \mathbf{v_d} \rangle - \langle \mathbf{x_i}, \mathbf{v_d} \rangle \langle \mathbf{v_d}, \mathbf{v_d} \rangle) \\
&= \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x_i} - (\mathbf{x_i} \cdot \mathbf{v_d}) \mathbf{v_d})(\langle \mathbf{x_i}, \mathbf{v_d} \rangle - \langle \mathbf{x_i}, \mathbf{v_d} \rangle) \qquad \text{Because } \langle \mathbf{v_d}, \mathbf{v_d} \rangle = 1 \\
&= 0
\end{aligned}$$

It's pretty simple that $\hat{\Sigma}\mathbf{v_k} = (\sum_{i=1}^{d} \lambda_i \mathbf{v_i}\mathbf{v_i}^T)\mathbf{v_k} = \sum_{i=1}^{d} \lambda_i \mathbf{v_i}\langle \mathbf{v_i}, \mathbf{v_k}\rangle = \lambda_k \mathbf{v_k}$

$$\begin{aligned}
\hat{\Sigma}^{(1)}\mathbf{v_k} &= (\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x_i} - (\mathbf{x_i}\cdot\mathbf{v_d})\mathbf{v_d})(\mathbf{x_i} - (\mathbf{x_i}\cdot\mathbf{v_d})\mathbf{v_d})^T))\mathbf{v_k} \\
&= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x_i} - (\mathbf{x_i}\cdot\mathbf{v_d})\mathbf{v_d})((\mathbf{x_i} - (\mathbf{x_i}\cdot\mathbf{v_d})\mathbf{v_d})^T)\mathbf{v_k}) \\
&= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x_i} - (\mathbf{x_i}\cdot\mathbf{v_d})\mathbf{v_d})\langle(\mathbf{x_i} - (\mathbf{x_i}\cdot\mathbf{v_d})\mathbf{v_d}), \mathbf{v_k}\rangle \\
&= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x_i} - (\mathbf{x_i}\cdot\mathbf{v_d})\mathbf{v_d})(\langle\mathbf{x_i}, \mathbf{v_k}\rangle - \langle\langle\mathbf{x_i}, \mathbf{v_d}\rangle\mathbf{v_d}, \mathbf{v_k}\rangle) \\
&= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x_i} - (\mathbf{x_i}\cdot\mathbf{v_d})\mathbf{v_d})(\langle\mathbf{x_i}, \mathbf{v_k}\rangle - \langle\mathbf{x_i}, \mathbf{v_d}\rangle\langle\mathbf{v_d}, \mathbf{v_k}\rangle) \\
&= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x_i} - (\mathbf{x_i}\cdot\mathbf{v_d})\mathbf{v_d})(\langle\mathbf{x_i}, \mathbf{v_k}\rangle - \langle\mathbf{x_i}, \mathbf{v_d}\rangle\cdot 0) \qquad \text{Because } \mathbf{v_d} \text{ and } \mathbf{v_k} \text{ are orthogonal} \\
&= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x_i} - (\mathbf{x_i}\cdot\mathbf{v_d})\mathbf{v_d})\langle\mathbf{x_i}, \mathbf{v_k}\rangle \\
&= \frac{1}{n}\sum_{i=1}^{n}\langle\mathbf{x_i}, \mathbf{v_k}\rangle\mathbf{x_i} - \frac{1}{n}\sum_{i=1}^{n}\langle\mathbf{x_i}, \mathbf{v_d}\rangle\langle\mathbf{x_i}, \mathbf{v_k}\rangle\mathbf{v_d} \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbf{x_i}^T\mathbf{v_k}\mathbf{x_i} - \frac{1}{n}\sum_{i=1}^{n}\mathbf{v_d}^T(\mathbf{x_i}\mathbf{x_i}^T)\mathbf{v_k}\mathbf{v_d} \\
&= (\frac{1}{n}\sum_{i=1}^{n}\mathbf{x_i}\mathbf{x_i}^T)\mathbf{v_k} - \mathbf{v_d}^T\hat{\Sigma}\mathbf{v_k}\mathbf{v_d} \\
&= \hat{\Sigma}\mathbf{v_k} - \mathbf{v_d}^T(\hat{\Sigma}\mathbf{v_k})\mathbf{v_d} \\
&= \lambda_k\mathbf{v_k} - \mathbf{v_d}^T(\lambda_k\mathbf{v_k})\mathbf{v_d} \\
&= \lambda_k\mathbf{v_k} - \lambda_k(\mathbf{v_d}^T\mathbf{v_k})\mathbf{v_d} \\
&= \lambda_k\mathbf{v_k} - 0 = \lambda_k\mathbf{v_k}
\end{aligned}$$

□

Now, $\hat{\Sigma}^{(1)}$ has eigenvalues $\lambda_1, \lambda_2 \cdots, \lambda_{d-1}$ with corresponding eigenvectors $\mathbf{v_1}, \mathbf{v_2}, \cdots, \mathbf{v_{d-1}}$. Then if we apply Rayleigh Quotient again to achieve maximum, $\mathbf{v_{d-1}}$ is the second principal component.

(b)

Proof. Consider when $k = 1$, then this is the case of v_d being the principal component. Note $\hat{\Sigma}^{(k)} = \frac{1}{n}\sum_{i=1}^{n}((\mathbf{x_i} - \sum_{j=1}^{k}(\mathbf{x_i}\cdot\mathbf{v_{d-j+1}})\mathbf{v_{d-j+1}})(\mathbf{x_i} - \sum_{j=1}^{k}(\mathbf{x_i}\cdot\mathbf{v_{d-j+1}})\mathbf{v_{d-j+1}})^T)$. If $k = t$ is true, that is the t'th principal component of the data is $\mathbf{v}_{d-t+1}$. Then we have the reduced empirical covariance matrix $\hat{\Sigma}^{(t)}$ has eigenvalues of $\lambda_1, \cdots, \lambda_{d-t}$, and the $(t+1)$'th principal component of the data is $\mathbf{v}_{d-t} = \mathbf{v}_{d-(t+1)+1}$, that is the $k = t+1$ case is true. Hence by induction, the n'th principal component of the data is $\mathbf{v}_{d-k+1}$. $\square$

Problem 4

Figure 1: Principal Component Analysis

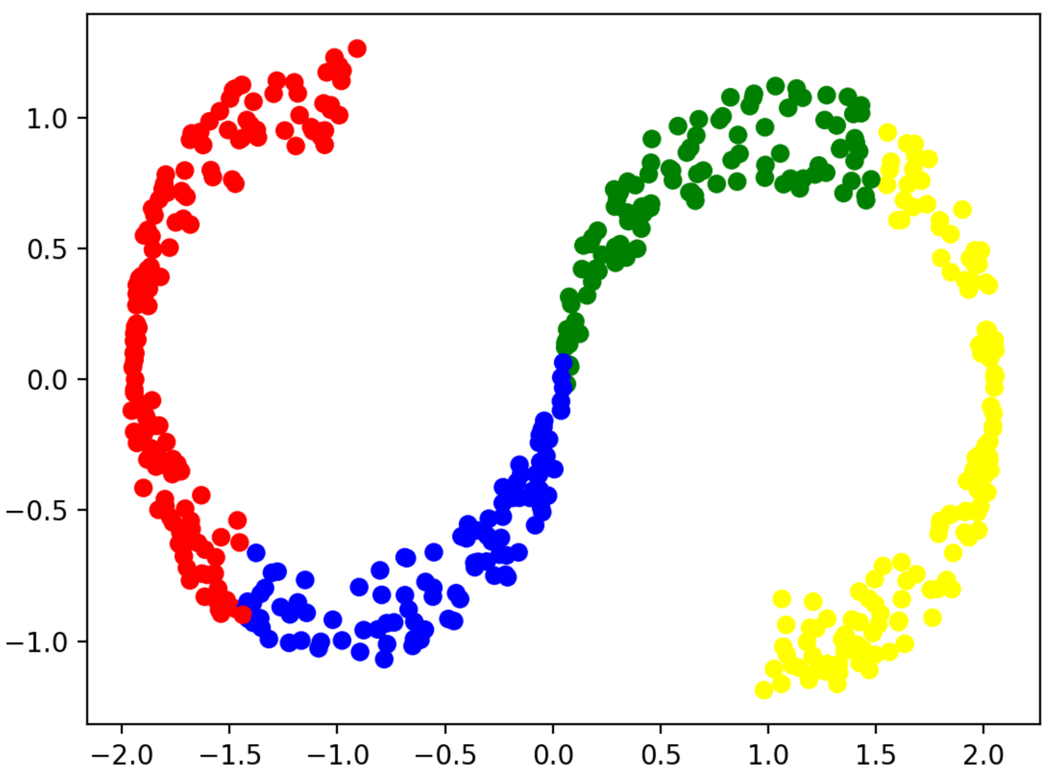
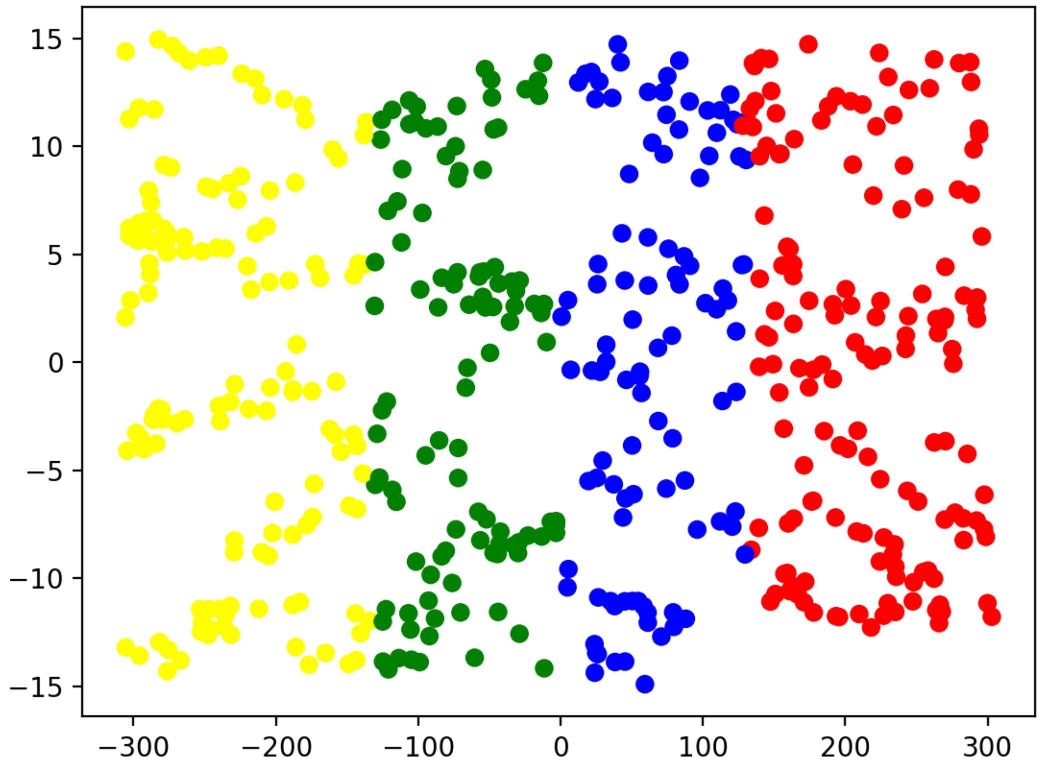


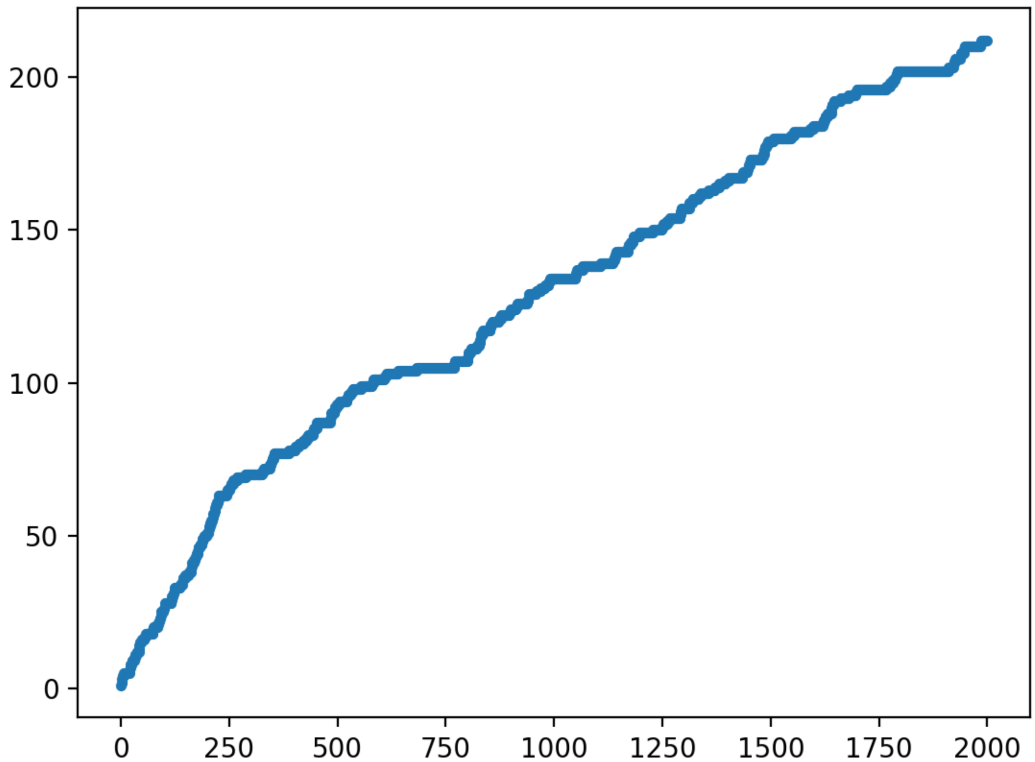
Figure 2: Isomap



Comment. The difference is that PCA is a linear algorithm that projects all the data points onto a principal hyperplane consisted of the first principal component and the second principal component while Isomap is a non-linear algorithm that stretched the three dimensional manifold into a two dimensional mesh that (approximately) preserves the distance between two data points.

Problem 5

Figure 3: Cumulative Plots of Mistakes



Here is the table of ε corresponds to M:

M	ε
1	0.1935
2	0.1405
3	0.199
4	0.2425
5	0.1615
6	0.1275
7	0.1275
8	0.1275
9	0.1275
10	0.1275

So, when $M = 6$, ε achieves its minimum.