Kernel Fisher Discriminant Analysis

Advanced Institute for Artificial Intelligence – AI2

https://advancedinstitute.ai

A técnica de *Kernel Fisher Discriminant Analysis* - (KFDA) é uma **generalização não linear** para LDA. Novamente, faremos uso dos *kernels* para tornar LDA uma técnica de projeção não linear. Inicialmente, iremos abordar a versão para classificação com duas classes.

Seja, então, $\mathcal{X} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_z, y_z)\}$ um conjunto de dados tal que $\boldsymbol{x}_i \in \mathbb{R}^n$ e $\phi : \mathbb{R}^n \rightarrow \mathbb{R}^{n'}$ uma função de mapeamento não linear tal que $n' > n$. Ademais, temos que $\mathcal{Y} = \{\omega_1, \omega_2\}$ de tal forma que $y_i \in \mathcal{Y}$. No KFDA busca-se maximizar o seguinte critério:

$$J(\boldsymbol{w}) = \frac{\boldsymbol{w}^T \boldsymbol{S}_B^{\phi} \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{S}_A^{\phi} \boldsymbol{w}}. \tag{1}$$

Note que o critério é o mesmo que LDA. O que muda, porém, é a maneira com a qual temos que calcular as matrizes de espalhamento interclasses $\boldsymbol{S}_B^{\phi}$ e intraclasses $\boldsymbol{S}_A^{\phi}$.

Neste caso, temos que:

$$\boldsymbol{S}_B^{\phi} = (\boldsymbol{\mu}_1^{\phi} - \boldsymbol{\mu}_2^{\phi})(\boldsymbol{\mu}_1^{\phi} - \boldsymbol{\mu}_2^{\phi})^T. \tag{2}$$

Além disso, temos que:

$$\boldsymbol{S}_A^{\phi} = \boldsymbol{S}_1^{\phi} + \boldsymbol{S}_2^{\phi}, \tag{3}$$

em que

$$\boldsymbol{S}_i^{\phi} = \sum_{\boldsymbol{x}_j \in \omega_i} (\phi(\boldsymbol{x}_j) - \boldsymbol{\mu}_i^{\phi})(\phi(\boldsymbol{x}_j) - \boldsymbol{\mu}_i^{\phi})^T,\ i \in \{1, 2\}. \tag{4}$$

Temos que a média de cada classe é dada por:

$$\boldsymbol{\mu}_i^{\phi} = \frac{1}{m_i} \sum_{j=1}^{m_i} \phi(\boldsymbol{x}_j),\ i \in \{1, 2\}, \tag{5}$$

em que m_i denota a quantidade de elementos do conjunto de treinamento $\mathcal{X}_1 \subset \mathcal{X}$ da classe ω_i tal que $|\mathcal{X}_1| = m = m_1 + m_2$.

Existe um teorema importante que diz que **qualquer** vetor solução $\boldsymbol{w} \in \mathbb{R}^{n'}$ para o problema de maximizar a Equação 1 deve fazer parte do espaço gerado por todas as amostras do conjunto de treinamento. Assim, temos que $\boldsymbol{w}$ é uma combinação linear das amostras de treinamento, ou seja:

$$\boldsymbol{w} = \sum_{i=1}^{m} \alpha_i \phi(\boldsymbol{x}_i), \quad \forall \boldsymbol{x}_i \in \mathcal{X}_1. \tag{6}$$

Temos que a projeção de $\boldsymbol{\mu}_1^{\phi}$ na direção do vetor $\boldsymbol{w}$ é dada por:

$$\boldsymbol{w}^T\boldsymbol{\mu}_1^{\phi} = \overbrace{\left(\sum_{i=1}^{m}\alpha_i\phi(\boldsymbol{x}_i)\right)^T}^{\boldsymbol{w}^T}\overbrace{\left(\frac{1}{m_1}\sum_{j=1}^{m_1}\phi(\boldsymbol{x}_j)\right)}^{\boldsymbol{\mu}_1^{\phi}} = \frac{1}{m_1}\sum_{i=1}^{m}\sum_{j=1}^{m_1}\alpha_i\overbrace{\phi(\boldsymbol{x}_i)^T\phi(\boldsymbol{x}_j)}^{\text{kernel}}$$

$$= \frac{1}{m_1}\sum_{i=1}^{m}\sum_{j=1}^{m_1}\alpha_i K(\boldsymbol{x}_i,\boldsymbol{x}_j). \tag{7}$$

De maneira análoga, temos que a projeção de $\boldsymbol{\mu}_2^\phi$ na direção do vetor $\boldsymbol{w}$ é dada por:

$$\boldsymbol{w}^T\boldsymbol{\mu}_2^\phi = \overbrace{\left(\sum_{i=1}^{m}\alpha_i\phi(\boldsymbol{x}_i)\right)^T}^{\boldsymbol{w}^T}\overbrace{\left(\frac{1}{m_2}\sum_{j=1}^{m_2}\phi(\boldsymbol{x}_j)\right)}^{\boldsymbol{\mu}_2^\phi} = \frac{1}{m_1}\sum_{i=1}^{m}\sum_{j=1}^{m_2}\alpha_i\overbrace{\phi(\boldsymbol{x}_i)^T\phi(\boldsymbol{x}_j)}^{\text{kernel}}$$

$$= \frac{1}{m_2}\sum_{i=1}^{m}\sum_{j=1}^{m_2}\alpha_i K(\boldsymbol{x}_i,\boldsymbol{x}_j). \tag{8}$$

Note, então, que as projeções dependem da matriz de *kernel* e $\boldsymbol{\alpha}$, e que $(\boldsymbol{w}^T\boldsymbol{\mu}_i^\phi) \in \mathbb{R}$, $i \in \{1,2\}$.

Seja $\boldsymbol{M}^1 \in \mathbb{R}^m$ tal que:

$$\boldsymbol{M}_i^1 = \frac{1}{m_1} \sum_{j=1}^{m_1} K(\boldsymbol{x}_i, \boldsymbol{x}_j), \tag{9}$$

em que $\boldsymbol{x}_j \in \omega_1$ e $\boldsymbol{x}_i \in \mathcal{X}_1$. Basicamente, a equação acima calcula a média da i-ésima linha da matriz de *kernel* da classe 1. O mesmo vale para o cálculo de $\boldsymbol{M}^2 \in \mathbb{R}^m$, ou seja:

$$\boldsymbol{M}_i^2 = \frac{1}{m_2} \sum_{j=1}^{m_2} K(\boldsymbol{x}_i, \boldsymbol{x}_j). \tag{10}$$

Desta forma, podemos reescrever a Equação 7 da seguinte forma:

$$
\begin{aligned}
\boldsymbol{w}^T\boldsymbol{\mu}_1^{\phi} &= \frac{1}{m_1}\sum_{i=1}^{m}\sum_{j=1}^{m_1}\alpha_i K(\boldsymbol{x}_i,\boldsymbol{x}_j) \\
&= \sum_{i=1}^{m}\frac{1}{m_1}\sum_{j=1}^{m_1}\alpha_i K(\boldsymbol{x}_i,\boldsymbol{x}_j) \\
&= \sum_{i=1}^{m}\alpha_i \underbrace{\frac{1}{m_1}\sum_{j=1}^{m_1}K(\boldsymbol{x}_i,\boldsymbol{x}_j)}_{\boldsymbol{M}_i^1} \\
&= \sum_{i=1}^{m}\alpha_i\boldsymbol{M}_i^1 = \boldsymbol{\alpha}\boldsymbol{M}^1,
\end{aligned}
\tag{11}
$$

em que $\boldsymbol{\alpha}\in\mathbb{R}^m$.

De modo análogo, podemos reescrever a Equação 8 da seguinte forma:

$$
\begin{aligned}
\boldsymbol{w}^T \boldsymbol{\mu}_2^{\phi} &= \frac{1}{m_2} \sum_{i=1}^{m} \sum_{j=1}^{m_2} \alpha_i K(\boldsymbol{x}_i, \boldsymbol{x}_j) \\
&= \sum_{i=1}^{m} \frac{1}{m_2} \sum_{j=1}^{m_2} \alpha_i K(\boldsymbol{x}_i, \boldsymbol{x}_j) \\
&= \sum_{i=1}^{m} \alpha_i \underbrace{\frac{1}{m_2} \sum_{j=1}^{m_2} K(\boldsymbol{x}_i, \boldsymbol{x}_j)}_{\boldsymbol{M}_i^2} \\
&= \sum_{i=1}^{m} \alpha_i \boldsymbol{M}_i^2 = \boldsymbol{\alpha} \boldsymbol{M}^2.
\end{aligned}
\tag{12}
$$

A ideia de toda essa manipulação é para reescrevermos a Equação 1 (Critério de Fisher).

Desta fora, conseguimos reescrever o numerador da Equação 1 da seguinte forma:

$$\begin{aligned}
\boldsymbol{w}^T \boldsymbol{S}_B^{\phi} \boldsymbol{w} &= \boldsymbol{w}^T (\boldsymbol{\mu}_1^{\phi} - \boldsymbol{\mu}_2^{\phi})(\boldsymbol{\mu}_1^{\phi} - \boldsymbol{\mu}_2^{\phi})^T \boldsymbol{w} \\
&= (\boldsymbol{w}^T \boldsymbol{\mu}_1^{\phi} - \boldsymbol{w}^T \boldsymbol{\mu}_2^{\phi})(\boldsymbol{\mu}_1^{\phi^T} \boldsymbol{w} - \boldsymbol{\mu}_2^{\phi^T} \boldsymbol{w}) \\
&= (\boldsymbol{\alpha}^T \boldsymbol{M}^1 - \boldsymbol{\alpha}^T \boldsymbol{M}^2)(\boldsymbol{M}^{1^T} \boldsymbol{\alpha} - \boldsymbol{M}^{2^T} \boldsymbol{\alpha}) \\
&= \boldsymbol{\alpha}^T (\boldsymbol{M}^1 - \boldsymbol{M}^2)(\boldsymbol{M}^1 - \boldsymbol{M}^2)^T \boldsymbol{\alpha} \\
&= \boldsymbol{\alpha}^T \boldsymbol{M} \boldsymbol{\alpha},
\end{aligned} \tag{13}$$

em que $\boldsymbol{M} = (\boldsymbol{M}^1 - \boldsymbol{M}^2)(\boldsymbol{M}^1 - \boldsymbol{M}^2)^T$.

Agora, reescreveremos o denominador da Equação 1, ou seja:

$$
\begin{aligned}
\boldsymbol{w}^T \boldsymbol{S}_A^{\phi} \boldsymbol{w} &= \left(\sum_{i=1}^{m} \alpha_i \phi(\boldsymbol{x}_i)\right)^T \boldsymbol{S}_A^{\phi} \left(\sum_{l=1}^{m} \alpha_i \phi(\boldsymbol{x}_l)\right) \\
&= \left(\sum_{i=1}^{m} \alpha_i \phi(\boldsymbol{x}_i)\right)^T \underbrace{\left(\sum_{j=1}^{2} \sum_{k=1}^{m_j} (\phi(\boldsymbol{x}_k) - \boldsymbol{\mu}_{\boldsymbol{j}}^{\phi})(\phi(\boldsymbol{x}_k) - \boldsymbol{\mu}_{\boldsymbol{j}}^{\phi})^T\right)}_{\boldsymbol{S}_A^{\phi} = \boldsymbol{S}_1^{\phi} + \boldsymbol{S}_2^{\phi}} \left(\sum_{l=1}^{m} \alpha_i \phi(\boldsymbol{x}_l)\right). \quad (14)
\end{aligned}
$$

Agrupando os somatórios e realizando algumas simplificações, temos que:

$$\boldsymbol{w}^T \boldsymbol{S}_A^{\phi} \boldsymbol{w} = \boldsymbol{\alpha}^T \boldsymbol{N} \boldsymbol{\alpha}, \tag{15}$$

em que $\boldsymbol{N} = \boldsymbol{N}_1 + \boldsymbol{N}_2 = \underbrace{\boldsymbol{K}_1(\boldsymbol{I}_1 - \mathbf{1}_{m_1})\boldsymbol{K}_1^T}_{\boldsymbol{N}_1} + \underbrace{\boldsymbol{K}_2(\boldsymbol{I}_2 - \mathbf{1}_{m_2})\boldsymbol{K}_2^T}_{\boldsymbol{N}_2}$. Neste caso, temos que $\boldsymbol{K}_1 \in \mathbb{R}^{m \times m_1}$ corresponde à matriz de *kernel* da classe ω_1, $\boldsymbol{I}_1 \in \mathbb{R}^{m_1 \times m_1}$ denota a matriz identidade correspondente à classe ω_1 e $\mathbf{1}_{m_1} \in \mathbb{R}^{m_1 \times m_1}$ representa a matriz com entradas $1/m_1$ também da classe ω_1. De maneira análoga, podemos definir $\boldsymbol{K}_2 \in \mathbb{R}^{m \times m_2}$, $\boldsymbol{I}_2 \in \mathbb{R}^{m_2 \times m_2}$ e $\mathbf{1}_{m_2} \in \mathbb{R}^{m_2 \times m_2}$ como sendo as matrizes de *kernel*, identidade e com entradas $1/m_2$ da classe ω_2, respectivamente.

Desta forma, conseguimos reescrever a Equação 1 da seguinte forma:

$$J(\boldsymbol{\alpha}) = \frac{\boldsymbol{\alpha}^T \boldsymbol{M} \boldsymbol{\alpha}}{\boldsymbol{\alpha}^T \boldsymbol{N} \boldsymbol{\alpha}}. \tag{16}$$

A condição necessária para otimizarmos a equação acima é dada por:

$$\frac{\partial J(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} = 0. \tag{17}$$

A solução da Equação 17 é dada por:

$$\frac{\partial J(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} \implies (\boldsymbol{N}^{-1}\boldsymbol{M})\boldsymbol{\alpha} = \lambda \boldsymbol{\alpha}. \tag{18}$$

Assim sendo, a solução é dada pelo autovetor $\boldsymbol{\alpha}$ associado ao maior autovalor da matriz $(\boldsymbol{N}^{-1}\boldsymbol{M})$, que é baseada nas matrizes de *kernel*.

Agora, dada uma amostra de teste $\boldsymbol{x} \in \mathcal{X}_2$, como realizamos a sua projeção LDA? Basta, então, projetarmos essa amostra no vetor $\boldsymbol{w}$ encontrado pela função de otimização dada pela Equação 16:

$$\hat{\boldsymbol{x}} = \phi(\boldsymbol{x})^T \boldsymbol{w}. \tag{19}$$

No entanto, a Equação 16 é escrita em termos de $\boldsymbol{\alpha}$, o que nos remete à Equação 6, ou seja:

$$\hat{\boldsymbol{x}} = \phi(\boldsymbol{x})^T \boldsymbol{w} = \phi(\boldsymbol{x})^T \sum_{i=1}^{m} \alpha_i \phi(\boldsymbol{x}) = \sum_{i=1}^{m} \alpha_i \phi(\boldsymbol{x})^T \phi(\boldsymbol{x}) = \sum_{i=1}^{m} \alpha_i K(\boldsymbol{x}, \boldsymbol{x}_i). \tag{20}$$

Assim sendo, uma amostra de teste pode ser projetada por meio de uma combinação linear das funções de *kernel* das amostras do conjunto de treinamento.

KFDA Multiclasses

Podemos generalizar o KFDA para problemas multiclasses também, ou seja, $\mathcal{Y} = \{\omega_1, \omega_2, \ldots, \omega_c\}$. Neste caso, da Equação 13, temos que $\boldsymbol{M} = (\boldsymbol{M}^1 - \boldsymbol{M}^2)(\boldsymbol{M}^1 - \boldsymbol{M}^2)^T$. Podemos, então, redefinir a matriz $\boldsymbol{M}$ da seguinte forma:

$$\boldsymbol{M} = \sum_{j=1}^{c} m_j (\boldsymbol{M}^j - \boldsymbol{M}^*)(\boldsymbol{M}^j - \boldsymbol{M}^*)^T, \tag{21}$$

em que $\boldsymbol{M}^j, \boldsymbol{M}^* \in \mathbb{R}^m$ e são calculados da seguinte forma:

$$M_i^j = \frac{1}{m_j} \sum_{k=1}^{m_j} K(\boldsymbol{x}_i, \boldsymbol{x}_k), \tag{22}$$

em que $\boldsymbol{x}_k$ varia entre todas as amostras da classe ω_j e $\boldsymbol{x}_i$ varia entre todas as amostras de treinamento.

Já $\boldsymbol{M}^*$ pode ser calculado da seguinte forma:

$$M_i^* = \frac{1}{m}\sum_{k=1}^{m} K(\boldsymbol{x}_i, \boldsymbol{x}_k), \tag{23}$$

$\boldsymbol{x}_i$ e $\boldsymbol{x}_m$ variam entre todas as amostras de treinamento. Já a matriz $\boldsymbol{N} = \boldsymbol{N}_1 + \boldsymbol{N}_2$ da Equação 15 também pode ser generalizada da seguinte forma:

$$\boldsymbol{N} = \boldsymbol{N}_1 + \boldsymbol{N}_2 + \ldots + \boldsymbol{N}_c = \sum_{j=1}^{c} \boldsymbol{K}_j(\boldsymbol{I} - \boldsymbol{1}_{m_j})\boldsymbol{K}_j^T. \tag{24}$$

Desta forma, a solução ótima para o KFDA multiclasses consiste, novamente, em resolvermos a Equação 18, ou seja, consiste em obter os d autovetores associados aos d maiores autovalores da matriz $\boldsymbol{N}^{-1}\boldsymbol{M}$. A nova base é então definida por:

$$\boldsymbol{W} = [\alpha_1, \alpha_2, \ldots, \alpha_d], \tag{25}$$

ou seja, cada coluna da matriz $\boldsymbol{W}$ representa um autovetor α_k. Já a projeção de uma amostra $\boldsymbol{x}$ é dada por:

$$\hat{\boldsymbol{x}} = \phi(\boldsymbol{x})^T \boldsymbol{w}_k = \sum_{i=1}^{m} \alpha_{ki}\phi(\boldsymbol{x})^T\phi(\boldsymbol{x}_i) = \sum_{i=1}^{m} \alpha_{ki}K(\boldsymbol{x}, \boldsymbol{x}_i), \tag{26}$$

em que $\boldsymbol{x}_i$ varia entre todas as amostras do conjunto de treinamento.

Um agradecimento especial ao **Prof. Alexandre Levada** do Centro de Ciências Exatas e de Tecnologia, Departamento de Computação, Universidade Federal de São Carlos, pelas notas de aula.