Tema 2. Análise discriminante linear e regresión linear

- LDA: linear discriminant analysis, 2 clases
- Proxecta a un espazo $v=\mathbf{w}^T\mathbf{x}$ unidimensional no cal:

a) Maximízase a separación entre as medias das 2 clases (entre clases): $m_i=\mathbf{w}^T\mathbf{m}_i$, i=1,2

$$\boldsymbol{m}_i=\frac{1}{N_i}\sum_{y_j=i}\boldsymbol{x}_j$$

N_i=nº patróns de clase i

b) Minimízase a separación entre os patróns da mesma clase (intraclase)

- Medida de separación: covarianza no espazo proxectado
- Intraclase s_i^2: $$s_i^2=\sum_{y_j=i}\left(v_j-m_i\right)^2=\sum_{y_j=i}\left(\boldsymbol{w}^T\boldsymbol{x}_j-\boldsymbol{w}^T\boldsymbol{m}_i\right)^2, i=1,2$$

Proxección do patrón

Proxección da media de clase

LDA para clasificación binaria (I)

- Covarianza entre clases en 1D: $(m_1-m_2)^2$
- Criterio de Fisher: $J(\boldsymbol{w})=\frac{(m_1-m_2)^2}{s_1^2+s_2^2}$
- Substituíndo m_i e s_i^2 vs. **w**: $J(\boldsymbol{w})=\frac{\boldsymbol{w}^T\boldsymbol{S}_B\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{S}_W\boldsymbol{w}}$ Cociente de Rayleigh
- $\mathbf{S}_B$: matriz de covarianza entre clases no espazo orixinal

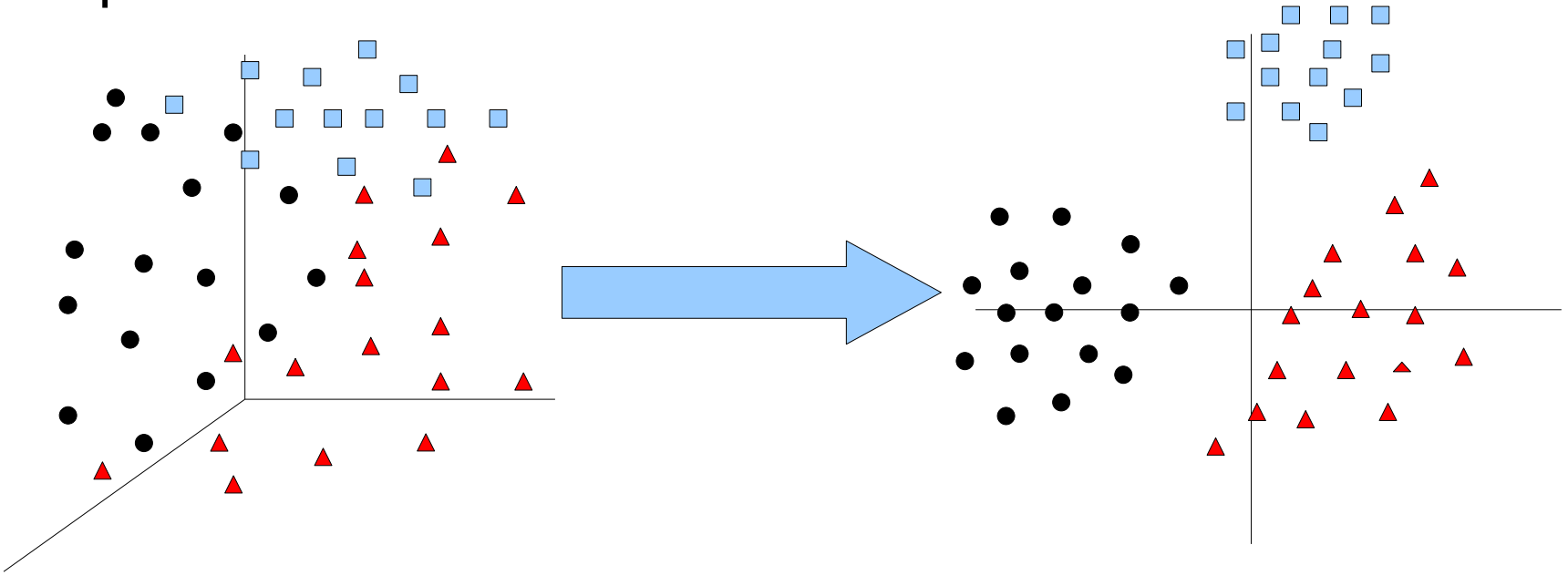
$$\boldsymbol{S}_B=(\boldsymbol{m}_2-\boldsymbol{m}_1)(\boldsymbol{m}_2-\boldsymbol{m}_1)^T \quad \boldsymbol{S}_W=\sum_{i=1}^{2}\sum_{y_j=i}(\boldsymbol{x}_j-\boldsymbol{m}_i)(\boldsymbol{x}_j-\boldsymbol{m}_i)^T$$

- $\mathbf{S}_W$: suma de matrices de covarianza intraclase

$\mathbf{S}_B$ e $\mathbf{S}_W$ cadradas de orde n

Exemplo de LDA

- Proxéctase a un espazo C-1-dimensional onde se maximiza a covarianza entre clases e minimízase a suma de covarianzas intraclase
- Exemplo con C=3 clases de $\mathbb{R}^3$ a $\mathbb{R}^2$



- Maximiza a covarianza entre clases e minimiza a covarianza intraclase

LDA para clasificación binaria (II)

- Para maximizar J(**w**), fago ∇J(**w**)=dJ(**w**)/d**w**=**0** e obteño:

$$(\mathbf{w}^T\mathbf{S}_B\mathbf{w})\mathbf{S}_W\mathbf{w}=(\mathbf{w}^T\mathbf{S}_W\mathbf{w})\mathbf{S}_B\mathbf{w}$$

- Consideremos que $\mathbf{S}_b\mathbf{w}=(\mathbf{m}_2-\mathbf{m}_1)(\mathbf{m}_2-\mathbf{m}_1)^T\mathbf{w}$.
- Como $(\mathbf{m}_2-\mathbf{m}_1)^T\mathbf{w}$ é un escalar, isto pódese escribir como $\lambda(\mathbf{m}_2-\mathbf{m}_1)$, obtendo

$$(\mathbf{w}^T\mathbf{S}_B\mathbf{w})\mathbf{S}_W\mathbf{w}=(\mathbf{w}^T\mathbf{S}_W\mathbf{w})\lambda(\mathbf{m}_2-\mathbf{m}_1)$$

- Suprimo os escalares $\mathbf{w}^T\mathbf{S}_B\mathbf{w}$ e $\mathbf{w}^T\mathbf{S}_W\mathbf{w}$ e o vector **w** é paralelo a $\mathbf{S}_W^{-1}(\mathbf{m}_2-\mathbf{m}_1)$
- Como |**w**| non importa, úsase $\mathbf{w}=\mathbf{S}_W^{-1}(\mathbf{m}_2-\mathbf{m}_1)$
- Se non existe $\mathbf{S}_W^{-1}$ regularízase: $(\mathbf{S}_W+\alpha\mathbf{1})^{-1}$, con $\alpha=0^+$
- Clasificación: y(**x**)=sign(**w**^{T}**x** - b): b é o umbral $b=\frac{\boldsymbol{w}^T\left(\boldsymbol{m}_1+\boldsymbol{m}_2\right)}{2}$

Discriminante linear de Fisher

- A miúdo úsase como sinónimo de LDA, pero hai lixeiras diferencias: o DL de Fisher non supón clases con distribucións normais nin iguais covarianzas de clase
- Se $\mathbf{m}_1$, $\mathbf{m}_2$ e $\mathbf{S}_1$, $\mathbf{S}_2$ son as medias e covarianzas das dúas clases, $\mathbf{w}^T\mathbf{x}$ ten media $\mathbf{w}^T\mathbf{m}_i$ e covarianza $\mathbf{w}^T\mathbf{S}_i\mathbf{w}$ para i=1,2
- Fisher definiu a separación S entre as dúas distribucións das clases como o cociente da varianza entre clases e a varianza intraclase:

$$S = \frac{\sigma_B^2}{\sigma_W^2} = \frac{\left(\boldsymbol{w}^T\boldsymbol{m}_2 - \boldsymbol{w}^T\boldsymbol{m}_1\right)^2}{\boldsymbol{w}^T\left(\boldsymbol{S}_1+\boldsymbol{S}_2\right)\boldsymbol{w}}$$

- É fácil demostrar que a máxima separación *S* acádase cando:

$$\boldsymbol{w} \propto \left(\boldsymbol{S}_1+\boldsymbol{S}_2\right)^{-1}\left(\boldsymbol{m}_2-\boldsymbol{m}_1\right)$$

- Se as distribucións son parecidas, un offset *b* axeitado é:

$$b = \frac{\boldsymbol{w}^T\left(\boldsymbol{m}_1+\boldsymbol{m}_2\right)}{2} = \frac{\boldsymbol{m}_1^T\boldsymbol{S}_1^{-1}\boldsymbol{m}_1 + \boldsymbol{m}_2\boldsymbol{S}_2^{-1}\boldsymbol{m}_2}{2}$$

- Se $\mathbf{S}_1=\mathbf{S}_2$ e as distribucións son normais, acádase o LDA

LDA multiclase (I)

- Con C>2 clases, LDA proxecta a espazo D-dimensional, con D≤C-1: usualmente D=C-1
- Suponse que as C clases teñen medias $\mathbf{m}_i$, i=1..C, e a mesma matriz de covarianza media $\mathbf{S}_W$:

$$\boldsymbol{S}_W = \frac{1}{C}\sum_{i=1}^{C}\frac{1}{N_i}\sum_{y_j=i}\left(\boldsymbol{x}_j-\boldsymbol{m}_i\right)\left(\boldsymbol{x}_j-\boldsymbol{m}_i\right)^T$$

- A covarianza $\mathbf{S}_B$ entre clases é: $\boldsymbol{S}_B = \frac{1}{C}\sum_{i=1}^{C}\left(\boldsymbol{m}_i-\boldsymbol{m}\right)\left(\boldsymbol{m}_i-\boldsymbol{m}\right)^T$
- A separación S entre clases na dirección**w** está dada por:

$$\boldsymbol{m} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i$$

$$S = \frac{\boldsymbol{w}^T\boldsymbol{S}_B\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{S}_W\boldsymbol{w}}$$

- O vector **w** é un autovector de $\mathbf{S}_W^{-1}\mathbf{S}_B$, e a separación S é o seu autovalor asociado

LDA multiclase (II)

- O producto exterior de 2 vectores n-dimensionais ($\mathbf{x}\mathbf{x}^T$, sendo $\mathbf{x}$ un vector columna) é unha matriz nxn de rango 1
- O rango da matriz $\mathbf{S}_B$ é $\leq$C-1 porque $\mathbf{S}_B$ é a suma de C matrices que son productos exteriores
- Como $\mathbf{S}_W^{-1}\mathbf{S}_B$ é diagonalizábel, a variabilidade das n entradas orixinais consérvase no subespazo de dimensión C-1 xerado polos C-1 meirandes autovalores de $\mathbf{S}_W^{-1}\mathbf{S}_B$
- No canto de por un vector $\mathbf{w}$ como na clasificación binaria, o LDA multiclase ven definido pola matriz $\mathbf{W}$, de orde (C-1)xn, que proxecta $\mathbf{x}_i$ a $\mathbf{v}_i=\mathbf{W}\mathbf{x}_i \in \mathbb{R}^{C-1}$, cun vector $\mathbf{w}_i$, con i=1..C-1, en cada fila
- O LDA emprégase para clasificación e tamén para reducción da dimensionalidade (de n a D$\leq$C-1)

LDA multiclase (III)

- No **programa** que veremos na clase **interactiva**:
- Matriz de covarianza de clase i: $\boldsymbol{S}_i = \frac{1}{N_i - 1} \sum_{y_j = i} (\boldsymbol{x}_j - \boldsymbol{m}_i)(\boldsymbol{x}_j - \boldsymbol{m}_i)^T$
- Matriz de covarianza total $\mathbf{S}_T$:

$$\boldsymbol{S}_T = \sum_{i=1}^{C} \frac{N_i - 1}{N - C} \boldsymbol{S}_i = \frac{1}{N - C} \sum_{i=1}^{C} \sum_{y_j = i} (\boldsymbol{x}_j - \boldsymbol{m}_i)(\boldsymbol{x}_j - \boldsymbol{m}_i)^T$$

- Offset b_i de clase i:

$$b_i = \log \frac{N_i}{N} - \frac{\boldsymbol{m}_i^T \boldsymbol{S}_T^{-1} \boldsymbol{m}_i}{2}, i = 1 \ldots C$$

- Vector $\mathbf{w}_i$ de clase i, con i=1..C: $\boldsymbol{w}_i = \boldsymbol{S}_T^{-1} \boldsymbol{m}_i$
- Activación $L_i(\mathbf{x})$ de clase i para patrón de teste **x**:

$$L_i(\boldsymbol{x}) = \boldsymbol{w}_i^T \boldsymbol{x} + b_i$$

- Clase predita z(**x**): $z(\boldsymbol{x}) = \underset{i=1 \ldots C}{argmax} \left\{ \frac{e^{L_i(\boldsymbol{x})}}{\sum_{c=1}^{C} e^{L_c(\boldsymbol{x})}} \right\}$

Función **softmax**: aproximación suavizada (derivábel) á función máximo

Regresión linear: mínimos cadrados (least mean squares, LMS)

- Función lineal para predecir a saída: $z=\mathbf{v}^T\mathbf{u}+b=\mathbf{w}^T\mathbf{x}$ onde $\mathbf{w}=(\mathbf{v},1)$,$\mathbf{x}=(\mathbf{u},1)$ dados $\{\mathbf{x}_i,y_i\}_{i=1}^N$

$$\boldsymbol{y}=\begin{bmatrix} y_1 \\ \cdots \\ y_N \end{bmatrix} \qquad \boldsymbol{X}=\begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \cdots & \cdots & \cdots \\ x_{N1} & \cdots & x_{Nn} \end{bmatrix}$$

- Minimizar o erro cadrático:

$$E(\boldsymbol{w})=\sum_{i=1}^{N}(y_i-\boldsymbol{w}^T\boldsymbol{x}_i)^2=|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}|^2=(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w})$$

- En forma matricial:

$$E(\boldsymbol{w})=\boldsymbol{y}^T\boldsymbol{y}-\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{w}-\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{y}+\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w}=\boldsymbol{y}^T\boldsymbol{y}-2\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{y}+\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w}$$

- Derivando e igualando a $\mathbf{0}$:

$$\nabla E(\boldsymbol{w})=\frac{\partial E(\boldsymbol{w})}{\partial \boldsymbol{w}}=-2\boldsymbol{X}^T\boldsymbol{y}+2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w}=\mathbf{0}\Rightarrow\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w}=\boldsymbol{X}^T\boldsymbol{y}$$

- A solución é: $\boxed{\boldsymbol{w}=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}=\boldsymbol{X}^\dagger\boldsymbol{y}}$

- A matriz $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ é a pseudoinversa de Moore-Penrose ($\mathbf{X}^\dagger$) de $\mathbf{X}$: función pinv(x) en Octave

- Función **lm** de paquete **stats** en R

Regresión linear regularizada (ridge regression)

- **Mínimos cadrados regularizados** ou **ridge** (Tikhonov): para evitar sobreaprendizaxe, pondérase tamén |**w**|, que mide a complexidade do modelo:

$$E(\boldsymbol{w}) = \sum_{i=1}^{N} (y_i - \boldsymbol{w}^T \boldsymbol{x}_i)^2 + \lambda |\boldsymbol{w}|^2 = |\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}|^2 + \lambda |\boldsymbol{w}|^2$$

- Orientada a problemas colineares (ill-posed) nos que a matriz **X** é singular, o cal é frecuente con datos anchos (n alto).
- Suma un valor positivo (parámetro de regularización, λ) á diagonal da matriz.
- Solución: $\boldsymbol{w} = (\lambda \boldsymbol{I}_n + \boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}$ ← **I**$_n$ é a matriz identidade cadrada de orde n
- Función **lm.ridge** en paquete **MASS** de R
- **Regresión loxística**: $y = \mathbf{w}^T \Phi(\mathbf{x}) + b$, onde $\Phi(\mathbf{x}) = (\Phi_1, \ldots, \Phi_m)$ e $\Phi_i(\mathbf{x}) = f((x_j - m_j)/\sigma_j)$, sendo $f(t) = 1/(1+e^{-t})$ a función sigmoide

Implementacións

- As implementacións en Octave non usan ningúnha función ou obxecto

	Python	**R**	**Weka-Java**
LDA	discriminant_analysis/ LinearDiscriminantAnalysis	MASS/lda	weka.classifiers.functions. LDA
Regresión linear	linear_model/LinearRegression	MASS/lm	weka.classifiers.functions. LinearRegression
Regresion regularizada	linear_model/Ridge	MASS/ lm.ridge	