CIS 530 Advanced Data Mining: Notes for PCA and LDA

Instructor: Thomas W Gyeera

1 PCA Algorithm Details

Assume we have a dataset $X = \{x_1, x_2, ..., x_n\}, x_i \in R^D$, we would like to learn a linear subspace (represented by w) where the variance of dataset X can be maximized. That is the variance of the low-dimensional representation $Y = \{w^T x_1, w^T x_2, ..., w^T x_n\}$ is maximized. To that end, we first centralize the dataset X.

$$\hat{x}_i = x_i - \mu_x, \text{ and } \mu_x = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{1}$$

where $\hat{x}_i$ is the centralized sample, and μ_x is the mean vector of dataset X. Next, we will find out the variance of Y in the low-dimensional space:

$$Var(Y) = \frac{1}{n}\sum_{i=1}^{n} (y_i - \mu_y)^2, \tag{2}$$

where μ_y is the mean of dataset Y. Let's skip the term $\frac{1}{n}$ as it won't affect our objective function. If we substitutes $w^T x$ for y in Eq. (2), we have the following derivations:

$$\begin{aligned}
\mathrm{Var}(Y) &= \sum_{i=1}^{n} (y_i - \mu_y)^2 = \sum_{i=1}^{n} (w^T x_i - w^T \mu_x)^2 = \sum_{i=1}^{n} w^T (x_i - \mu_x)(x_i - \mu_x)^T w \\
&= w^T \left(\sum_{i=1}^{n} (x_i - \mu_x)(x_i - \mu_x)^T \right) w = w^T \left(\sum_{i=1}^{n} \hat{x}_i \hat{x}_i^T \right) w = w^T \hat{X}\hat{X}^T w,
\end{aligned} \tag{3}$$

where $\hat{X} = [\hat{x}_1, \hat{x}_2, ..., \hat{x}_n] \in R^{D\times n}$.

Therefore, maximizing $\mathrm{Var}(Y)$ is equal to maximizing $w^T \hat{X}\hat{X}^T w$. However, the magnitude of w will affect the variance because when $w \to \infty$, $w^T \hat{X}\hat{X}^T w \to \infty$. So we need an additional constraint for the learning objective, i.e., $\|w\|_2^2 = 1$. By introducing this constraint, the objective function of PCA can be formulated as:

$$\max_w w^T \hat{X}\hat{X}^T w \text{ subject to } \|w\|_2^2 = 1, \tag{4}$$

which is a constrained quadratic optimization problem with the global solution. Here we use Lagrangian Multiplier Method for solutions. We first introduce the multiplier λ and then convert the original problem to an unconstrained optimization problem:

$$\max_w w^T \hat{X}\hat{X}^T w + \lambda(\|w\|_2^2 - 1). \tag{5}$$

By setting the its first-order derivative w.r.t w to 0, we have the following equivalent eigen-decomposition problem:

$$\hat{X}\hat{X}^T w = \lambda w, \tag{6}$$

where the eigenvector w is the solution for PCA. Usually we have more than one solutions for w. If we have d eigenvectors, we can reduce the dimensionality of x from D to d.

2 LDA Algorithm Details

In LDA, we would like to find a subspace w where the data from the same class are close to each other while data from different classes are distant. For the first target, we need to minimize the within-class scatter, while for the second target, we need to maximize the between-class scatter.

2.1 Within-Class Scatter

Assume the we have c classes of data $X = \{X_1, X_2, ..., X_c\}$, $X_i \in R^{D\times n_i}$, where n_i is the number of samples in class i. For class i, in the learned subspace w, we will minimize the following value:

$$\begin{aligned}
\sum_{x_j\in X_i} (w^T(x_j - \mu_i))^2 &= \sum_{x_j\in X_i} w^T(x_j - \mu_i)(x_j - \mu_i)^T w^T = w^T \left(\sum_{x_j\in X_i} (x_j - \mu_i)(x_j - \mu_i)^T \right) w = w^T S_i w, \\
\text{where } S_i &= \sum_{x_j\in X_i} (x_j - \mu_i)(x_j - \mu_i)^T
\end{aligned} \tag{7}$$

is the within-class scatter matrix for class i, and μ_i is the center of X_i.

Then, the within-class scatter matrix for c classes can be computed by:

$$S_w = \sum_{i=1}^{c} S_i = \sum_{i=1}^{c} \sum_{x_j\in X_i} (x_j - \mu_i)(x_j - \mu_i)^T. \tag{8}$$

2.2 Between-Class Scatter

For between-class scatter, we will maximize distance between the center of each class, and the center of all data. Assume the center of all data is μ, then in the subspace w, we will maximize the following value:

$$\begin{aligned}
\sum_{i=1}^{c} n_i(w^T(\mu_i - \mu))^2 &= \sum_{i=1}^{c} n_i w^T(\mu_i - \mu)(\mu_i - \mu)^T w = w^T \left(\sum_{i=1}^{c} n_i(\mu_i - \mu)(\mu_i - \mu)^T \right) w = w^T S_b w \\
\text{where } S_b &= \sum_{i=1}^{c} n_i(\mu_i - \mu)(\mu_i - \mu)^T.
\end{aligned} \tag{9}$$

2.3 Fisher Criterion

According to the Fisher Criterion, we will jointly optimize the Eq. (8) and (9) to learn the subspace w by the following objective function:

$$\max_w \frac{w^T S_b w}{w^T S_w w}, \tag{10}$$

which is usually converted to the following equivalent problem:

$$\max_w w^T S_b w \text{ subject to } w^T S_w w = 1. \tag{11}$$

Similar to the solutions of PCA, we still use the Lagrangian Multiplier Method to solve the constrained optimization problem. And this is equal to solving the following eigen-decomposition problem:

$$S_w^{-1} S_b w = \lambda w, \tag{12}$$

where the eigenvector is the intended subspace w. Note that typically we can only obtain $c-1$ eigenvectors according to the definitions of within- and between-class scatter matrices.