Homework 6

ESE 402/542

December 17, 2020

Solution 1. (Principal Component Analysis)

(a) Solution:

The Two PCA Components are:

[[0.70710678 0.70710678]

 [0.70710678 -0.70710678]]

The Transformed Dataset using the first Principal Component is:

[[-1.87305312]

 [-1.30527815]

 [-0.73750317]

 [0.28355177]

 [0.34079927]

 [0.85132674]

 [2.44015666]]

(b) Solution:

The Two PCA Components are:

[[0.65908697 0.75206673]

 [0.75206673 -0.65908697]]

The Transformed Dataset is:

[[-2.37927216]

 [-1.72018519]

 [-1.06109821]

 [0.44303524]

 [0.35005549]

 [1.10212221]

 [3.26534264]]

The above results show that Principal Component Analysis is not scale invariant, that is, multiplying each feature vector by a non-zero number changes the Principal Components.

Therefore, it is essential to standardize the dataset before using PCA.

Solution 2. (k-means is suboptimal)

Consider an example with 4 data points in 2D plane: $(\sqrt{t},1),(-\sqrt{t},1),(\sqrt{t},-1),(-\sqrt{t},-1)$ where $t>1$. If $k=2$, we know that optimal partition is $C_1=\{(\sqrt{t},1),(\sqrt{t},-1)\}, C_2=\{(-\sqrt{t},1),(-\sqrt{t},-1)\}$, $\text{OPT}=\min_{c_1,c_2,\cdots,c_k}\sum_{i=1}^{n}||x_i-c(x_i)||_2^2=4$ with centers $(\sqrt{t},0),(-\sqrt{t},0)$. Now if we initialize on centers: $(0,1),(0,-1)$, we can no longer update clusters, algorithm reaches end, and $\min_{c_1,c_2,\cdots,c_k}\sum_{i=1}^{n}||x_i-c(x_i)||_2^2=4t=t\times\text{OPT}$.

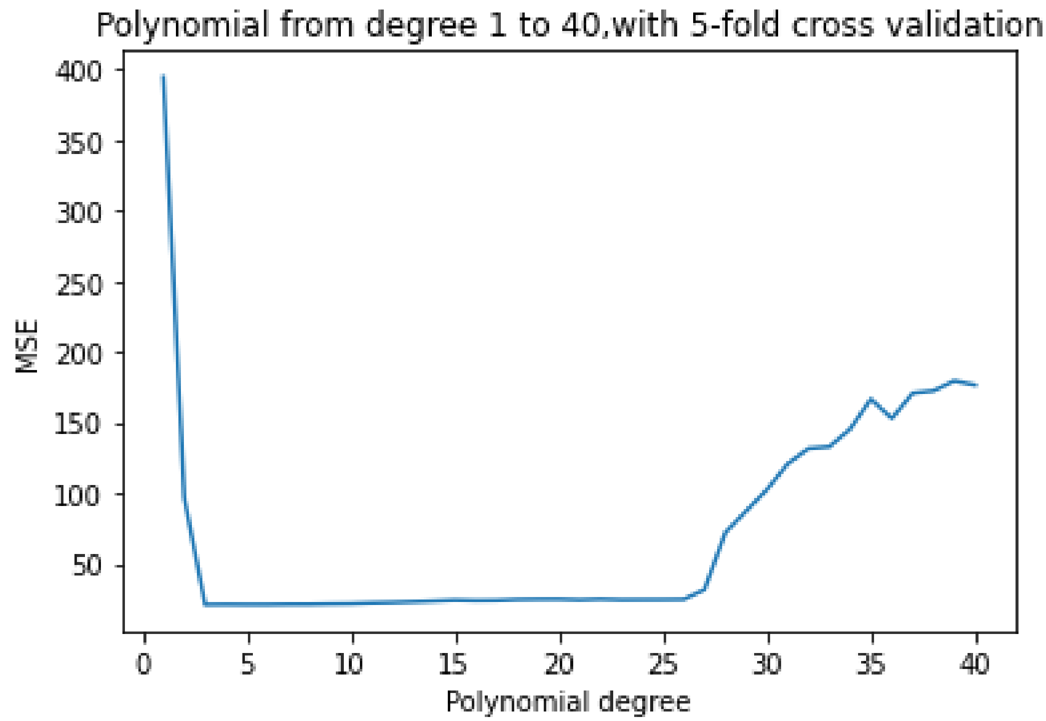
We can easily generalize this example for p-dimension:
$(\sqrt{t},1,0,0,\cdots,0)$, $(-\sqrt{t},1,0,0,\cdots,0)$, $(\sqrt{t},-1,0,0,\cdots,0)$, $(-\sqrt{t},-1,0,0,\cdots,0)$.

Or N data points:
$(\sqrt{t},0)$,$(-\sqrt{t},0)$,$(\sqrt{t},2)$,$(-\sqrt{t},2)$,$(\sqrt{t},4)$,$(-\sqrt{t},4)\cdots$,$(\sqrt{t},2n)$,$(-\sqrt{t},2n)$ with initialized centers $(0,0),(0,2),\cdots,(0,n)$.
There are $N=4n$ data points and $k=2n$ clusters. For $N=4n+1/4n+2/4n+3$, we can just add 1/2/3 data points which are very far away from any other points, (Like $(10^8t,0),(-10^8t,0),(0,-10^8n)$ etc.) and $k=2n+1/2n+2/2n+3$.

Or k clusters:
Above example, $k=2n$ or $2n+1$.

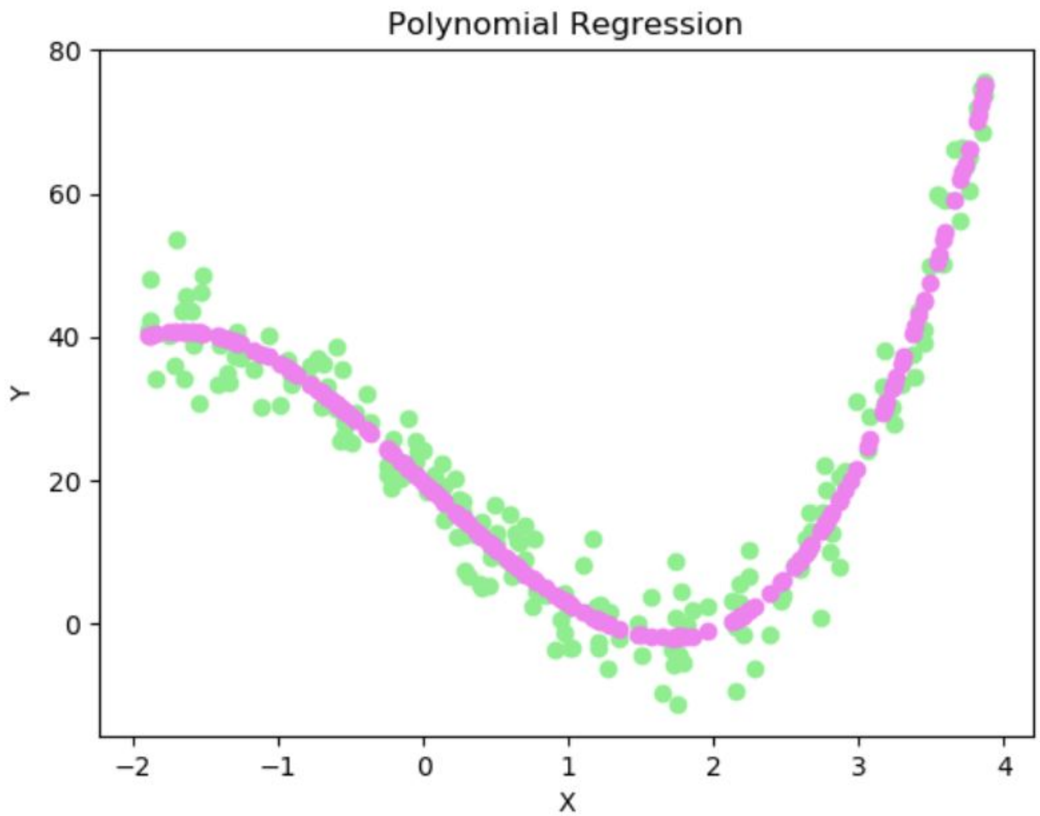
Solution 3. (Polynomial Regression)

(a) We examine the polynomials from 1 to 40 and choose k = 5.



It can be seen that the lowest MSE is when the degree is 3.

(b) Since the lowest classification error was when the degree of the polynomial was 3, I chose the cubic model for part b.
After running polynomial regression, the coefficients calculated are as follows:
0, -19.04905889, -0.09961775, 2.24871642.
There are 4 coefficients since it's a cubic polynomial. (B0, B1, B2 and B3)
The scatter plot is shown below:



The green scatter plot is of the original data. The violet scatter plot is of the cubic curve (that is, $y_{hat} = B_0 + B_1 * x + B_2 * x^2 + B_3 * x^3$).

Solution 4. (Classification)

1. Solution:

We can re-state given equations into:

$$P(X=x \mid Y=y) = \frac{1}{\sqrt{2\pi}} exp(-\frac{(x-5y)^2}{2})$$

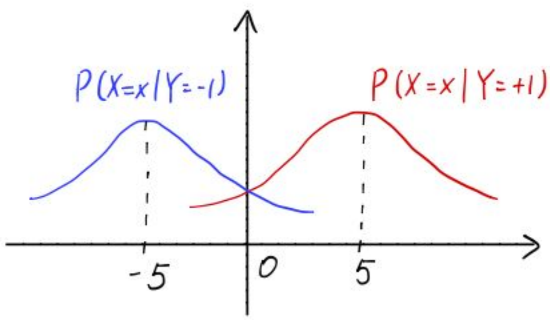
$$\therefore \quad P(X=x, Y=y)$$

$$= P(Y=y)\, P(X=x \mid Y=y)$$

$$= \frac{1}{2} \cdot \frac{1}{\sqrt{2\pi}} exp(-\frac{(x-5y)^2}{2})$$

$$= \frac{1}{2\sqrt{2\pi}} exp(-\frac{(x-5y)^2}{2})$$

2. Solution:



3. Solution:

$$P(Y=y \mid X=x) = \frac{P(X=x, Y=y)}{P(X=x)}$$

$$= P(Y=y \mid X=x) \cdot \frac{P(Y=y)}{P(X=x)}$$

$$\therefore \quad \hat{y} = h(x) = \arg\max_y \; P(Y=y \mid X=x)$$

$$= \arg\max_y \; P(Y=y \mid X=x) \cdot \frac{P(Y=y)}{P(X=x)}$$

$$= \arg\max_y \; P(X=x \mid Y=y)$$

$$= \begin{cases} +1, & \text{if } x > 0 \\ -1, & \text{otherwise} \end{cases} \quad \text{by plot}$$

4. Solution:

$$
\begin{aligned}
&\text{error} \\
&= \mathbb{E}_{(X,Y)\sim P}\left[\mathbb{1}\left(h(x) \neq Y\right)\right] \\
&= P_{(X,Y)\sim P}\left(h(x) \neq Y\right) \\
&= P(X>0)\,P(Y=-1 \mid X=x) + P(X<0)\,P(Y=+1 \mid X<0) \\
&= P(Y=-1, X>0) + P(Y=+1, X<0) \\
&= P(Y=-1)\,P(X>0 \mid Y=-1) + P(Y=+1)\,P(X<0 \mid Y=+1) \\
&= 2 \cdot \frac{1}{2}\left(1-\Phi\left(\frac{5}{1}\right)\right) \\
&= 1-\Phi(5)
\end{aligned}
$$

5. Solution:

Estimate: $P(Y=+1 \mid X=x)$, $P(Y=-1 \mid X=x)$

Model: $P(Y=+1 \mid X=x) = \dfrac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$

$$P(Y=-1 \mid X=x) = \frac{1}{1+e^{\beta_0+\beta_1 x}}$$

Optimization:

$$
\begin{aligned}
L(\beta_0, \beta_1 \mid x) &= Pr(Y \mid X; \beta_0, \beta_1) \\
&= \prod_{i=1}^{n} P(Y=y_i \mid X=x_i; \beta_0, \beta_1)
\end{aligned}
$$

$$\max_{\beta_0, \beta_1} L(\beta_0, \beta_1 \mid x)$$

6. Solution:

From: $P(X=x \mid Y=+1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-5)^2}{2}\right)$, $P(Y=+1)=\frac{1}{2}$

$P(X=x \mid Y=-1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x+5)^2}{2}\right)$, $P(Y=-1)=\frac{1}{2}$

$$
\begin{aligned}
& P(Y=+1 \mid X=x) \\
&= \frac{P(X=x,\ Y=+1)}{P(X=x,\ Y=+1)+P(X=x,\ Y=-1)} \\
&= \frac{P(X=x \mid Y=+1)P(Y=+1)}{P(X=x \mid Y=+1)P(Y=+1)+P(X=x \mid Y=-1)P(Y=-1)} \\
&= \frac{\frac{1}{2}\cdot\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x-5)^2}{2}\right)}{\frac{1}{2}\cdot\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x-5)^2}{2}\right)+\frac{1}{2}\cdot\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x+5)^2}{2}\right)} \\
&= \frac{\exp\left(-\frac{(x-5)^2}{2}\right)}{\exp\left(-\frac{(x-5)^2}{2}\right)+\exp\left(-\frac{(x+5)^2}{2}\right)} \\
&= \frac{1}{\exp\left(\frac{(x-5)^2}{2}-\frac{(x+5)^2}{2}\right)+1} \\
&= \frac{1}{\exp(-10x)+1}
\end{aligned}
$$

Note that logistic regression would try estimate

$$P(Y=+1 \mid X=x) = \frac{1}{1+\exp(-\beta_1 x-\beta_0)}$$

When n is large, which means there is enough data to estimate probability accurately,

So $\beta_0 = 0$, $\beta_1 = 10$.

Solution 5. (Extra Credit)

(a) Some eigenfaces associated with the first 25 principal components:

(b) Reconstructed faces:

Original faces:

It can be seen that the reconstructed faces are very similar to the original faces. This is because the amount of capture of data energy/variation reduces as we go to the lower principal components. Here, we considered the first 150 components. So, these components capture the most energy (PC1 has the highest, followed by PC2, and so on).