CW4

IMPERIAL COLLEGE LONDON

DEPARTMENT OF COMPUTING

CO496

Author:
Chengyi Zhang (CID: 01604256)

Date: November 29, 2018

Q1 From Fig. 1 and Fig. 2, the error rate of PCA is higher than whitening PCA.

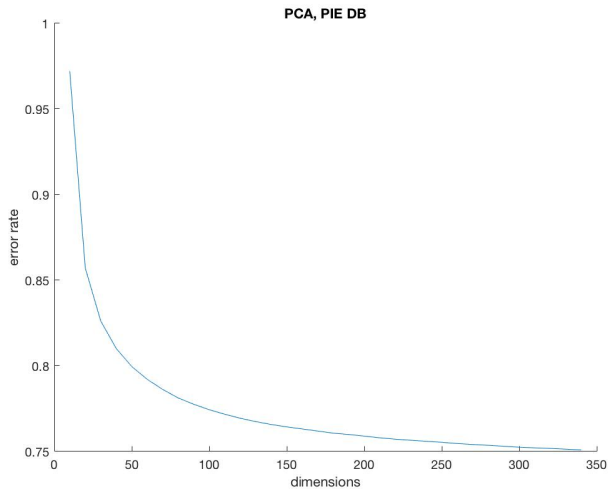

Figure 1: PCA

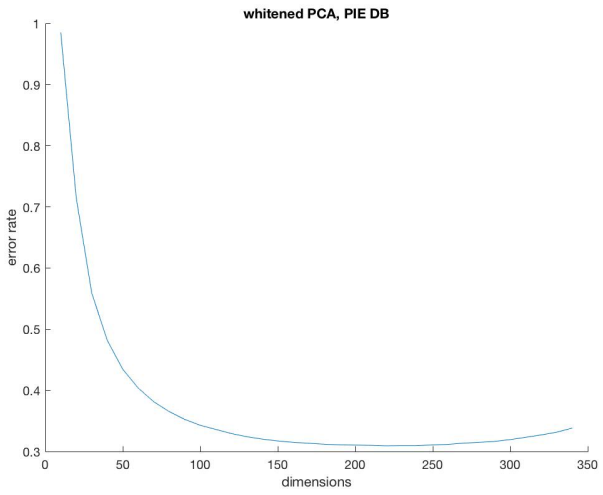

Figure 2: wPCA

When the dimension grows, both of them have similar change. The error rate decreases rapidly at start, and then the change gets slower as dimension is getting larger. For WPCA, the error rate even increases at the end. The rapid reduction at start is because the column eigenvectors in the base are placed corresponding to the eigenvalues sorted in descending order, which means that the eigenvectors in lower dimensions (top principle components) contain more information of the data than the ones in higher dimension. In WPCA, the data is normalised to have unit variance (the co-variance matrix is identity). Thus, the model with WPCA has better performance. Also, in WPCA, the smallest eigenvalues are close to zeros, and thus may cause numerical instability. The increase of error rate at the end of the curve WPCA may be because of the numerical instability by the small eigenvalues.

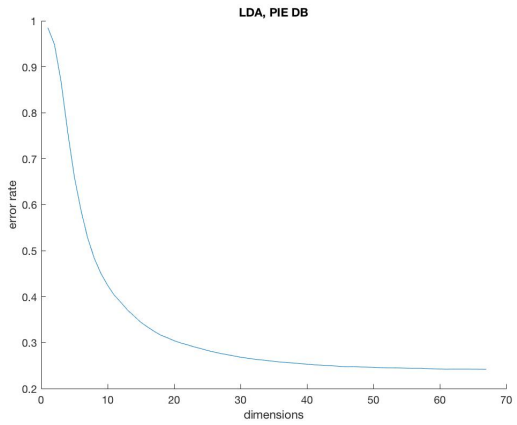

Figure 3: LDA

LDA has best performance among these three analysis algorithms. Compared to PCA, LDA is explicitly defined for classification problems, and is designed to define a latent space that helps classify data. Since PCA is also performed on projected data $\tilde{\boldsymbol{X}}_b = \boldsymbol{U}^\top \boldsymbol{X}\boldsymbol{M}$ in LDA, the columns at lower dimensions contain more information. And the performance curve shows similar pattern to that of PCA: rapidly decreases at start, and then becomes flat.

Q2

2.1

$$\min_{R,\ \boldsymbol{\alpha},\ \xi_i} R^2 \ + \ C\sum_{i=1}^{n}\xi_i \tag{1}$$

$$\text{subject to } (\boldsymbol{x}_i-\boldsymbol{\alpha})^\top(\boldsymbol{x}_i-\boldsymbol{\alpha}) \le R^2+\xi_i, \forall \xi_i \ge 0.$$

Set $f(R,\ \boldsymbol{\alpha},\ \xi_i) = R^2 \ + \ C\sum_{i=1}^{n}\xi_i$, $g_i(R,\ \boldsymbol{\alpha},\ \xi_i) = (\boldsymbol{x}_i-\boldsymbol{\alpha})^\top(\boldsymbol{x}_i-\boldsymbol{\alpha}) - R^2 - \xi_i$, and $h_i(R,\ \boldsymbol{\alpha},\ \xi_i) = -\xi_i$, then rewrite problem (1) to:

$$\begin{gathered}\min_{R,\ \boldsymbol{\alpha},\ \xi_i} f(R,\ \boldsymbol{\alpha},\ \xi_i)\\ \text{subject to } g_i(R,\ \boldsymbol{\alpha},\ \xi_i) \le 0,\\ h_i(R,\ \boldsymbol{\alpha},\ \xi_i) \le 0,\ \forall i = 1...n\end{gathered}$$

The Lagrangian is:

$$L(R,\ \boldsymbol{\alpha},\ \xi_i, \mu_i, \lambda_i) = f(R,\ \boldsymbol{\alpha},\ \xi_i) + \sum_{i=1}^{n}\mu_i g_i(R,\ \boldsymbol{\alpha},\ \xi_i) + \sum_{i=1}^{n}\lambda_i h_i(R,\ \boldsymbol{\alpha},\ \xi_i) \tag{2}$$

The primal is:

$$\min_{R,\ \boldsymbol{\alpha},\ \xi_i}\max_{\mu_i,\lambda_i\ge 0} L(R,\ \boldsymbol{\alpha},\ \xi_i, \mu_i, \lambda_i)$$

And the dual is:

$$\max_{\mu_i,\lambda_i\ge 0}\min_{R,\ \boldsymbol{\alpha},\ \xi_i} L(R,\ \boldsymbol{\alpha},\ \xi_i, \mu_i, \lambda_i)$$

Then find the partial derivatives:

$$\frac{\partial L}{\partial R} = 2R - \sum_{i=1}^{n}\mu_i 2R = 0 \ \Rightarrow \ \sum_{i=1}^{n}\mu_i = 0$$

$$\nabla_{\boldsymbol{\alpha}} L = -2\sum_{i=1}^{n}\mu_i(\boldsymbol{x}_i-\boldsymbol{\alpha}) = 0 \ \Rightarrow \ \sum_{i=1}^{n}\mu_i(\boldsymbol{x}_i-\boldsymbol{\alpha}) = \sum_{i=1}^{n}\mu_i\boldsymbol{x}_i - \sum_{i=1}^{n}\mu_i\boldsymbol{\alpha} = \sum_{i=1}^{n}\mu_i\boldsymbol{x}_i - \boldsymbol{\alpha} = 0$$

$$\frac{\partial L}{\partial \xi_i} = C - \mu_i - \lambda_i = 0$$

Set $\boldsymbol{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_n]$, and then note that

$$\boldsymbol{x}_i - \boldsymbol{\alpha} = \boldsymbol{x}_i - \sum_{j=1}^{n}\mu_j\boldsymbol{x}_j = \boldsymbol{x}_i - \boldsymbol{X}\boldsymbol{\mu}$$

Plug these values back to the dual problem:

$$\max_{\mu_i,\lambda_i\ge 0}\sum_{i=1}^{n}\mu_i(\boldsymbol{x}_i-\boldsymbol{X}\boldsymbol{\mu})^\top(\boldsymbol{x}_i-\boldsymbol{X}\boldsymbol{\mu})$$

which is equivalent to:

$$\min_{\mu_i \geq 0} \sum_{i=1}^{n} -\mu_i (\boldsymbol{x}_i - \boldsymbol{X}\boldsymbol{\mu})^\top (\boldsymbol{x}_i - \boldsymbol{X}\boldsymbol{\mu})$$

$$\text{s.t } \mu_i \leq C, \sum_{i=1}^{n} \mu_i = 1$$

$$\begin{aligned}
\sum_{i=1}^{n} \mu_i (\boldsymbol{x}_i - \boldsymbol{X}\boldsymbol{\mu})^\top (\boldsymbol{x}_i - \boldsymbol{X}\boldsymbol{\mu}) &= \sum_{i=1}^{n} \mu_i (\boldsymbol{x}_i^\top \boldsymbol{x}_i - 2\boldsymbol{x}_i^\top \boldsymbol{X}^\top \boldsymbol{\mu} + \boldsymbol{\mu}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\mu}) \\
&= \sum_{i=1}^{n} \mu_i \boldsymbol{x}_i^\top \boldsymbol{x}_i - 2\sum_{i=1}^{n} \mu_i \boldsymbol{x}_i^\top \boldsymbol{X}^\top \boldsymbol{\mu} + \sum_{i=1}^{n} \mu_i \boldsymbol{\mu}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\mu} \\
&= \sum_{i=1}^{n} \mu_i \boldsymbol{x}_i^\top \boldsymbol{x}_i - 2\boldsymbol{\mu}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\mu} + \boldsymbol{\mu}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\mu} \sum_{i=1}^{n} \mu_i \\
&= \sum_{i=1}^{n} \mu_i \boldsymbol{x}_i^\top \boldsymbol{x}_i - \boldsymbol{\mu}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\mu}
\end{aligned}$$

The problem is now

$$\min_{\boldsymbol{\mu}} \sum_{i=1}^{n} -\mu_i \boldsymbol{x}_i^\top \boldsymbol{x}_i + \boldsymbol{\mu}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\mu} \tag{3}$$

$$\text{s.t } 0 \leq \mu_i \leq C, \sum_{i=1}^{n} \mu_i = 1$$

Transform to the formula for quadprog():

$$\min_{\boldsymbol{g}} \boldsymbol{f}^\top \boldsymbol{g} + \frac{1}{2}\boldsymbol{g}^\top \boldsymbol{H}\boldsymbol{g}$$

$$\text{s.t } \boldsymbol{A}\boldsymbol{g} \leq \boldsymbol{c},\ \boldsymbol{A}_e\boldsymbol{g} = \boldsymbol{c}_e,\ \boldsymbol{g}_l \leq \boldsymbol{g} \leq \boldsymbol{g}_u$$

where $\boldsymbol{g} = \boldsymbol{\mu},\ \boldsymbol{f} = -\boldsymbol{diag}(\boldsymbol{X}^\top \boldsymbol{X}),\ \boldsymbol{H} = 2\boldsymbol{X}^\top \boldsymbol{X},\ \boldsymbol{A} = 0,\ \boldsymbol{c} = 0,\ \boldsymbol{A}_e = \mathbf{1}^\top, \boldsymbol{g}_l = \mathbf{0}, \boldsymbol{g}_u = \mathbf{1} \cdot C$

Plug back to Lagrangian (2), and get

$$\boldsymbol{\alpha} = \sum_{i=1}^{n} \mu_i \boldsymbol{x}_i$$

Now we need to find optimal R and ξ_i's. Note that when R^2 is larger than the square of distance from the $\lceil \frac{1}{C} \rceil$-th farthest data point to α, the decrease of R^2 is more significant than the increase of $C\sum_{i=1}^{n} \xi_i$ (one unit decrease in R^2 causes $C \cdot \lfloor \frac{1}{C} \rfloor$ unit increase in $C\sum_{i=1}^{n} \xi_i$), and thus

$$SV = \{\boldsymbol{x}_i's \ with \ largest \ \lceil 1/C \rceil \ \mu_i's\}, \text{ the set of support vectors}$$

$$R = \sqrt{\min_{\boldsymbol{x}_i \in SV} (\boldsymbol{x}_i - \boldsymbol{\alpha})^\top (\boldsymbol{x}_i - \boldsymbol{\alpha})}$$

And the centre and radius are found.

2.2

Similar to **2.1**. Replace $\boldsymbol{x}_i$ with $\phi(\boldsymbol{x}_i)$, and $\boldsymbol{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_n]$ with $\boldsymbol{\Phi} = [\phi(\boldsymbol{x}_1), ..., \phi(\boldsymbol{x}_n)]$ in (3), then set $\boldsymbol{K}_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \phi(\boldsymbol{x}_i)^\top \phi(\boldsymbol{x}_j)$ (note that $\boldsymbol{K}$ is positive definite since kernel k is positive definite), and get

$$\min_{\boldsymbol{\mu}} \sum_{i=1}^{n} -\mu_i \phi(\boldsymbol{x}_i)^\top \phi(\boldsymbol{x}_i) + \boldsymbol{\mu}^\top \boldsymbol{\Phi}^\top \boldsymbol{\Phi} \boldsymbol{\mu} \tag{4}$$

$$= \min_{\boldsymbol{\mu}} \sum_{i=1}^{n} -\mu_i k(\boldsymbol{x}_i,\ \boldsymbol{x}_i) + \boldsymbol{\mu}^\top \boldsymbol{K} \boldsymbol{\mu} \tag{5}$$

$$\text{s.t } 0 \le \mu_i \le C, \sum_{i=1}^{n} \mu_i = 1$$

Transform to the formula for quadprog():

$$\min_{\boldsymbol{g}} \boldsymbol{f}^\top \boldsymbol{g} + \frac{1}{2} \boldsymbol{g}^\top \boldsymbol{H} \boldsymbol{g}$$
$$\text{s.t } \boldsymbol{A}\boldsymbol{g} \le \boldsymbol{c},\ \boldsymbol{A}_e \boldsymbol{g} = \boldsymbol{c}_e,\ \boldsymbol{g}_l \le \boldsymbol{g} \le \boldsymbol{g}_u$$
$$\text{where } \boldsymbol{g} = \boldsymbol{\mu},\ \boldsymbol{f} = -\boldsymbol{diag}(\boldsymbol{K}),\ \boldsymbol{H} = 2\boldsymbol{K},\ \boldsymbol{A} = 0,\ \boldsymbol{c} = 0,\ \boldsymbol{A}_e = \boldsymbol{1}^\top, \boldsymbol{g}_l = \boldsymbol{0}, \boldsymbol{g}_u = \boldsymbol{1} \cdot C$$

The rest is similar to to **2.1**.

$$\boldsymbol{\alpha} = \sum_{i=1}^{n} \mu_i \phi(\boldsymbol{x}_i)$$
$$SV = \{\boldsymbol{x}_i' s \ with \ largest \ \lceil 1/C \rceil \ \mu_i' s\}, \text{ the set of support vectors}$$
$$R = \sqrt{\min_{\boldsymbol{x}_i \in SV} (\phi(\boldsymbol{x}_i) - \boldsymbol{\alpha})^\top (\phi(\boldsymbol{x}_i) - \boldsymbol{\alpha})}$$

2.3

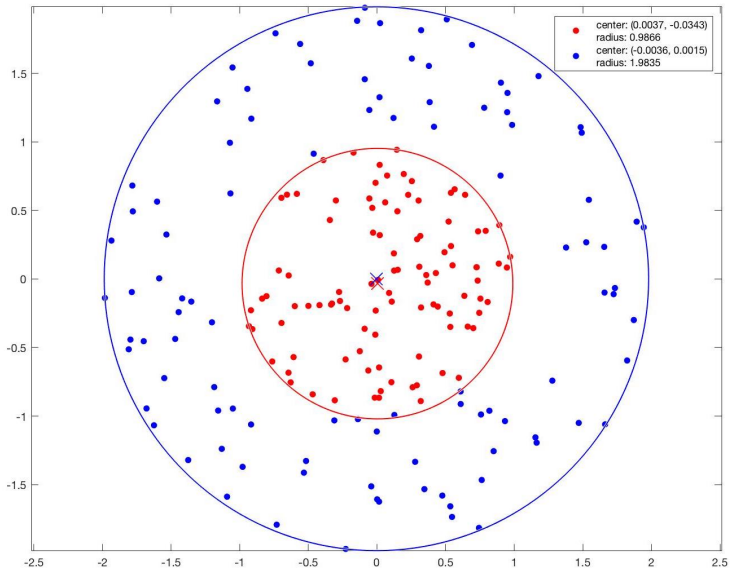

Figure 4: SVM, C=1

With $C = 1$ (no allowance of error):
Class 1: Centred at$(0.0037, -0.0343)$ with radius 0.9866
Class 2: Centred at$(-0.0036, 0.0015)$ with radius 1.9835

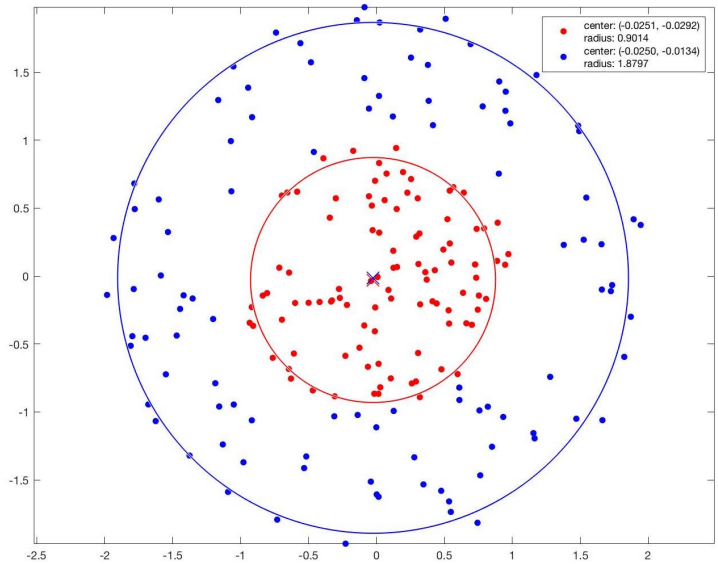

Figure 5: SVM, C=0.05

With $C = 0.05$:
Class 1: Centred at$(-0.0251, -0.0292)$ with radius 0.9014
Class 2: Centred at$(-0.0250, -0.0134)$ with radius 1.8797