COURSEWORK

IMPERIAL COLLEGE LONDON

DEPARTMENT OF COMPUTING

Mathematics for Machine Learning

Author:
James Dorricott (CID: 01009425)
jd3114@ic.ac.uk

Date: December 2, 2019

1

Figures 1, 2 and 3 show plots of the error rate versus number of components for three dimensionality reduction algorithms: PCA, whitened PCA and LDA, respectively. All three techniques show an exponentially decreasing error as the number of components increases. However, the best would appear to be LDA, where the error rate reaches a minimum of 25% compared with 30% for whitened PCA and 75% for PCA. Moreover, for whitened PCA, the error rate starts to increase as the number of components approaches the original dimensions of the problem.

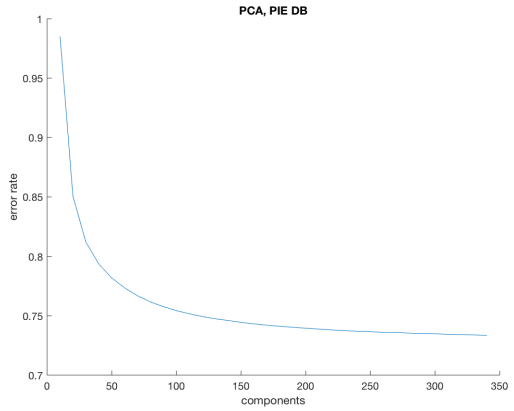

Figure 1: PCA reconstruction error rate as a function of the number of components.

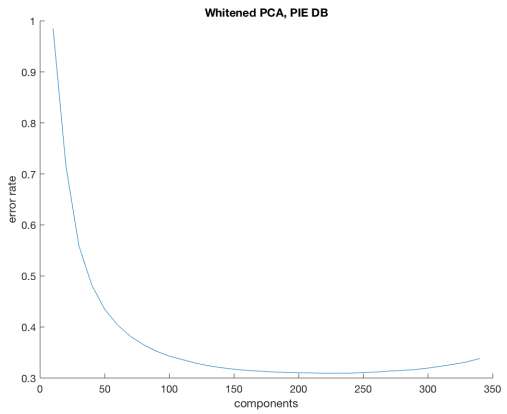

Figure 2: Whitened PCA reconstruction error rate as a function of the number of components.

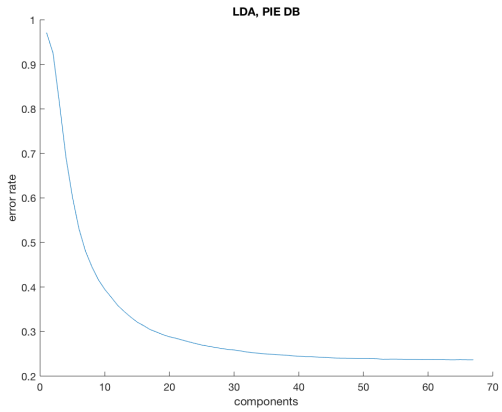

Figure 3: LDA reconstruction error rate as a function of the number of components.

2

2.1

The primal Lagrangian for the soft margin SVM optimisation problem stated in the specification is formulated as

$$L(\boldsymbol{w}, b, \xi, a, r) = \frac{1}{2}\boldsymbol{w}^T \boldsymbol{S}_t \boldsymbol{w} + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N} a_i(y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) - 1 + \xi_i) - \sum_{i=1}^{N} r_i\xi_i \tag{1}$$

where $a_i \geq 0$ and $r_i \geq 0$ are Lagrange multipliers corresponding to the constraints

$$y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 - \xi_i \tag{2}$$

$$\xi_i \geq 0. \tag{3}$$

Differentiating the Lagrangian with respect to the three primal variables $\boldsymbol{w}$, b and ξ, respectively, yields

$$\frac{\partial L}{\partial \boldsymbol{w}} = \boldsymbol{S}_t\boldsymbol{w} - \sum_{i=1}^{N} a_i y_i \boldsymbol{x}_i \tag{4}$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{N} a_i y_i \tag{5}$$

$$\frac{\partial L}{\partial \xi} = C - a_i - r_i. \tag{6}$$

If we set each partial equal to zero, we derive an equation for the optimal value of $\boldsymbol{w}$

$$\boldsymbol{w}^* = \sum_{i=1}^{N} a_i y_i \boldsymbol{S}_t^{-1}\boldsymbol{x}_i \tag{7}$$

and constraints on the sum of $a_i y_i$ and the values of the Lagrange multipliers a_i

$$\sum_{i=1}^{N} a_i y_i = 0 \tag{8}$$

$$0 \leq a_i \leq C. \tag{9}$$

Plugging equations (7), (8) and (9) back into to the primal optimisation problem, the dual optimisation problem is expressed (in matrix form) as

$$\max_a -\frac{1}{2}\boldsymbol{a}^T \boldsymbol{K}^y \boldsymbol{a} + \mathbf{1}^T \boldsymbol{a} \tag{10}$$

or equivalently

$$\min_a \frac{1}{2}\boldsymbol{a}^T \boldsymbol{K}^y \boldsymbol{a} - \mathbf{1}^T \boldsymbol{a} \tag{11}$$

i.e., only in terms of a_i and subject to constraints (8) and (9). $\boldsymbol{K}^y$ is the matrix containing the inner products between pairs of examples

$$\boldsymbol{K}^y = \boldsymbol{y}\boldsymbol{y}^T \odot \boldsymbol{X}^T \boldsymbol{S}_t^{-1} \boldsymbol{X}. \tag{12}$$

Using the optimal weight vector $\boldsymbol{w}^*$, we obtain the optimal value of b with

$$b^* = \frac{1}{N_s}(y_i - \boldsymbol{w}^* \boldsymbol{x}_i) \tag{13}$$

where N_s is the number of non-zero Lagrange multipliers in $\boldsymbol{a}$.

2.2

For the case when $\boldsymbol{S}_t$ is singular, we could diagonalize it according to

$$\boldsymbol{S}_t = \boldsymbol{U}\Lambda\boldsymbol{U}^T \tag{14}$$

keeping only the $r < (n-1)$ dimensions of $\boldsymbol{S}_t$ (and $\boldsymbol{X}$) corresponding to the non-zero eigenvalues of $\boldsymbol{S}_t$, such that $\boldsymbol{S}_t$ becomes full rank (call it $\boldsymbol{S}_t'$). This can always be performed as $\boldsymbol{S}_t$ is symmetric. We would then carry out the optimisation in this subspace.