Similarity Based Constraint Score For Feature Selection

Mid-term Research Project Presentation

Gadegbeku Fabio

Supervised By : Dr. Ludovic Macaire

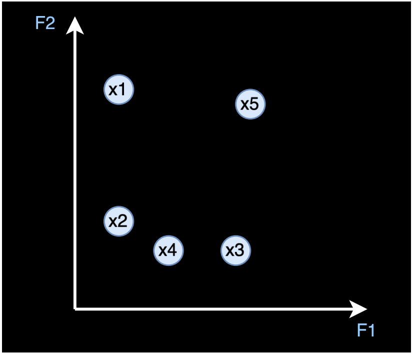
January 30, 2024

Table of Contents

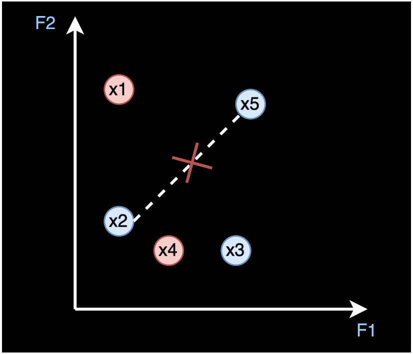
Introduction

Semi Supervised Learning

- We can define **must link** and **cannot link** constraints
  - must link : When two samples belong to the same class
  - cannot link : When two samples belong to different classes
- **Constraint Scores** to evaluate how well each feature matches the constraints



(a) Unsupervised



(b) Semi Supervised

Introduction

Schema

- Read and understand the papers on the subjects
- Implement an unsupervised score and two semi supervised scores described in [1]
- Compare results on these scores on multiple datasets

1

[1] Mariam Kalakech et al. "Constraint Scores for Semi-Supervised Feature Selection: A Comparative Study". In: *Pattern Recognition Letters* 32.5 (Apr. 2011), pp. 656–665. ISSN: 01678655. DOI: 10.1016/j.patrec.2010.12.014. (Visited on 09/24/2023).

Laplacian Score (Unsupervised)

Our n samples of d features

$$X = \begin{bmatrix} x_{11} & \dots & x_{1r} & \dots & x_{1d} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & \dots & x_{ir} & \dots & x_{id} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{nr} & \dots & x_{nd} \end{bmatrix}$$

A sample of our Data

$$x_i = (x_{i1}, \dots, x_{ir}, \dots, x_{id})^T \in \mathbb{R}^d$$

A feature vector

$$f_r = (x_{1r}, \dots, x_{ir}, \dots, x_{nr})^T \in \mathbb{R}^n$$

Laplacian Score (Unsupervised)

Similarity Matrix

$$W = \begin{bmatrix} 1 & w_{12} & \dots & w_{1n} \\ w_{21} & 1 & \dots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \dots & 1 \end{bmatrix}$$

Similarity between two samples

$$w_{ij} = S(x_i, x_j)$$

For example :

$$S(x_i, x_j) = \exp(-\frac{||x_i - x_j||^2}{2\sigma^2})$$

Laplacian Score (Unsupervised)

Degree Matrix

$$D = \begin{bmatrix} d_{11} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & d_{nn} \end{bmatrix}$$

Where :

$$d_{ii} = \sum_{j=1}^{n} w_{ij}$$

Laplacian Score (Unsupervised)

Laplacian Matrix

$$L = D - W$$

Laplacian Score of feature r

$$L_r = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} (x_{ir} - x_{jr})^2 s_{ij}}{\sum_{i=1}^{n} (x_{ir} - \bar{f}_r) p_i}$$

Where :

$$pi = \frac{d_i}{\sum_{k=1}^{n} d_k}$$

And we have :

$$L_r = \frac{f_r^T L f_r}{f_r^T D f_r}$$

Constraint Score A (Semi Supervised)

Constraints

$$\mathcal{M} = \{(x_i, x_j) \in X \times X \,|\, \text{such that } x_i \text{ and } x_j \text{ belong to the same class}\}$$

$$\mathcal{C} = \{(x_i, x_j) \in X \times X \,|\, \text{such that } x_i \text{ and } x_j \text{ belong to different classes}\}$$

Binary Constraint Matrices

$$w_{ij}^{\mathcal{M}} = \begin{cases} 1 & \text{if } (x_i, x_j) \in \mathcal{M} \text{ or } (x_j, x_i) \in \mathcal{M} \\ 0 & \text{else} \end{cases}$$

$$w_{ij}^{\mathcal{C}} = \begin{cases} 1 & \text{if } (x_i, x_j) \in \mathcal{C} \text{ or } (x_j, x_i) \in \mathcal{C} \\ 0 & \text{else} \end{cases}$$

Constraint Score A (Semi Supervised)

Minimize

$$\sum_{i=1}^{n}\sum_{j=1}^{n}(x_{ir}-x_{jr})^2 w_{ij}^{\mathcal{M}}$$

Maximize

$$\sum_{i=1}^{n}\sum_{j=1}^{n}(x_{ir}-x_{jr})^2 w_{ij}^{\mathcal{C}}$$

Constraint Score A (Semi Supervised)

Laplacian Matrix with constraints

$$L^{\mathcal{M}} = D^{\mathcal{M}} - W^{\mathcal{M}} \quad \text{and} \quad L^{\mathcal{C}} = D^{\mathcal{C}} - C^{\mathcal{C}}$$

Where :

$$D_{ii}^{\mathcal{M}} = \sum_{j=1}^{n} w_{ij}^{\mathcal{M}} \quad D_{ii}^{\mathcal{C}} = \sum_{j=1}^{n} w_{ij}^{\mathcal{C}}$$

Constraint Score A of feature r

$$SC_r^A = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} (x_{ir} - x_{jr})^2 w_{ij}^{\mathcal{M}}}{\sum_{i=1}^{n} \sum_{j=1}^{n} (x_{ir} - x_{jr})^2 w_{ij}^{\mathcal{C}}} = \frac{f_r^T L^{\mathcal{M}} f_r^T}{f_r^T L^{\mathcal{C}} f_r^T}$$

Constraint Score B (Semi Supervised)

Constraint Score B of feature r

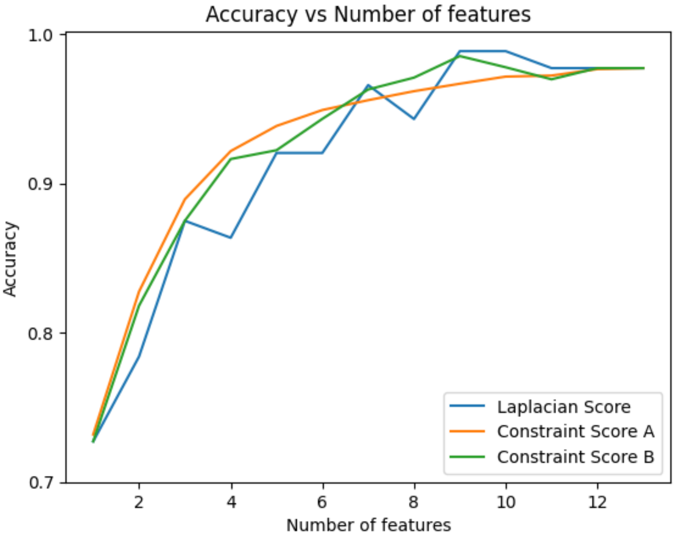
$$SC_r^B = \frac{f_r{}^T L f_r}{f_r{}^T D f_r} \cdot \frac{f_r{}^T L^{\mathcal{M}} f_r}{f_r{}^T L^{\mathcal{C}} f_r} = SL_r \cdot SC_r^A$$

Experimental Results

Wine Database

- 178 samples characterized by 13 features (n=178, d=13)
- 3 classes (k=3)
  - 59 class 1
  - 71 class 2
  - 48 class 3
- We select 30, 36, and 24 instances from each class to constitute the training set.
- 1-NN classifier to measure accuracy
- 9 labels avaiable (3 prototypes per class)
  - $k \times p \times (p-1)$ Must link constraints
  - $k \times (k-1) \times p$ Cannot link constraints

Experimental Results



Experimental Results

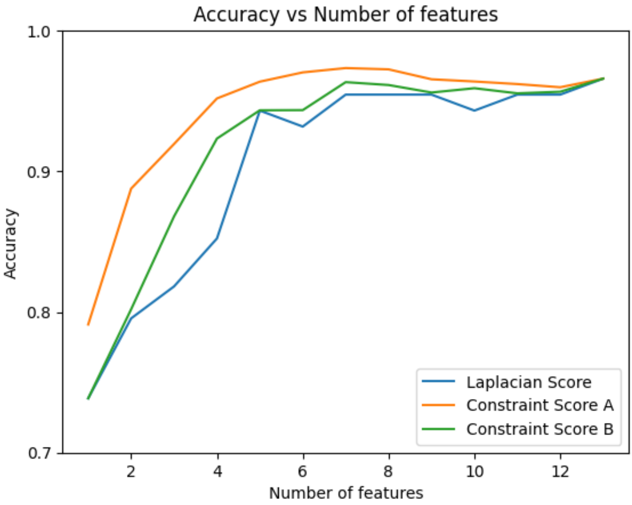


Figure: For 6 labels available per class

Conclusion

Perspectives

- Measure different criteria (ex : sensitivity to constraints)
- Implement the Similarity Based Constraint Score (SBCS) as described by [2]
- Compare results of the SBCS with the previous scores
- Improve the SBCS by using constraints directly instead of available labels to generate the constraints

2

[2] Abderezak Salmi, Kamal Hammouche, and Ludovic Macaire. "Similarity-Based Constraint Score for Feature Selection". In: *Knowledge-Based Systems* 209 (Dec. 2020), p. 106429. ISSN: 09507051. DOI: 10.1016/j.knosys.2020.106429. (Visited on 09/23/2023).