

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, $\mathcal{X} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_z, y_z)\}$ um conjunto de dados tal que $\mathbf{x}_i \in \mathbb{R}^n$ 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:

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

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 \mathbf{S}_B^ϕ e intraclasses \mathbf{S}_A^ϕ .

Neste caso, temos que:

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

Além disso, temos que:

$$\mathbf{S}_A^\phi = \mathbf{S}_1^\phi + \mathbf{S}_2^\phi, \quad (3)$$

em que

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

Temos que a média de cada classe é dada por:

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

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 $\mathbf{w} \in \mathbb{R}^{n'}$ 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 \mathbf{w} é uma combinação linear das amostras de treinamento, ou seja:

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

Temos que a projeção de μ_1^ϕ na direção do vetor w é dada por:

$$\begin{aligned} w^T \mu_1^\phi &= \overbrace{\left(\sum_{i=1}^m \alpha_i \phi(\mathbf{x}_i) \right)^T}^{w^T} \overbrace{\left(\frac{1}{m_1} \sum_{j=1}^{m_1} \phi(\mathbf{x}_j) \right)}^{\mu_1^\phi} = \frac{1}{m_1} \sum_{i=1}^m \sum_{j=1}^{m_1} \alpha_i \overbrace{\phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)}^{\text{kernel}} \\ &= \frac{1}{m_1} \sum_{i=1}^m \sum_{j=1}^{m_1} \alpha_i K(\mathbf{x}_i, \mathbf{x}_j). \end{aligned} \tag{7}$$

De maneira análoga, temos que a projeção de μ_2^ϕ na direção do vetor w é dada por:

$$\begin{aligned}
 w^T \mu_2^\phi &= \overbrace{\left(\sum_{i=1}^m \alpha_i \phi(\mathbf{x}_i) \right)^T}^{w^T} \overbrace{\left(\frac{1}{m_2} \sum_{j=1}^{m_2} \phi(\mathbf{x}_j) \right)}^{\mu_2^\phi} = \frac{1}{m_1} \sum_{i=1}^m \sum_{j=1}^{m_2} \alpha_i \overbrace{\phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)}^{\text{kernel}} \\
 &= \frac{1}{m_2} \sum_{i=1}^m \sum_{j=1}^{m_2} \alpha_i K(\mathbf{x}_i, \mathbf{x}_j).
 \end{aligned} \tag{8}$$

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

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

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

em que $\mathbf{x}_j \in \omega_1$ e $\mathbf{x}_i \in \mathcal{X}_1$. 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 $\mathbf{M}^2 \in \mathbb{R}^m$, ou seja:

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

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

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

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

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

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

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:

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

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

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

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

Agrupando os somatórios e realizando algumas simplificações, temos que:

$$\mathbf{w}^T \mathbf{S}_A^\phi \mathbf{w} = \boldsymbol{\alpha}^T \mathbf{N} \boldsymbol{\alpha}, \quad (15)$$

em que $\mathbf{N} = \mathbf{N}_1 + \mathbf{N}_2 = \underbrace{\mathbf{K}_1(\mathbf{I}_1 - \mathbf{1}_{m_1})\mathbf{K}_1^T}_{\mathbf{N}_1} + \underbrace{\mathbf{K}_2(\mathbf{I}_2 - \mathbf{1}_{m_2})\mathbf{K}_2^T}_{\mathbf{N}_2}$. Neste caso, temos que

$\mathbf{K}_1 \in \mathbb{R}^{m_1 \times m_1}$ corresponde à matriz de *kernel* da classe ω_1 , $\mathbf{I}_1 \in \mathbb{R}^{m_1 \times m_1}$ 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 $\mathbf{K}_2 \in \mathbb{R}^{m_2 \times m_2}$, $\mathbf{I}_2 \in \mathbb{R}^{m_2 \times m_2}$ 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:

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

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

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

A solução da Equação 17 é dada por:

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

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

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

$$\hat{\mathbf{x}} = \phi(\mathbf{x})^T \mathbf{w}. \quad (19)$$

No entanto, a Equação 16 é escrita em termos de α , o que nos remete à Equação 6, ou seja:

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

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.

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

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

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

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

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

Já M^* pode ser calculado da seguinte forma:

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

\mathbf{x}_i e \mathbf{x}_k variam entre todas as amostras de treinamento. Já a matriz $N = N_1 + N_2$ da Equação 15 também pode ser generalizada da seguinte forma:

$$N = N_1 + N_2 + \dots + N_c = \sum_{j=1}^c K_j(I - \mathbf{1}_{m_j})K_j^T. \quad (24)$$

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

$$\mathbf{W} = [\alpha_1, \alpha_2, \dots, \alpha_d], \quad (25)$$

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

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

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

Um agradecimento especial ao **Prof. João Paulo Papa** do Departamento de Computação, Universidade Estadual Paulista (UNESP), pelas notas de aula.