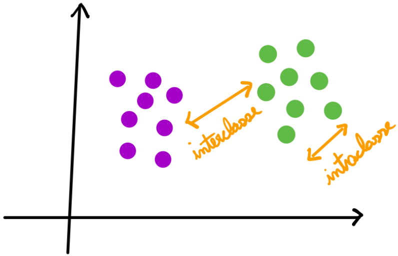
Análise Discriminante Linear

Advanced Institute for Artificial Intelligence – AI2

https://advancedinstitute.ai

Introdução

Análise Discriminante Linear, do inglês *Linear Discriminant Analysis* - LDA, é uma técnica **supervisionada** que objetiva **maximizar a separabilidade** entre as classes. A ideia é gerar agrupamentos em que a distância intraclasse (elementos de mesma classe) seja pequena, enquanto que a distância interclasse (elementos de classes distintas) seja grande. **Queremos, então, gerar agrupamentos compactos e distantes.**



Seja $\mathcal{X} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_z, y_z)\}$ um conjunto de dados rotulado tal que $\boldsymbol{x}_i \in \mathbb{R}^n$. Temos que $\mathcal{X}_1 \subset \mathcal{X}$ e $\mathcal{X}_2 \subset \mathcal{X}$ representam duas partições que denotam os conjuntos de treinamento e teste, respectivamente. Ademais, seja $\mathcal{Y} = \{\omega_1, \omega_2, \ldots, \omega_c\}$ o conjunto de rótulos possíveis. Em nosso exemplo, vamos assumir que temos apenas duas classes, isto é, ω_1 e ω_2.

Podemos calcular as médias de cada agrupamento da seguinte forma:

$$\boldsymbol{\mu}_1 = \frac{1}{m_1} \sum_{\boldsymbol{x}_i \in \omega_1} \boldsymbol{x}_i \quad \text{e} \quad \boldsymbol{\mu}_2 = \frac{1}{m_2} \sum_{\boldsymbol{x}_j \in \omega_2} \boldsymbol{x}_j,$$

em que m_1 e m_2 correspondem ao número de amostras de treinamento das classes 1 e 2, respectivamente. Já $\boldsymbol{\mu}_1$ e $\boldsymbol{\mu}_2$ denotam os centros dos agrupamentos das amostras das classes 1 e 2, respectivamente.

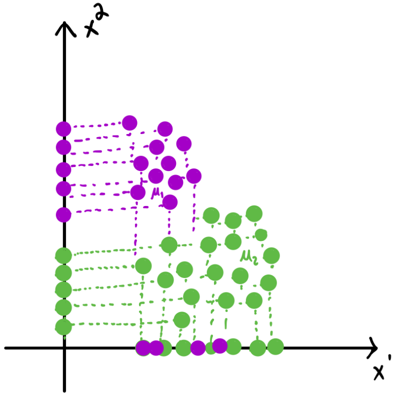
Objetivo: dado um conjunto de treinamento $\mathcal{X}_1$, achar um vetor direção $\boldsymbol{w} \in \mathbb{R}^n$ de tal forma que, quando projetarmos nossas amostras nesta direção, maximizaremos a sua separabilidade.

Sejam $\hat{\boldsymbol{\mu}}_1$ e $\hat{\boldsymbol{\mu}}_2$ as médias dos dados das classes 1 e 2, respectivamente, projetados na direção $\boldsymbol{w}$:

$$
\begin{aligned}
\hat{\boldsymbol{\mu}}_1 &= \frac{1}{m_1} \sum_{\boldsymbol{x}_i \in \omega_1} \boldsymbol{w}^T \boldsymbol{x}_i \\
&= \underbrace{\boldsymbol{w}^T}_{\text{constante}} \left(\frac{1}{m_1} \sum_{\boldsymbol{x}_i \in \omega_1} \boldsymbol{x}_i \right) \\
&= \boldsymbol{w}^T \boldsymbol{\mu}_1.
\end{aligned}
\tag{1}
$$

$$
\begin{aligned}
\hat{\boldsymbol{\mu}}_2 &= \frac{1}{m_2} \sum_{\boldsymbol{x}_j \in \omega_2} \boldsymbol{w}^T \boldsymbol{x}_j \\
&= \underbrace{\boldsymbol{w}^T}_{\text{constante}} \left(\frac{1}{m_2} \sum_{\boldsymbol{x}_j \in \omega_2} \boldsymbol{x}_j \right) \\
&= \boldsymbol{w}^T \boldsymbol{\mu}_2.
\end{aligned}
\tag{2}
$$

Um critério interessante seria maximizar a diferença entre as médias projetadas, ou seja, $|\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2|$:



Neste exemplo, a distância entre as médias é maior no eixo x^1, mas a separabilidade é maior no eixo x^2.

Por que isto ocorre? A abordagem acima não leva em consideração a variância, ou seja, o **espalhamento** entre as classe.

Dado o nosso conjunto de dados $\mathcal{X} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_z, y_z)\}$, temos que sua média amostral é dada por:

$$\boldsymbol{\mu} = \frac{1}{z}\sum_{i=1}^{z} \boldsymbol{x}_i, \tag{3}$$

e o seu espalhamento (*scatter*) é calculado como segue:

$$\boldsymbol{s}^2 = \sum_{i=1}^{z} (\boldsymbol{x}_i - \boldsymbol{\mu})^2. \tag{4}$$

Ademais, seja $\hat{\boldsymbol{x}}_i = \boldsymbol{w}^T \boldsymbol{x}_i$ a projeção da amostra $\boldsymbol{x}_i$ na direção do vetor $\boldsymbol{w}$.

Temos que os espalhamentos no espaço projetado para as classes ω_1 e ω_2 são definidos como segue:

$$\hat{\boldsymbol{s}}_1^2 = \sum_{\hat{\boldsymbol{x}}_i \in \omega_1} (\hat{\boldsymbol{x}}_i - \hat{\mu}_1)^2, \tag{5}$$

e

$$\hat{\boldsymbol{s}}_2^2 = \sum_{\hat{\boldsymbol{x}}_j \in \omega_2} (\hat{\boldsymbol{x}}_j - \hat{\mu}_2)^2. \tag{6}$$

Assim, desejamos **maximizar a distância entre as médias** e **minimizar o espalhamento entre as classes**. O Critério de Fisher atende à esta nossa exigência.

Temos, então, que maximizar o Critério de Fisher, dado por:

$$J(\boldsymbol{w}) = \frac{(\hat{\mu}_1 - \hat{\mu}_2)^2}{\hat{\boldsymbol{s}}_1^2 + \hat{\boldsymbol{s}}_2^2}. \tag{7}$$

Desta forma, queremos encontrar o vetor $\boldsymbol{w}$ que maximiza o critério, acima, ou seja:

$$\mathbf{w}^* = \arg\max_{\boldsymbol{w}} J(\boldsymbol{w}). \tag{8}$$

No entanto, precisamos reescrever tanto o numerador quanto o denominador em termos de $\boldsymbol{w}$.

Antes da projeção, podemos obter as matrizes de espalhamento (*scatter*) das classes ω_1 e ω_2 como segue:

$$\boldsymbol{S}_1 = \sum_{\boldsymbol{x}_i \in \omega_1} (\boldsymbol{x}_i - \boldsymbol{\mu}_1)(\boldsymbol{x}_i - \boldsymbol{\mu}_1)^T \tag{9}$$

e

$$\boldsymbol{S}_2 = \sum_{\boldsymbol{x}_j \in \omega_2} (\boldsymbol{x}_j - \boldsymbol{\mu}_2)(\boldsymbol{x}_j - \boldsymbol{\mu}_2)^T. \tag{10}$$

Seja $S_A = S_1 + S_2$ a matriz de espalhamento **intraclasse**. Podemos reescrever a Equação 5 como segue:

$$\hat{\boldsymbol{s}}_1^2 = \sum_{\hat{\boldsymbol{x}}_i \in \omega_1} (\hat{\boldsymbol{x}}_i - \hat{\boldsymbol{\mu}}_1)^2 = \sum_{\boldsymbol{x}_i \in \omega_1} (\boldsymbol{w}^T \boldsymbol{x}_i - \boldsymbol{w}^T \boldsymbol{\mu}_1)^2 = \sum_{\boldsymbol{x}_i \in \omega_1} [\boldsymbol{w}^T (\boldsymbol{x}_i - \boldsymbol{\mu}_1)]^T [\boldsymbol{w}^T (\boldsymbol{x}_i - \boldsymbol{\mu}_1)]. \tag{11}$$

Rearranjando os termos, temos que:

$$
\begin{aligned}
\hat{\boldsymbol{s}}_1^2 &= \sum_{\boldsymbol{x}_i \in \omega_1} [\boldsymbol{w}^T(\boldsymbol{x}_i - \boldsymbol{\mu}_1)]^T [\boldsymbol{w}^T(\boldsymbol{x}_i - \boldsymbol{\mu}_1)] \\
&= \sum_{\boldsymbol{x}_i \in \omega_1} [(\boldsymbol{x}_i - \boldsymbol{\mu}_1)^T \boldsymbol{w}]^T [(\boldsymbol{x}_i - \boldsymbol{\mu}_1)^T \boldsymbol{w}] \\
&= \sum_{\boldsymbol{x}_i \in \omega_1} \boldsymbol{w}^T(\boldsymbol{x}_i - \boldsymbol{\mu}_1)(\boldsymbol{x}_i - \boldsymbol{\mu}_1)^T \boldsymbol{w} \\
&= \boldsymbol{w}^T \underbrace{\sum_{\boldsymbol{x}_i \in \omega_1} (\boldsymbol{x}_i - \boldsymbol{\mu}_1)(\boldsymbol{x}_i - \boldsymbol{\mu}_1)^T}_{\boldsymbol{S}_1} \boldsymbol{w} \\
&= \boldsymbol{w}^T \boldsymbol{S}_1 \boldsymbol{w}.
\end{aligned}
\tag{12}
$$

De maneira análoga, podemos escrever:

$$\hat{\boldsymbol{s}}_2^2 = \boldsymbol{w}^T \boldsymbol{S}_2 \boldsymbol{w}. \tag{13}$$

Substituindo-se as Equações 12 e 13 na Equação 7, temos que:

$$J(\boldsymbol{w}) = \frac{(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)^2}{\hat{\boldsymbol{s}}_1^2 - \hat{\boldsymbol{s}}_2^2} = \frac{(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)^2}{\boldsymbol{w}^T \boldsymbol{S}_1 \boldsymbol{w} + \boldsymbol{w}^T \boldsymbol{S}_2 \boldsymbol{w}} = \frac{(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)^2}{\boldsymbol{w}^T \underbrace{(\boldsymbol{S}_1 + \boldsymbol{S}_2)}_{\boldsymbol{S}_A} \boldsymbol{w}} = \frac{(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)^2}{\boldsymbol{w}^T \boldsymbol{S}_A \boldsymbol{w}}, \tag{14}$$

a qual é uma forma quadrática em $\boldsymbol{w}$.

Seja, agora, $\boldsymbol{S}_B$ a matriz de espalhamento interclasses, que pode ser definida como segue:

$$\boldsymbol{S}_B = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T, \tag{15}$$

a qual mede a separação entre os vetores média antes da projeção. Note que o numerador da Equação 14 pode ser expressado como:

$$(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)^2 = (\boldsymbol{w}^T\boldsymbol{\mu}_1 - \boldsymbol{w}^T\boldsymbol{\mu}_2)^2 = [\boldsymbol{w}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]^T[\boldsymbol{w}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]. \tag{16}$$

Rearranjando os produtos internos, temos que:

$$(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)^2 = [(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\boldsymbol{w}^T]^T[(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\boldsymbol{w}^T] = \boldsymbol{w}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{w} = \boldsymbol{w}^T\boldsymbol{S}_B\boldsymbol{w}. \tag{17}$$

Podemos reescrever o numerador da Equação 14 utilizando o resultado da Equação 17, obtendo uma expressão final para o Critério de Fisher:

$$J(\boldsymbol{w}) = \frac{\boldsymbol{w}^T \boldsymbol{S}_B \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{S}_A \boldsymbol{w}}. \tag{18}$$

Precisamos, agora, maximizar a Equação 18 com relação à $\boldsymbol{w}$, ou seja, calculamos $\frac{\partial J(\boldsymbol{w})}{\partial \boldsymbol{w}}$, resultando em uma equação fechada para o cálculo de $\boldsymbol{w}$, dada por:

$$\boldsymbol{w}^* = \boldsymbol{S}_A^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \tag{19}$$

que é, basicamente, a diferença entre as médias modulada pela matriz de espalhamento.

Suponha, agora, que tenhamos um problema de classificação por múltiplas classes, isto é, $\mathcal{Y} = \{\omega_1, \omega_2, \ldots, \omega_c\}$. este caso, podemos reduzir a dimensionalidade do espaço original para, no máximo, $c-1$ classes, utilizando a Análise Discriminante Múltipla, do inglês *Multiple Discriminant Analysis* (MDA). Temos que a operação de transformação é dada por:

$$\hat{\boldsymbol{x}}_i = \boldsymbol{W}^T \boldsymbol{x}_i, \tag{20}$$

em que, novamente, $\boldsymbol{x}_i \in \mathbb{R}^n$ e $\boldsymbol{W} \in \mathbb{R}^{n \times d}$ é uma matriz de projeção para um espaço com $d < c$ dimensões. Temos que cada coluna de $\boldsymbol{W}$ é uma direção ortogonal $\boldsymbol{w}_j$.

Relembrando que temos as seguintes definições:

- m_i: número de elementos do conjunto de treinamento da classe ω_i.
- $\boldsymbol{\mu}_i$: média dos elementos da classe classe ω_i.
- $\boldsymbol{\mu}$: média global, isto é, de todo o conjunto de treinamento.

A função objetivo (Critério de Fisher) a ser maximizada é dada por:

$$J(\boldsymbol{W}) = \frac{|\boldsymbol{W}^T \boldsymbol{S}_B \boldsymbol{W}|}{|\boldsymbol{W}^T \boldsymbol{S}_A \boldsymbol{W}|}, \tag{21}$$

em que $|\cdot|$ calcula o determinante de uma matriz.

Já as matrizes $\boldsymbol{S}_A$ e $\boldsymbol{S}_B$ são calculadas como segue:

$$\boldsymbol{S}_A = \sum_{i=1}^{c} \boldsymbol{S}_i = \sum_{i=1}^{c} \sum_{\boldsymbol{x}_j \in \omega_i} (\boldsymbol{x}_j - \boldsymbol{\mu}_i)(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^T, \quad (22)$$

em que $\boldsymbol{S}_i$ representa a matriz de espalhamento da classe ω_i, e

$$\boldsymbol{S}_B = \sum_{i=1}^{c} m_i(\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T. \quad (23)$$

Pode-se mostrar que o posto (rank) da matriz $\boldsymbol{S}_B$ é $c-1$, ou seja, o número de linhas ou colunas linearmente independentes é limitado superiormente à $c-1$ (número de direções discriminantes do método).

A condição para maximização de nosso Critério de Fisher é dada por:

$$\frac{\partial J(\boldsymbol{W})}{\partial \boldsymbol{W}} = 0.$$

Resolvendo analiticamente a equação acima, chegamos na seguinte formulação:

$$(\boldsymbol{S}_B \boldsymbol{W}) = \lambda \boldsymbol{S}_A \boldsymbol{W}. \tag{24}$$

Caso $\boldsymbol{S}_A$ admita inversa, temos que:

$$(\boldsymbol{S}_A^{-1} \boldsymbol{S}_B)\boldsymbol{W} = \lambda \boldsymbol{W}. \tag{25}$$

Para resolver o problema, basta selecionarmos, no máximo, $c-1$ autovetores de $(\boldsymbol{S}_A^{-1}\boldsymbol{S}_B)$ associados aos $c-1$ maiores autovalores.

Algumas limitações do LDA/MDA:

- Não é interessante para problemas com poucas classes e muitas características.
- Quanto a média das classes são muito próximas, o numerador da Equação 7 tende a 0.
- Situações em que a função de custo (objetivo) assume valores muito altos (classes com formas não lineares).