Práctica 3

Selección de características mediante PCA y MDA

Reducción de dimensionalidad

Objetivo:

- Reducir el número de características (asumiendo vectores columna):

$$\mathbf{x}_k \ (d \text{ características}) \quad \Rightarrow \quad \mathbf{y}_k = \underbrace{\mathbf{W}^T}_{d' \times d} \mathbf{x}_k \ (d' \text{ características})$$

  - **OJO !!!** : en la práctica trabajamos con vectores fila y por lo tanto la matriz **W** ha de multiplicar por la derecha

- **Eso ayuda a…**
  - Simplificar la estructura del clasificador
  - Minimizar el coste computacional
  - Eliminar información redundante

- **A tener en cuenta…**
  - Hay que diseñar de manera adecuada la matriz de reducción **W** a partir de la BdD de training

Matriz de dispersión

$$\mathbf{m}_i = \frac{1}{N_i}\sum_{\mathbf{x}\in D_i}\mathbf{x}$$ *Media de los datos de la clase i*

$$\mathbf{m} = \frac{1}{N}\sum_{\mathbf{x}\in\{D_1,\ldots,D_c\}}\mathbf{x} = \frac{1}{N}\sum_{i=1}^{c} N_i\mathbf{m}_i$$ *Media de todos los datos*

$$\mathbf{S}_T = \sum_{\mathbf{x}\in\{D_1,\ldots,D_c\}}(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^T$$ *Dispersión total de los datos*

$$\mathbf{S}_T = \underbrace{\sum_{i=1}^{c}\sum_{\mathbf{x}\in D_i}(\mathbf{x}-\mathbf{m}_i)(\mathbf{x}-\mathbf{m}_i)^T}_{\mathbf{S}_C} + \underbrace{\sum_{i=1}^{c}\sum_{\mathbf{x}\in D_i}(\mathbf{m}_i-\mathbf{m})(\mathbf{m}_i-\mathbf{m})^T}_{\mathbf{S}_B}$$

Suma de matrices de dispersión intra-clases

Matriz de dispersión inter-clases

PCA (Principal Component Analysis)

Objetivo:

- Maximizar: $traza\left(\mathbf{W}^T \mathbf{S}_T \mathbf{W}\right)$
- Restricciones: $\mathbf{w}_i^T \mathbf{w}_i = E$

Solución (función pca.m de MATLAB):

- Columnas de **W**: autovectores asociados a los máximos d' autovalores de $\mathbf{S}_T$:

$$\lambda_i \mathbf{w}_i = \mathbf{S}_T \mathbf{w}_i$$

Problema:

- Aunque minimiza el error cuadrático en la aproximación, no garantiza separabilidad entre las clases

MDA (Multiple Discriminant Analysis)

Objetivo:

- Maximizar la separación entre las clases a la vez que se intenta minimizar la dispersión dentro de cada clase
- Medimos la separación y la dispersión mediante los volúmenes de los elipsoides suponiendo Gaussianidad

Formulación:

- Maximización: $\mathbf{W} = \arg\max_{\mathbf{w}} \frac{\left|\mathbf{W}^T \mathbf{S}_B \mathbf{W}\right|}{\left|\mathbf{W}^T \mathbf{S}_C \mathbf{W}\right|}$

Solución (función mda_clp.m de MATLAB):

- d' ≤ min(d,c-1) (c: número de clases, ya que $\mathbf{S}_B$ tiene rango c-1)
- Columnas de **W**: autovectores asociados los autovalores máximos:

$$\mathbf{S}_B \mathbf{w}_j = \sigma_j \mathbf{S}_C \mathbf{w}_j \quad \Rightarrow \quad \mathbf{S}_C^{-1} \mathbf{S}_B \mathbf{w}_j = \sigma_j \mathbf{w}_j$$

Práctica

Parte I:

- Generación de BdD Gaussiana
- c=3 clases, d=3 características
- Evaluar PCA y MDA con d'=1, 2
- Comparar visualmente y en términos de probabilidad de error

Parte II:

- Utilización de BdD de fonemas (práctica 2)
- c=5 clases, d=256 características
- Con PCA evaluar d'=1,…,256
- Con MDA evaluar d'=1,…,4
- Comparar en términos de probabilidad de error