Dimensionality Reduction through LDA

- We studied dimensionality reduction through PCA.
- $\boxed{\text{LDA can be used for dimensionality reduction.}}$
  It can be considered as a “supervised PCA”.

- Given $\mathcal{D} = \mathcal{D}_1 \sqcup \mathcal{D}_2 \sqcup \cdots \sqcup \mathcal{D}_s$ (disjoint union),
  set

$$N_t := \#(\mathcal{D}_t), \quad t = 1, 2, \ldots, s, \qquad N := N_1 + \cdots + N_s.$$

- Main Idea: Shrink class $\mathcal{D}_t$ into a single point $\boldsymbol{\mu}_t$ and do PCA.

  $\rightsquigarrow$ Principal directions for classes.

  Choose

  $$\boldsymbol{\mu}_t = \frac{1}{N_t} \sum_{\boldsymbol{x} \in \mathcal{D}_t} \boldsymbol{x} \quad \text{and} \quad \boldsymbol{\mu} = \frac{1}{N} \sum_{\boldsymbol{x} \in \mathcal{D}} \boldsymbol{x}.$$

  However, the point $\boldsymbol{\mu}_t$ must have multiplicity N_t.

- Find a feature vector $\boldsymbol{a}$ which maximizes

  $$\boldsymbol{a}^\top \left(\sum_{t=1}^{s} N_t (\boldsymbol{\mu}_t - \boldsymbol{\mu})(\boldsymbol{\mu}_t - \boldsymbol{\mu})^\top \right) \boldsymbol{a}.$$

- A feature vector $\boldsymbol{a}$ may capture more variability from one class than from the other classes.

  To normalize, we fix the sum of variances. That is,

$$\sum_{t=1}^{s} \boldsymbol{a}^\top \left(\sum_{\boldsymbol{x}\in\mathcal{D}_t} (\boldsymbol{x}-\boldsymbol{\mu}_t)(\boldsymbol{x}-\boldsymbol{\mu}_t)^\top \right) \boldsymbol{a} = 1.$$

- Define

$$B = \sum_{t=1}^{s} N_t(\boldsymbol{\mu}_t - \boldsymbol{\mu})(\boldsymbol{\mu}_t - \boldsymbol{\mu})^\top$$

$$W = \sum_{t=1}^{s} \sum_{\boldsymbol{x} \in \mathcal{D}_t} (\boldsymbol{x} - \boldsymbol{\mu}_t)(\boldsymbol{x} - \boldsymbol{\mu}_t)^\top.$$

- Task: Maximize $\boldsymbol{a}^\top B \boldsymbol{a}$, subject to $\boldsymbol{a}^\top W \boldsymbol{a} = 1$.

- Lagrange Multiplier

$$F = \boldsymbol{a}^\top B \boldsymbol{a} - \lambda(\boldsymbol{a}^\top W \boldsymbol{a} - 1)$$

- We obtain

$$\nabla F = 2B\boldsymbol{a} - 2\lambda W\boldsymbol{a} = 0 \quad \Longleftrightarrow \quad W^{-1}B\boldsymbol{a} = \lambda\boldsymbol{a}$$

  Critical points are eigenvectors of $W^{-1}B$.

  Note that $W^{-1}B$ is symmetric.

- If $\boldsymbol{a}$ is an eigenvector of $W^{-1}B$ such that $\boldsymbol{a}^\top W\boldsymbol{a} = 1$, then

$$\boldsymbol{a}^\top B\boldsymbol{a} = \boldsymbol{a}^\top WW^{-1}B\boldsymbol{a} = \lambda\boldsymbol{a}^\top W\boldsymbol{a} = \lambda.$$

  Thus an eigenvector of the largest eigenvalue is the first principal direction.

- Take k'-many principal directions for $k' < k$.

  Project data points onto the subspace of the principal directions.

  $\rightsquigarrow$ Dimensionality Reduction