Derivación del problema de optimización de PCA

Percepción - ETSInf

Equivalencia del error de reconstrucción

La reconstrucción de la muestra de entrenamiento $\hat{\mathbf{x}}_n$ se formula como:

$$\hat{\mathbf{x}}_n = \overline{\mathbf{x}} + WW^t(\mathbf{x}_n - \overline{\mathbf{x}}) \tag{1}$$

donde $W \in \mathbb{R}^{D\times k}$ es la matriz proyección a k dimensiones, $(\mathbf{x}_n - \overline{\mathbf{x}}) \in \mathbb{R}^{D\times 1}$, y $\hat{\mathbf{x}}_n \in \mathbb{R}^{D\times 1}$. Teniendo en cuenta que $W = (\mathbf{w}_1\mathbf{w}_2 \dots \mathbf{w}_k)$ es ortonormal ($\mathbf{w}_i \cdot \mathbf{w}_j = 0$ si $i \neq j$ y $\mathbf{w}_j\mathbf{w}_j = 1$), tendremos:

$$W \cdot W^t(\mathbf{x}_n - \overline{\mathbf{x}}) = \sum_{j=1}^{k} \mathbf{w}_j^t(\mathbf{x}_n - \overline{\mathbf{x}})\mathbf{w}_j$$

siendo $\mathbf{w}_j \in \mathbb{R}^{D\times 1}$ y $(\mathbf{x}_n - \overline{\mathbf{x}}) \in \mathbb{R}^{D\times 1}$, el resultado de ese primer producto es un escalar resultado de proyectar $(\mathbf{x}_n - \overline{\mathbf{x}})$ por $\mathbf{w}_j^t$. Al escalar resultante le llamaremos z_j^n, y es el valor de la dimensión $j \in \{1, \cdots, k\}$ de la proyección de $\mathbf{x}_n$. Por tanto, reescribimos Ec. 1 como

$$\hat{\mathbf{x}}_n = \overline{\mathbf{x}} + \sum_{j=1}^{k} z_j^n \mathbf{w}_j$$

donde se suma el producto de z_j^n por $\mathbf{w}_j$ para todo $j \in \{1, \cdots, k\}$, y finalmente se suma $\overline{\mathbf{x}}$.

Al ser W ortonormal, cualquier $\mathbf{x}_n$ puede ser reconstruido *sin error* con D vectores $\mathbf{w}_j$. Es decir, tomando $k = D$ tendremos que $W \in \mathbb{R}^{D\times D}$ y por tanto

$$\mathbf{x}_n = \hat{\mathbf{x}}_n = \overline{\mathbf{x}} + \sum_{j=1}^{D} z_j^n \mathbf{w}_j^t$$

Es decir, los vectores reconstruidos son los originales[1]. Esto se puede conseguir, por ejemplo, con $W = I$ (matriz identidad).

El error de reconstucción para $k < D$ dimensiones puede calcularse como

$$\text{error}_k = \sum_{n=1}^{N} \|\mathbf{x}_n - \hat{\mathbf{x}}_n\|^2$$

[1] Realmente lo que se obtiene son los vectores originales expresados en los ejes de coordenadas definidos por la matriz de proyección W completa.

Teniendo en cuenta que $\mathbf{x}_n = \overline{\mathbf{x}} + \sum_{j=1}^{D} z_j^n \mathbf{w}_j$ y $\hat{\mathbf{x}}_n = \overline{\mathbf{x}} + \sum_{j=1}^{k} z_j^n \mathbf{w}_j$

$$\begin{aligned}
\mathrm{error}_k &= \sum_{n=1}^{N} \|\mathbf{x}_n - \hat{\mathbf{x}}_n\|^2 \\
&= \sum_{n=1}^{N} \left\| \left(\overline{\mathbf{x}} + \sum_{j=1}^{D} z_j^n \mathbf{w}_j \right) - \left(\overline{\mathbf{x}} + \sum_{j=1}^{k} z_j^n \mathbf{w}_j \right) \right\|^2 \\
&= \sum_{n=1}^{N} \left\| \sum_{j=k+1}^{D} z_j^n \mathbf{w}_j \right\|^2 \\
&= \sum_{n=1}^{N} \left(\sum_{j=k+1}^{D} z_j^n \mathbf{w}_j \right)^t \left(\sum_{j=k+1}^{D} z_j^n \mathbf{w}_j \right) \\
&= \sum_{n=1}^{N} \left(\sum_{j=k+1}^{D} z_j^n \mathbf{w}_j^t \right) \left(\sum_{j=k+1}^{D} z_j^n \mathbf{w}_j \right)
\end{aligned}$$

Desarrollando el sumatorio

$$\mathrm{error}_k = \sum_{n=1}^{N} \left(z_{k+1}^n \mathbf{w}_{k+1}^t + \cdots + z_D^n \mathbf{w}_D^t \right) \left(z_{k+1}^n \mathbf{w}_{k+1} + \cdots + z_D^n \mathbf{w}_D \right)$$

Aplicando la propiedad distributiva y usando la ortonormalidad tendremos:

$$\begin{aligned}
\mathrm{error}_k = \sum_{n=1}^{N} & \left(z_{k+1}^n \mathbf{w}_{k+1}^t \right) \left(z_{k+1}^n \mathbf{w}_{k+1} + z_{k+2}^n \mathbf{w}_{k+2} + \ldots + z_D^n \mathbf{w}_D \right) + \\
& \left(z_{k+2}^n \mathbf{w}_{k+2}^t \right) \left(z_{k+1}^n \mathbf{w}_{k+1} + z_{k+2}^n \mathbf{w}_{k+2} + \ldots + z_D^n \mathbf{w}_D \right) + \ldots + \\
& \left(z_D^n \mathbf{w}_D^t \right) \left(z_{k+1}^n \mathbf{w}_{k+1} + z_{k+2}^n \mathbf{w}_{k+2} + \ldots + z_D^n \mathbf{w}_D \right)
\end{aligned}$$

$$= \sum_{n=1}^{N} \left(z_{k+1}^{n}\right)^2 \mathbf{w}_{k+1}^{t}\mathbf{w}_{k+1} + z_{k+1}^{n} z_{k+2}^{n} \mathbf{w}_{k+1}^{t}\mathbf{w}_{k+2} + \ldots + z_{k+1}^{n} z_{D}^{n} \mathbf{w}_{k+1}^{t}\mathbf{w}_{D} +$$

$$z_{k+2}^{n} z_{k+1}^{n} \mathbf{w}_{k+2}^{t}\mathbf{w}_{k+1} + \left(z_{k+2}^{n}\right)^2 \mathbf{w}_{k+2}^{t}\mathbf{w}_{k+2} + \ldots + z_{k+2}^{n} z_{D}^{n} \mathbf{w}_{k+2}^{t}\mathbf{w}_{D} + \ldots +$$

$$z_{D}^{n} z_{k+1}^{n} \mathbf{w}_{D}^{t}\mathbf{w}_{k+1} + z_{D}^{n} z_{k+2}^{n} \mathbf{w}_{D}^{t}\mathbf{w}_{k+2} + \ldots + \left(z_{D}^{n}\right)^2 \mathbf{w}_{D}^{t}\mathbf{w}_{D}$$

$$= \sum_{n=1}^{N} \left(z_{k+1}^{n}\right)^2 + \left(z_{k+2}^{n}\right)^2 + \ldots + \left(z_{D}^{n}\right)^2 = \sum_{n=1}^{N} \sum_{j=k+1}^{D} \left(z_{j}^{n}\right)^2$$

Retomando el valor de $z_j^n = \mathbf{w}_j^t \left(\mathbf{x}_n - \overline{\mathbf{x}}\right)$

$$\begin{aligned} \text{error}_k &= \sum_{n=1}^{N} \sum_{j=k+1}^{D} \left(z_{j}^{n}\right)^2 = \sum_{n=1}^{N} \sum_{j=k+1}^{D} \left(\mathbf{w}_j^t \left(\mathbf{x}_n - \overline{\mathbf{x}}\right)\right)^2 \\ &= \sum_{n=1}^{N} \sum_{j=k+1}^{D} \mathbf{w}_j^t \left(\mathbf{x}_n - \overline{\mathbf{x}}\right)\left(\mathbf{x}_n - \overline{\mathbf{x}}\right)^t \mathbf{w}_j \end{aligned}$$

Intercambiando sumatorios

$$\text{error}_k = \sum_{j=k+1}^{D} \sum_{n=1}^{N} \mathbf{w}_j^t \left(\mathbf{x}_n - \overline{\mathbf{x}}\right)\left(\mathbf{x}_n - \overline{\mathbf{x}}\right)^t \mathbf{w}_j$$

Como las $\mathbf{w}_j$ son independientes de n se pueden sacar del sumatorio interno, con lo cual

$$\text{error}_k = \sum_{j=k+1}^{D} \mathbf{w}_j^t \left(\sum_{n=1}^{N} \left(\mathbf{x}_n - \overline{\mathbf{x}}\right)\left(\mathbf{x}_n - \overline{\mathbf{x}}\right)^t\right) \mathbf{w}_j$$

Dado que la matriz de covarianzas de los datos $\Sigma_{\mathcal{X}} = \frac{1}{N}\sum_{n=1}^{N} \left(\mathbf{x}_n - \overline{\mathbf{x}}\right)\left(\mathbf{x}_n - \overline{\mathbf{x}}\right)^t$, queda finalmente que

$$\text{error}_k = N \sum_{j=k+1}^{D} \mathbf{w}_j^t \Sigma_{\mathcal{X}} \mathbf{w}_j$$

Para la derivación del error, se optimiza respecto a la matriz de proyección W. Por tanto, no hay dependencia del número de datos N, por lo que generalmente usaremos:

$$\text{error}_k = \sum_{j=k+1}^{D} \mathbf{w}_j^t \Sigma_{\mathcal{X}} \mathbf{w}_j$$

Derivada del error de reconstrucción

Nuestra función objetivo a minimizar es el error de reconstrucción $\text{error}_k = \sum_{j=k+1}^{D} \mathbf{w}_j^t \Sigma_{\mathcal{X}} \mathbf{w}_j$ mediante la matriz de proyección W a k dimensiones. Así, el problema de optimización se formula como

$$\widehat{W} = \operatorname*{argmin}_{W \in \mathbb{R}^{D \times k}} \sum_{j=k+1}^{D} \mathbf{w}_j^t \Sigma_{\mathcal{X}} \mathbf{w}_j = \operatorname*{argmax}_{W \in \mathbb{R}^{D \times k}} \sum_{j=1}^{k} \mathbf{w}_j^t \Sigma_{\mathcal{X}} \mathbf{w}_j$$

Simplificando el problema anterior al caso de la proyección del conjunto de entrenamiento a una única dimensión dado por $\mathbf{w}$, esto se puede formular como

$$\hat{\mathbf{w}} = \operatorname*{argmax}_{\mathbf{w} \in \mathbb{R}^D} \ \mathbf{w}^t \Sigma_{\mathcal{X}} \mathbf{w} \quad \text{sujeto a que} \quad \mathbf{w}^t \mathbf{w} = 1$$

Este es un problema de maximización con restricciones que puede formularse a través de los multiplicadores de Lagrange, de forma que queda como

$$\hat{\mathbf{w}} = \operatorname*{argmax}_{\mathbf{w} \in \mathbb{R}^D} \max_{\lambda \in \mathbb{R}} \ \mathbf{w}^t \Sigma_{\mathcal{X}} \mathbf{w} + \lambda \left(1 - \mathbf{w}^t \mathbf{w}\right)$$

Para la optimización se procede a derivar e igualar a cero. Para ello, primero debemos identificar el tipo de función E que vamos a derivar, que en este caso es

$$E = \mathbf{w}^t \Sigma_{\mathcal{X}} \mathbf{w} + \lambda \left(1 - \mathbf{w}^t \mathbf{w}\right)$$

Por análisis de dimensiones se puede observar que ambos términos son escalares. Ahora se debe proceder a la derivación respecto a cada una de las variables de optimización: el vector de proyección $\mathbf{w}$ y el multiplicador de Lagrange λ.

Derivada respecto a λ

La derivada respecto al multiplicador de Lagrange λ, que es un escalar, es trivial. La derivada del primer término es cero y la derivada del segundo resulta en

$$\frac{\partial E}{\partial \lambda} = 1 - \mathbf{w}^t \mathbf{w} = 0 \rightarrow \mathbf{w}^t \mathbf{w} = 1$$

Simplemente nos recuerda que $\mathbf{w}$ debe ser de módulo unitario

Derivada respecto a $\mathbf{w}$

La derivada de la función escalar E respecto al vector $\mathbf{w} = (w_1 \cdots w_D)^t$ es

$$\frac{\partial E}{\partial \mathbf{w}} = \left(\frac{\partial E}{\partial w_1}, \frac{\partial E}{\partial w_2}, \cdots, \frac{\partial E}{\partial w_D} \right)$$

Resulta en un vector fila donde la componente j es la derivada de E respecto a w_j. Por simplicidad, hagamos la derivada respecto a w_1 desarrollando E

$$\frac{\partial E}{\partial w_1} = \frac{\partial\, \mathbf{w}^t \Sigma_{\mathcal{X}} \mathbf{w} + \lambda\,(1-\mathbf{w}^t\mathbf{w})}{\partial w_1} = \frac{\partial \sum_{i=1}^{D}\sum_{j=1}^{D} w_i \Sigma_{ij} w_j}{\partial w_1} + \frac{\partial \lambda \left(1-\sum_{i=1}^{D} w_i^2\right)}{\partial w_1}$$

Haciendo primero la derivada del segundo sumando

$$\begin{aligned}\frac{\partial \lambda \left(1-\sum_{i=1}^{D} w_i^2\right)}{\partial w_1} &= \lambda \frac{\partial \left(1-\sum_{i=1}^{D} w_i^2\right)}{\partial w_1} \\ &= \lambda \frac{\partial \left(1-(w_1^2+w_2^2+\cdots+w_D^2)\right)}{\partial w_1} = \lambda\,(-2w_1) = -2\lambda w_1\end{aligned}$$

En general, tendríamos $\frac{\partial \lambda \left(1-\sum_{i=1}^{D} w_i^2\right)}{\partial w_j} = -2\lambda w_j$, con lo cual, si realizáramos la derivada del segundo término respecto al vector $\mathbf{w}$ completo se obtiene

$$\frac{\partial\, \lambda \left(1-(w_1^2+w_2^2+\cdots+w_D^2)\right)}{\partial \mathbf{w}} = (-2\lambda w_1, -2\lambda w_2, \cdots, -2\lambda w_D) = -2\lambda \mathbf{w}^t$$

Respecto a la derivada del primer sumando

$$\begin{aligned}&\frac{\partial \sum_{i=1}^{D}\sum_{j=1}^{D} w_i \Sigma_{ij} w_j}{\partial w_1} = \\ &= \frac{\partial \left(\sum_{j=1}^{D} w_1 \Sigma_{1j} w_j + \sum_{j=1}^{D} w_2 \Sigma_{2j} w_j + \ldots + \sum_{j=1}^{D} w_D \Sigma_{Dj} w_j\right)}{\partial w_1} \\ &= \frac{\partial\,(w_1\Sigma_{11}w_1 + w_1\Sigma_{12}w_2 + \ldots + w_1\Sigma_{1D}w_D)}{\partial w_1} + \\ &\quad \frac{\partial\,(w_2\Sigma_{21}w_1 + w_2\Sigma_{22}w_2 + \ldots + w_2\Sigma_{2D}w_D)}{\partial w_1} + \ldots + \\ &\quad \frac{\partial\,(w_D\Sigma_{D1}w_1 + w_D\Sigma_{D2}w_2 + \ldots + w_D\Sigma_{DD}w_D)}{\partial w_1} \\ &= (2w_1\Sigma_{11} + \Sigma_{12}w_2 + \ldots + \Sigma_{1D}w_D) + w_2\Sigma_{21} + \ldots + w_D\Sigma_{D1} \\ &= (\Sigma_{11}w_1 + \Sigma_{12}w_2 + \ldots + \Sigma_{1D}w_D) + (w_1\Sigma_{11} + w_2\Sigma_{21} + \ldots + w_D\Sigma_{D1}) \\ &= \sum_{j=1}^{D} \Sigma_{1j} w_j + \sum_{i=1}^{D} w_i \Sigma_{i1}\end{aligned}$$

Teniendo en cuenta que $\Sigma_{\mathcal{X}}$ es simétrica y, por tanto, $\Sigma_{ij} = \Sigma_{ji}$, queda como

$$2\sum_{i=1}^{D} w_i \Sigma_{i1} = 2\mathbf{w}^t \Sigma_{\bullet 1}$$

Donde $\Sigma_{\bullet i}$ indica la i-ésima columna de $\Sigma_{\mathcal{X}}$. Por tanto, en general para cualquier w_k tendríamos que

$$\frac{\partial \sum_{i=1}^{D} \sum_{j=1}^{D} w_i \Sigma_{ij} w_j}{\partial w_k} = 2\mathbf{w}^t \Sigma_{\bullet k}$$

Por tanto, si realizáramos la derivada del primer término respecto al vector $\mathbf{w}$ completo tendremos

$$\frac{\partial \sum_{i=1}^{D} \sum_{j=1}^{D} w_i \, \Sigma_{ij} \, w_j}{\partial \mathbf{w}} = \left(2\mathbf{w}^t \Sigma_{\bullet 1}, 2\mathbf{w}^t \Sigma_{\bullet 2}, \cdots, 2\mathbf{w}^t \Sigma_{\bullet D}\right) = 2\mathbf{w}^t \Sigma_{\mathcal{X}}$$

En consecuencia, la derivada de E respecto a $\mathbf{w}$ queda como

$$\frac{\partial E}{\partial \, \mathbf{w}} = 2\mathbf{w}^t \Sigma_{\mathcal{X}} - 2\lambda \mathbf{w}^t$$

Al igualar a cero

$$2\mathbf{w}^t \Sigma_{\mathcal{X}} - 2\lambda \mathbf{w}^t = 0 \rightarrow 2\mathbf{w}^t \Sigma_{\mathcal{X}} = 2\lambda \mathbf{w}^t \rightarrow \mathbf{w}^t \Sigma_{\mathcal{X}} = \lambda \mathbf{w}^t \rightarrow \Sigma_{\mathcal{X}} \mathbf{w} = \lambda \mathbf{w}$$

Es decir, $\mathbf{w}$ es un vector propio con λ valor propio asociado.