ML Homework 5

Qi Luo
A02274095

November 11, 2019

1 SVR

1 (a). Since $y_i - (w^T x_i + b) \le \varepsilon + \xi_i^+ \; \forall i$ and $\xi_i^+ \ge 0 \; \forall i$,

then $\xi_i^+ \ge \max\{0,\; y_i - (w^T x_i + b) - \varepsilon\}$

Also, since $w^T x_i + b - y_i \le \varepsilon + \xi_i^- \; \forall i$ and $\xi_i^- \ge 0 \; \forall i$,

then $\xi_i^- \ge \max\{0,\; -y_i + (w^T x_i + b) - \varepsilon\}$

Therefore, we can get $\xi_i^+ + \xi_i^- \ge \max\{0,\; |y_i - (w^T x_i + b)| - \varepsilon\}$

Since, the objective function is minimizing,

$$\xi_i^+ + \xi_i^- = \max\{0,\; |y_i - (w^T x_i + b)| - \varepsilon\} = \ell_\varepsilon(y_i, w^T x_i + b)$$

Then,
$$\frac{1}{2}\|w\|^2 + \frac{c}{n}\sum_{i=1}^{n}(\xi_i^+ + \xi_i^-)$$
$$= \frac{1}{2}\|w\|^2 + \frac{c}{n}\sum_{i=1}^{n}\ell_\varepsilon(y_i, w^T x_i + b)$$
$$= C\left(\frac{1}{2C}\|w\|^2 + \frac{1}{n}\sum_{i=1}^{n}\ell_\varepsilon(y_i, w^T x_i + b)\right).$$

Therefore, the appropriate choice of λ is $\frac{1}{2C}$, SVR solves

$$\min_{w,b} \frac{1}{n}\sum_{i=1}^{n}\ell_\varepsilon(y_i, w^T x_i + b) + \lambda\|w\|^2 .$$

(b)
$$\min_{w,b,\xi^+,\xi^-} \frac{1}{2}\|w\|^2 + \frac{c}{n}\sum_{i=1}^{n}(\xi_i^+ + \xi_i^-)$$
s.t.
$$y_i - w^T x_i - b - \varepsilon - \xi_i^+ \le 0 \quad \forall i$$
$$w^T x_i + b - y_i - \varepsilon - \xi_i^- \le 0 \quad \forall i$$
$$-\xi_i^+ \le 0 \quad \forall i$$
$$-\xi_i^- \le 0 \quad \forall i$$

$$L(w, b, \xi^+, \xi^-, \alpha, \beta, \lambda, \nu) = \frac{1}{2}\|w\|^2 + \frac{c}{n}\sum_{i=1}^{n}(\xi_i^+ + \xi_i^-) + \sum_{i=1}^{n}\alpha_i(y_i - w^T x_i - b - \varepsilon - \xi_i^+) + \sum_{i=1}^{n}\beta_i(w^T x_i + b - y_i - \varepsilon - \xi_i^-) - \sum_{i=1}^{n}\lambda_i\xi_i^+ - \sum_{i=1}^{n}\nu_i\xi_i^-$$

Then $L_D(\alpha, \beta, \lambda, \nu) = \min_{w,b,\xi^+,\xi^-} L(w, b, \xi^+, \xi^-, \alpha, \beta, \lambda, \nu)$

$$\Rightarrow \max_{\alpha,\beta,\lambda,\nu} -\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i - \beta_i)(\alpha_j - \beta_j)\langle x_i, x_j\rangle + \sum_{i=1}^{n}(\alpha_i - \beta_i)y_i - \sum_{i=1}^{n}(\alpha_i + \beta_i)\varepsilon$$

s.t. $\alpha_i \ge 0,\; \beta_i \ge 0,\; \lambda_i \ge 0,\; \nu_i \ge 0 \quad \forall i$

$$\sum_{i=1}^{n}\alpha_i = \sum_{i=1}^{n}\beta_i$$

$$\frac{c}{n} - \alpha_i + \lambda_i = 0 \quad \Rightarrow \quad 0 \le \alpha_i \le \frac{c}{n}$$
$$\frac{c}{n} - \beta_i - \nu_i = 0 \quad\quad\quad 0 \le \beta_i \le \frac{c}{n}$$

(c) Let α^*, β^* be optimal dual variables.

By KTT condiction, we can get $w^* = \sum_{i=1}^n (\alpha_i^* - \beta_i^*) x_i$

where $0 < \alpha_i^* < \frac{c}{n}$ and $0 < \beta_i^* < \frac{c}{n}$ and $\sum_{i=1}^n \alpha_i = \sum_{i=1}^n \beta_i \;\; \forall i$

Also, we can get
$$y_i - w^{*T} x_i - b^* = \varepsilon + \xi_i^{*+}$$
$$w^{*T} x_i + b^* - y_i = \varepsilon + \xi_i^{*-}$$

Since $\alpha_i^* < \frac{c}{n}$ and $\beta_i^* < \frac{c}{n}$, we can get
$$\frac{c}{n} - \alpha_i^* = \lambda_i^* > 0 \Rightarrow \xi_i^{*+} = 0$$
$$\frac{c}{n} - \beta_i^* = \nu_i^* > 0 \Rightarrow \xi_i^{*-} = 0$$

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^n \alpha_i x_i + \sum_{i=1}^n \beta_i x_i = 0$$
$$\frac{\partial L}{\partial b} = -b\sum_{i=1}^n \alpha_i + b\sum_{i=1}^n \beta_i = 0$$
$$\frac{\partial L}{\partial \xi^+} = \frac{c}{n} - \alpha_i - \lambda_i = 0$$
$$\frac{\partial L}{\partial \xi^-} = \frac{c}{n} - \beta_i - \nu_i = 0$$

Then
$$\begin{aligned}
L_D(\alpha, \beta, \lambda, \nu) &= \frac{1}{2}\Big(\sum_{i=1}^n \alpha_i x_i - \sum_{i=1}^n \beta_i x_i\Big)^T\Big(\sum_{i=1}^n \alpha_i x_i - \sum_{i=1}^n \beta_i x_i\Big) \\
&\quad + \frac{c}{n}\sum_{i=1}^n \xi_i^+ - \sum_{i=1}^n \alpha_i \xi_i^+ - \sum_{i=1}^n \lambda_i \xi_i^+ + \frac{c}{n}\sum_{i=1}^n \xi_i^- - \sum_{i=1}^n \beta_i \xi_i^- \\
&\quad - \sum_{i=1}^n \nu_i \xi_i^- + \sum_{i=1}^n \alpha_i y_i - \sum_{i=1}^n \alpha_i w^T x_i - \sum_{i=1}^n \alpha_i b - \sum_{i=1}^n \alpha_i \varepsilon \\
&\quad - \sum_{i=1}^n \beta_i y_i + \sum_{i=1}^n \beta_i w^T x_i + \sum_{i=1}^n \beta_i b - \sum_{i=1}^n \beta_i \varepsilon \\
&= \frac{1}{2}\Big(\sum_{i=1}^n (\alpha_i - \beta_i) x_i\Big)^T\Big(\sum_{i=1}^n (\alpha_i - \beta_i) x_i\Big) + \sum_{i=1}^n (\alpha_i - \beta_i) y_i \\
&\quad - \sum_{i=1}^n (\alpha_i - \beta_i)\Big(\sum_{i=1}^n (\alpha_i - \beta_i) x_i\Big)^T x_i - \sum_{i=1}^n (\alpha_i + \beta_i)\varepsilon \\
&= \frac{1}{2}\sum_{i,j=1}^n (\alpha_i - \beta_i)(\alpha_j - \beta_j)\langle x_i, x_j\rangle + \sum_{i=1}^n (\alpha_i - \beta_i) y_i \\
&\quad - \sum_{i,j=1}^n (\alpha_i - \beta_i)(\alpha_j - \beta_j)\langle x_i, x_j\rangle - \sum_{i=1}^n (\alpha_i + \beta_i)\varepsilon
\end{aligned}$$

$$\left.\begin{aligned} &\Rightarrow y_i - w^{*T} x_i - b^* = \varepsilon \\ &\quad\;\; w^{*T} x_i + b^* - y_i = \varepsilon \end{aligned}\right\} \Rightarrow \begin{aligned} b^* &= y_i - w^{*T} x_i \\ &= y_i - \sum_{i,j=1}^n (\alpha_i^* - \beta_i^*)\langle x_i, x_j\rangle \end{aligned}$$

For kernelized regression function:
$$\begin{aligned}
f(x) &= \text{sign}\{\langle w^*, x\rangle + b^*\} \\
&= \text{sign}\Big\{\langle \sum_{i,j=1}^n (\alpha_i^* - \beta_i^*)\langle x_i, x_j\rangle + b^*\Big\}
\end{aligned}$$

where $0 < \alpha_i^* < \frac{c}{n}$, $0 < \beta_i^* < \frac{c}{n}$, and $\sum_{i=1}^n \alpha_i^* = \sum_{i=1}^n \beta_i^*$

and $b^* = y_i - \sum_{i,j=1}^n (\alpha_i^* - \beta_i^*)\langle x_i, x_j\rangle$

(d). From complimetary slackness,
$$\alpha_i^*(y_i - w^{*T} x_i - b^* - \varepsilon - \xi_i^{*+}) = 0$$
$$\beta_i^*(w^{*T} x_i + b^* - y_i - \varepsilon - \xi_i^{*-}) = 0$$

For support vector x_i must satisfied $\begin{cases} y_i - w^{*T} x_i - b^* - \varepsilon - \xi_i^{*+} = 0 \\ w^{*T} x_i + b^* - y_i - \varepsilon - \xi_i^{*-} = 0 \end{cases}$

If x_i is not a support vector, then $\alpha_i^* = \beta_i^* = 0$, therefore $w^* = \sum_{i=1}^n (\alpha_i^* - \beta_i^*) x_i$ depends only on a subset of training examples and characterize those training examples.

2 PCA

2. (a) Since the solution for PCA is $u = \bar{X}$, $A = [u_1, \cdots, u_k]$

$\theta_i = A^T(X_i - \bar{X})$ without loss of generality, then we assume $\bar{X} = 0$.

$$\min_{u, A, \{\theta_i\}} \sum_{i=1}^{n} \| X_i - u - A\theta_i \|^2$$

$$= \sum_{i=1}^{n} \| X_i - AA^T(X_i - \bar{X}) \|^2$$

$$= \sum_{i=1}^{n} \| X_i - AA^T X_i \|^2$$

$$= \sum_{i=1}^{n} \| X_i \|^2 - \sum_{i=1}^{n} \| AA^T X_i \|^2$$

$$= \sum_{i=1}^{n} tr(X_i X_i^T) - \sum_{i=1}^{n} tr(AA^T X_i X_i^T)$$

$$= tr\left(\sum_{i=1}^{n} X_i X_i^T\right) - tr\left(\sum_{i=1}^{n} AA^T X_i X_i^T\right)$$

$= n\, tr(S) - n\, tr(AA^T S)$ where S is the sample covariance matrix

$$= n\, tr(S) - n\, tr(A^T S A)$$

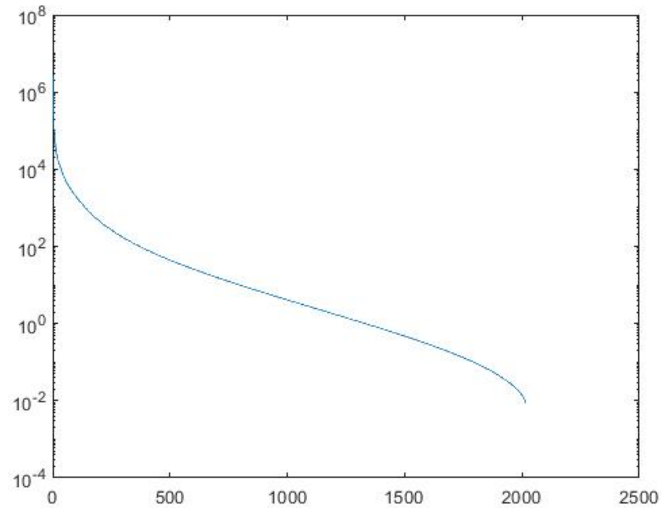
$$= n\, tr(S) - n\, tr\left(\sum_{i=1}^{k} u_i^T S u_i\right)$$

$$= n \sum_{i=k+1}^{d} \lambda_i$$

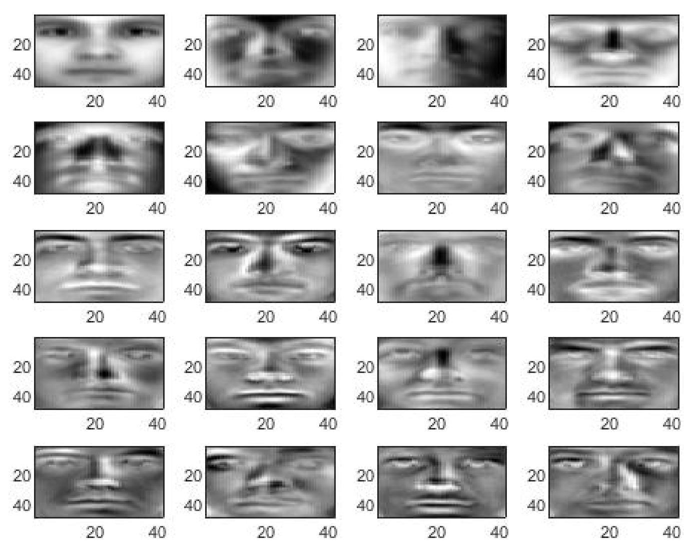
(b). $S = U \Lambda U^T$ Let S be the eigenvectors with the largest eigenvalues to form the $d \times k$ dimention matrix.

3 Eigenfaces

(a) 95% variation captured: 43 and percentage reduction in dimension: 97.87%
99% variation captured: 167 and percentage reduction in dimension: 91.72%



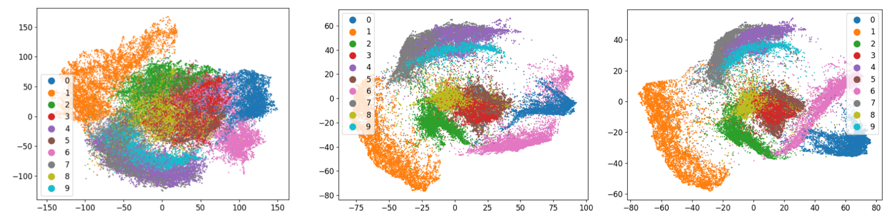
(b) The second principal component is capturing the case in which the lighting is located on the left side of cheeks. The zeroth principal component is capturing nose.



(c) code

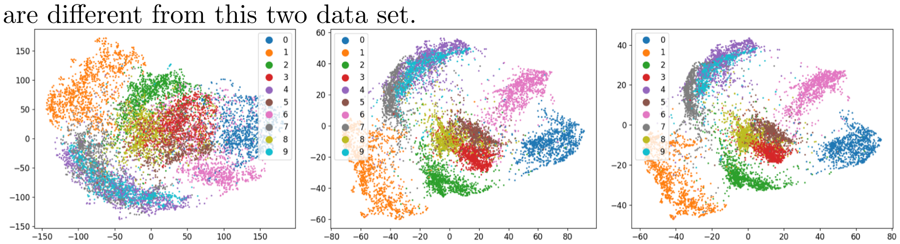
4 PHATE and Clustering

(a) The optimal t is 25 for training data. The optimal t gives a better separation between labels.



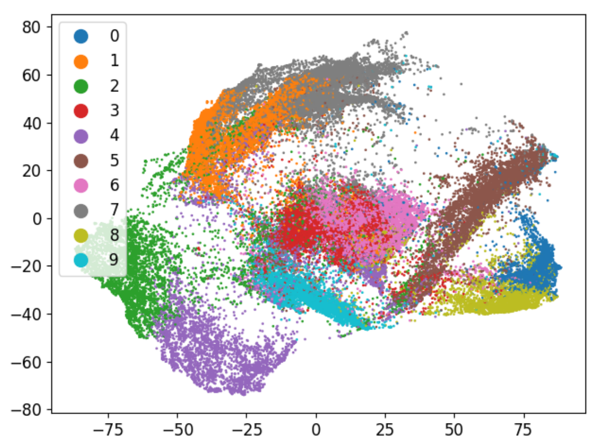
t = 5 | Optimal t = 25 | t = 35

(b) The optimal t is 25 for testing data. For training data and testing data, the separation for t =5 looks very similar, and most likely all labels did not separate well. For t = 25 and t = 35, for training and testing data are looks similar but they both did better on separation. Another difference is the labels locations are different from this two data set.

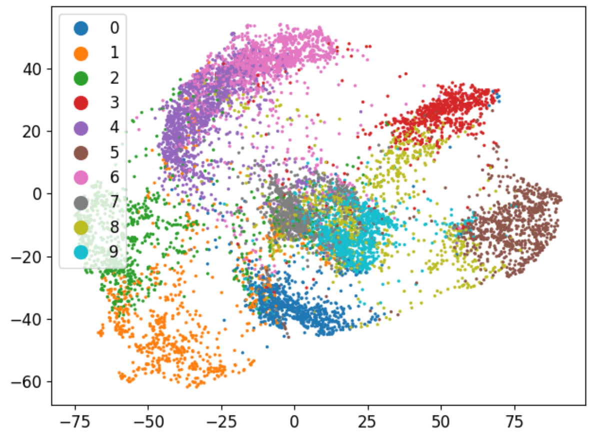


t = 5 | Optimal t = 25 | t = 35

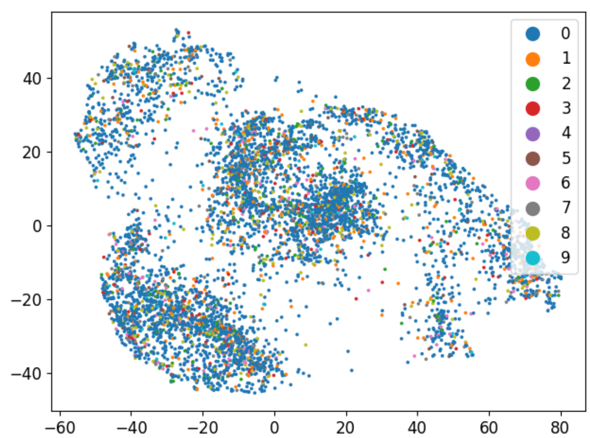
(c) The average ARI for 20 subsampling is 0.361 on training data. I do not think k-means match the true labels well.



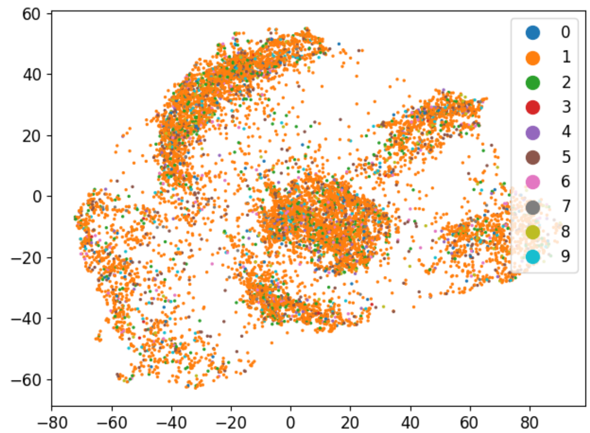
(d) The average ARI for 20 subsampling is 0.381 on training data. The similarity for this two data set is they do not separate labels very well, some labels are layer up. The differences is the labels positions are different from training data and testing data.



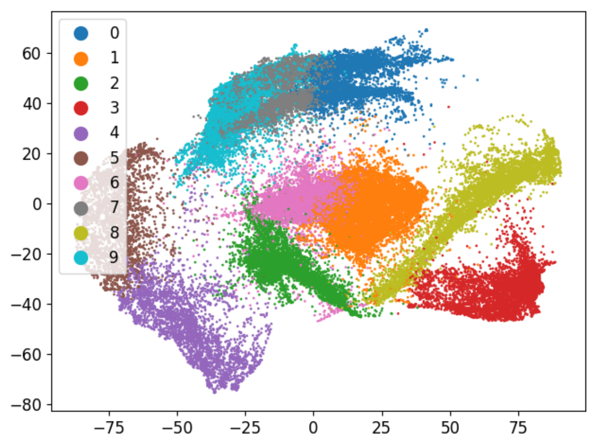
(e) The average ARI for 20 subsampling is 0.00037 on training data when gamma is 1.6. Obviously, k-mean did better than spectral when we use rbf kernel. If we use nearest neighbors, the result will be better than k-means.



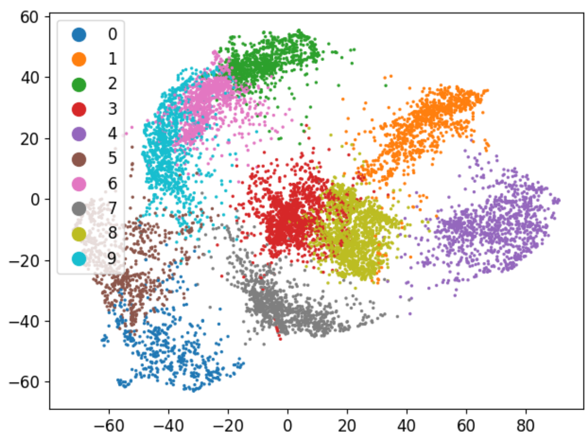
(f) The average ARI for 20 subsampling is 7.815e-05 on testing data.Obviously, k-mean did better than spectral when we use rbf kernel. If we use nearest neighbors, the result will be better than k-means.



(g) The optimal t is 26. ARI is 0.673 on training data. Obviously, this cluster approaches does the best.



(h) The optimal t is 26. ARI is 0.635 on testing data. The similarity is they do separate labels better than other clustering, but the difference is the labels position are not the same.



(i) For unsupervised learning, the most common method is cluster analysis. Maybe we can use neural network that learn the topology and distribution of the data and tune the bandwidth based on the accuracy.
(j) code

5 Ncut and Normalized Spectral Clustering

5. $K=2$. $\quad Ncut(A,\bar{A}) = \frac{1}{2}\sum_{k=1}^{k} \frac{C(A_k,\bar{A}_k)}{vol(A_k)} = \frac{1}{2}\left[\frac{C(A,\bar{A})}{vol(A)} + \frac{C(\bar{A},A)}{vol(\bar{A})}\right]$

$A \subseteq \{1,\cdots,n\}$, defined $f_A = (f_{A1},\cdots,f_{An})^T \in \mathbb{R}^n$ by $\qquad vol(A) = \sum_{i\in A}\sum_{j\in V} w_{ij}$

$$f_{Ai} = \begin{cases} +\sqrt{\frac{|\bar{A}|}{|A|}} & : i \in A \\ -\sqrt{\frac{|A|}{|\bar{A}|}} & : i \notin A \end{cases} \qquad vol(\bar{A}) = \sum_{i\in \bar{A}}\sum_{j\in V} w_{ij}$$

$$f_A^T L f_A = \cancel{RatioCut(A,\bar{A})} = \frac{1}{2}\sum_{i,j} w_{ij}(f_{Ai} - f_{Aj})^2$$
$$= \frac{1}{2}\sum_{i\in A, j\in\bar{A}} w_{ij}\left(\sqrt{\frac{vol(\bar{A})}{vol(A)}} + \sqrt{\frac{vol(A)}{vol(\bar{A})}}\right)^2 + \frac{1}{2}\sum_{i\in\bar{A}, j\in A} w_{ij}\left(-\sqrt{\frac{vol(A)}{vol(\bar{A})}} - \sqrt{\frac{vol(\bar{A})}{vol(A)}}\right)^2$$
$$= \frac{1}{2}C(A,\bar{A})\left[\frac{vol(\bar{A})}{vol(A)} + \frac{vol(A)}{vol(\bar{A})} + 2\right] + \frac{1}{2}C(\bar{A},A)\left[\frac{vol(A)}{vol(\bar{A})} + \frac{vol(\bar{A})}{vol(A)} + 2\right]$$
$$= C(A,\bar{A})\left[\frac{vol(\bar{A}) + vol(A)}{vol(A)} + \frac{vol(A) + vol(\bar{A})}{vol(\bar{A})}\right]$$
$$= (vol(\bar{A}) + vol(A))\left[\frac{C(A,\bar{A})}{vol(A)} + \frac{C(A,\bar{A})}{vol(\bar{A})}\right]$$
$$= 2(vol(\bar{A}) + vol(A))\,Ncut(A,\bar{A})$$

$$\mathbb{1}^T D f_A = \sum_{i=1}^{n} d_i f_{Ai} = \sum_{i\in A} d_i\sqrt{\frac{vol(\bar{A})}{vol(A)}} - \sum_{i\in\bar{A}} d_i\sqrt{\frac{vol(A)}{vol(\bar{A})}}$$

$$\sum_{i\in A} d_i = \sum_{i\in A}\sum_{j=1}^{n} w_{ij} = vol(A)$$

$$\sum_{i\in\bar{A}} d_i = \sum_{i\in\bar{A}}\sum_{j=1}^{n} w_{ij} = vol(\bar{A})$$

Then, $\mathbb{1}^T D f_A = vol(A)\cdot\sqrt{\frac{vol(\bar{A})}{vol(A)}} - vol(\bar{A})\sqrt{\frac{vol(A)}{vol(\bar{A})}} = 0$

$$f_A^T D f_A = \sum_{i=1}^{n} d_i f_{Ai}^2 = \sum_{i\in A} d_i\frac{vol(\bar{A})}{vol(A)} + \sum_{i\in\bar{A}} d_i\frac{vol(A)}{vol(\bar{A})} = vol(A)\cdot\frac{vol(\bar{A})}{vol(A)} + vol(\bar{A})\frac{vol(A)}{vol(\bar{A})} = vol(\bar{A}) + vol(A)$$

Therefore, Ncut can be written as
$$\min_{A\subset\{1,\cdots,n\}} f_A^T L f_A \quad \text{s.t.} \quad \mathbb{1}^T D f_A = 0, \quad f_A^T D f_A = vol(\bar{A}) + vol(A).$$

A relaxation of Ncut is
$$\min_{f\in\mathbb{R}^n} f^T L f \quad \text{s.t.} \quad \mathbb{1}^T D f = 0, \quad f^T D f = vol(\bar{A}) + vol(A).$$

Since the normalized graph Laplacian $\tilde{L} := D^{-1}L$, then the relaxation can be written as
$$\min f^T(D\tilde{L})f \quad \text{s.t.} \quad \mathbb{1}^T D f = 0, \quad f^T D f = vol(\bar{A}) + vol(A).$$

Then let $g = D^{\frac{1}{2}} f$, relaxation can be rewrite as:
$$\min g^T \tilde{L} g \quad \text{s.t.} \quad \mathbb{1}^T D^{\frac{1}{2}} g = 0, \quad g^T g = vol(\bar{A}) + vol(A)$$

Transform back to f: $\quad \min f^T \tilde{L} f \quad \text{s.t.} \quad \mathbb{1}^T D^{\frac{1}{2}} f = 0, \quad f^T f = vol(\bar{A}) + vol(A) \Rightarrow f = \sqrt{vol(\bar{A}) + vol(A)}$