**CHALLENGE 1**

INTRODUZIONE

In questa challenge dobbiamo studiare un database contenente informazioni su immagini di banconote.

Il database è ordinato rispetto alla variabile obbiettivo, il che può creare problemi quando andiamo a selezionare il test-set se non facciamo attenzione a prendere casualmente i punti. Inoltre, poiché usiamo metodi basati su distanze (PCA, T-SNE, K-MEANS, K-NN …), risulta utile fare dello scaling.

UNSUPERVISED LEARNING

Come primo approccio vediamo se una proiezione con PCA ci aiuta nel clustering con K-MEANS.

La proiezione di PCA con le sole prime 2 componenti non è linearmente separabile.

Immagine che contiene mappa, testo, diagramma

Il contenuto generato dall'IA potrebbe non essere corretto.Immagine che contiene schermata

Il contenuto generato dall'IA potrebbe non essere corretto.Inoltre K-MEANS sia con 2 che con tutte le componenti non riesce a separare bene i 2 gruppi, probabilmente perché gli insiemi non sono convessi.

Andiamo ora a vedere come t-SNE con DBSCAN trattino i dati. Notiamo che t-SNE separa le classi in 2 zone dense ben separate da un vuoto centrale. Infatti DBSCAN separa correttamente i 2 gruppi principali, con il solo problema di separare in altri sottogruppi il cluster viola.

Immagine che contiene mappa, diagramma

Il contenuto generato dall'IA potrebbe non essere corretto.

SUPERVISED LEARNING

-PERFORMANCE COMPARISON

Come possiamo vedere dalla tabella tutti i modelli performano bene.

Decision Tree e K-NN sono i migliori visto che i dati sono stati scalati e le relazioni sono non lineari (ulteriore conferma dei risultati trovati con PCA).

Logistic Regression anche funziona molto bene, questo significa che a dimensioni più alte i dati sono linearmente separabili. Naïve Bayes è quello che funziona meno, poiché c’è una forte correlazione tra le variabili. Le prime 3 componenti di PCA infatti hanno il 96% di varianza spiegata. Come miglioria proporrei quindi di usare le prime 3 componenti di PCA per Naïve Bayes

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Logistic Regression | Decision Tree | Naïve Bayes | Naïve Bayes-PCA | K-NN |
| Accuracy | 0.97 | 0.99 | 0.85 | 0.90 | 1.00 |
| Precision | 0.97 | 0.99 | 0.85 | 0.90 | 1.00 |
| Recall | 0.97 | 0.99 | 0.85 | 0.90 | 1.00 |
| F1-score | 0.97 | 0.99 | 0.85 | 0.90 | 1.00 |

-REGULARIZATION ANALISYS

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Classica | Ridge | | Lasso | | EL-Net | |
|  |  | Lambda = 1 | Lambda = 0.001 | Lambda = 1 | Lambda = 0.001 | Lambda = 1 | Lambda = 0.001 |
| Accuracy | 0.97 | 0.86 | 0.97 | 0.44 | 0.97 | 0.85 | 0.97 |
| Precision | 0.97 | 0.86 | 0.97 | 0.44 | 0.97 | 0.86 | 0.97 |
| Recall | 0.97 | 0.86 | 0.97 | 0.44 | 0.97 | 0.85 | 0.97 |
| F1-score | 0.97 | 0.86 | 0.97 | 0.44 | 0.97 | 0.85 | 0.97 |

Come notato precedentemente nelle considerazioni sulla regressione logistica le classi sono linearmente separabili. Un learning rate troppo alto può provocare quindi un under-fitting.

Notiamo come i valori migliori corrispondano a quelli senza regolarizzazione. E i grafici delle loss ci confermano l’impatto quasi nullo della regolarizzazione.

|  |  |  |
| --- | --- | --- |
| Classica | Ridge | |
|  | |  | |
| Lasso | | **EL-Net** | |
|  | |  | |

CONSIDERAZIONI FINALI

Dallo studio fatto con K-MEANS e DBSCAN notiamo che i cluster hanno una forma concava, e quindi non riconoscibile da K-MEANS, DBSCAN invece riesce a separare bene le 2 classi.

I modelli predittivi funzionano tutti bene, ad eccezione di Naïve Bayes per il quale serve un piccolo accorgimento data la correlazione tra le variabili.

La regolarizzazione non ha un effetto particolarmente positivo, anzi, si rischia un under-fitting dei pesi e quindi un peggioramento dei risultati