Report del progetto di Machine Learning

Ferrario Tommaso Matr. 869005 (@TommasoFerrario18)

Terzi Telemaco Matr. 865981(@Tezze2001)

Vendramini Simone Matr. 866229(@Svendra4MySelf)

18 gennaio 2024

Indice

Capitolo 1

Introduzione

Questo è un progetto per l'esame di Machine Learning del primo anno del corso di laurea magistrale in informatica dell'Università degli Studi di Milano-Bicocca.

L'intero progetto si basa sul riconoscimento della presenza di un tumore al cervello data l'immagine di una risonanza magnetica. Il dataset scelto per questo progetto è scaricabile dal seguente link.

Per il riconoscimento del tumore sono stati allenati i seguenti modelli di machine learning:

- **SVM**: è stato scelto questo modello vista la buona capacità teorica nel geralizzare.
- **Naive Bayes Gaussiano**: è stato scelto questo modello dal momento che è l'unico ad essere probabilistico.
- **Rete neurale**: è stato scelto questo modello per confrontare i primi due con una soluzione neurale.

La relazione è stata suddivisa nei seguenti capitoli:

- **Introduzione**: descrizione del dominio e presentazione dei modelli che verranno presi in considerazione per questo progetto.
- **Dataset**: descrizione di come è stato costruito il dataset a partire dalle immagini, ovvero come sono state ricavate le features, e analisi esplorativa.
- **Rete neurale**: descrizione e analisi delle performance del percettrone.
- **SVM**: descrizione e analisi delle performance delle SVM.
- **Naive Bayes Gaussiano**: descrizione e analisi delle performance per Naive Bayes Gaussian.
- **Analisi dei risultati**: analisi comparata dei risultati tra i tre modelli considerati.
- **Conclusioni**: conclusioni sull'elaborato.

Capitolo 2

Dataset

Il dataset è stato a partire da un set di 3762 immagini ottenute dalla risonanza magnetica del cervello di altrettante persone, ettichettato manualmente da professionisti del settore nelle rispettive classi:

- **presenza del tumore**: $T = 1$
- **assenza del tumore**: $T = 0$

Il valore della label cade sotto al colonna *Class*.

Le features del dataset sono state ottenute calcolando i **momenti Hu** sulle immagini della risonanza magnetica. I momenti Hu catturano le informazioni di base sull'immagine come l'area dell'oggetto, il centroide, l'orientamento e altre proprietà.

Le feature sulle immagini si dividono in base a 2 gruppi[1]:

- **First Order Features**: forniscono informazioni legate alla distribuzione dei livelli di grigio dell'immagine. Queste features corrispondono alle statistiche descrittive calcolate sui valori di ciascun pixel dell'imamgine:
  - **media**
  - **varianza**
  - **deviazione standard**
  - **indice di asimmetria**
  - **indice di kurtosis**
- **Second Order Features**: forniscono informazioni a livello di composizione della texture dell'immagine.
  - **contrast**: misura la variazione locale dei livelli di grigio dei pixel. Maggiore sarà il valore allora maggiore sarà il contrasto dell'immagine.
  - **energy**: misura l'etereogenità o la variazione dell'intensità dell'immagine. Più piccolo è il valore allora meno variazioni di intesità e più omogenea sarà la texture, al contrario, più la texture è irregolare allora maggiore sarà il valore.
  - **ASM**: misura come sono distribuiti uniformemente i livelli di grigio nell'immagine. Più uniformi e distribuiti sono i livelli di grigio allora minore sarà il valore dell'indice.
  - **entropy**: misura la randomicità dei livelli di grigio, quindi più piccolo è il valore, più la texture sarà uniforme.
  - **homogeneous**: misura secondaria del contrasto. Più alto sarà l'indice allora minore sarà il contrasto dell'immagine.
  - **dissimilarity**: misura quanto spesso differenti combinazioni dei valori di intensità dei pixel occorrono nell'immagine. Un valore alto dell'indice indica che l'immagine ha una maggior variazione delle intesità dei pixel vicini, quindi più complessa sarà la texture.
  - **correlation**: misura la correlazione tra pixel nelle due diverse direzioni.
  - **coarseness**: misura quanto l'immagine è composta da regioni di intensità omogenea, ovvero quanto è granulare o fine la texture.

2.1 Analisi descrittiva

Il dataset è formato da un totale composto da un totale 3762 esempi, ciascuno descritto da 15 attributi: 13 legati alle feature (presentate precedentemente) calcolate sull'immagine associata, mentre l'attributo *Image* specifica l'id dell'immagine a cui si riferisce l'esempio, infine, l'attributo *Class* specifica l'etichetta dell'esempio.

Quando è stato importato il dataset nel dataframe, è stato opportuno controllare che il tipo inferito in automatico sulla colonna fosse corretto, di conseguenza è stata convertita tutta la colonna associata alla label in *categorico*, infine, tutti gli altri attributi sono stati mantenuti di tipo *float*, essendo valori continui. Per altro si evidenza l'assenza di valori nulli, evitando di cancellare i record associati.

Successivamente sono state calcolate le statistiche descrittive del dataframe.

TODO:immagine

Ciò ha permesso innanzitutto di notare che i dati non sono standardizzati dal momento che nessun dato ha media 0 e deviazione standard 1. La standardizzazione delle feature è stata fatta, utilizzando l'equatione 2.1, solo per le SVM e il modello neurale, mentre per Gaussian Bayes la standardizzazione viene fatta in automatico dalla libreria del modello.

$$F_{\mu,\sigma} = \frac{F - \mu}{\sigma} \tag{2.1}$$

In aggiunta, per ogni feature è stato disegnato il barplot per poter visionare se le distribuzioni sono normali.

TODO:immagine

Dai grafici si evince che le feature *Energy*, *Homogeneity* e *Coarseness* non seguono una distribuzione normale, mentre le altre si possono assumere normali. In ogni caso anche se non seguono l'ipotesi di normalità abbiamo deciso in ogni ciaso di utilizzarli, anche se stiamo andando contro le assunzioni dei modelli che vogliamo utilizzare.

Dalle statistiche descrittive e dal barplot si può notare come la feature di *Coarseness* assume valori molto bassi, quindi è stato pensato di convertire questa feature ad una scala logaritmica, permettendo di aumentare la significatività dei valori. In ogni caso si può notare che, anche con questa operazione, si ha comunque una deviazione standard pari a circa 0.02, valore non particolarmente significativo, quindi questo ci permette di escludere questa feature perché sicuramente non sarà la feature più discriminante.

Una delle operazioni preliminari di analisi del dataset è controllare il bilanciamento delle etichette di ciascun esempio, in questo modo è possibile valutare se il dataset in esame è buono per essere utilizzato per l'apprendimento supervisionato. In questo caso, disegnando lo scatter plot delle label, si può notare che le classi sono bilanciate, infatti, circa il 55% degli esempi sono negativi (assenza del tumore), contro circa il 45% degli esempi sono positivi (presenza del tumore).

TODO:immagine

Successivamente risulta importante analizzare se le distribuzioni si avvicinano ad una normale standard, di conseguenza sono stati disegnati i barplot delle frequenze di ciascuna feature. Dai grafici si evince che tutte le feature eccetto *Energy* e *Homogeneity* sono distribuite normalmente.

Dal barplot delle features creato precedentemente, si possono ottenere ulteriori informazioni sulla distribuzione potenzialmente normale delle altre feature. Infatti, la distribuzione della *Standard deviation* è simmentrica, mentre tutte le altre sono assimmetriche.

Ricapitolando, da questo primo studio descrittivo è stato modificato il dataset rimuovendo la feature *Coarseness*.

TODO:immagine

2.1.1 Analisi delle correlazioni

Il passo successivo è calcolare le correlazioni tra le feature, in modo tale da ridurre la dimensionalità dei dati considerando solo una delle feature che sono tra di loro una a una correlate.

Si potrebbe, innanzitutto, cominciare dalla forte correlazione positiva tra **media** e **variabilità**, più precisamente tra le feature *Mean*, *variance* e *standard deviation*. La correlazione tra varianza è deviazione standard è banale ($SD[P] = \sqrt{VAR[P]}$), invece, la correlazione tra media e variabilità può essere indotta da come sono composte le immagini. Più precisamente analizzando le immagini prodotte dalle risonanze magnetiche, si evince che, essendo in bianco e nero, se la media tende a 1 (colore bianco) allora deve aumentare la variabilità, perché ci sono diversi pixel bianchi allora le transizioni dal nero assoluto al bianco assoluto neccessitano di regioni di pixel maggiore rispetto ad una transizione tra nero assoluto e grigio (0.5).

TODO:immagine

Una seconda forte correlazione positiva è tra le feature che misurano l'**uniformità dei livelli di grigio** dei pixel, più precisamente tra le feature *Entropy*, *ASM*, *Homogeneity* ed *Energy*. Queste feature quantificano delle informazioni legate alla texture dell'immagine, quindi la forte correlazione positiva può essere facilmente spiegata

analizzando le texture delle immagini su cui vengono calcolate. Più precisamente se si ha un valore molto alto della feature *Entropy*, significa che la texture non è uniforme, ovvero si hanno strutture complesse e irregolari, quindi meno uniforme sarà la distribuzione dei livelli di grigio, aumentando l'indice di *ASM*, comportando di conseguenza un aumento delle variazioni di intensità dei livelli di grigio, aumentando di conseguenza anche l'indice di *Energy*, infine, (**cosa c'entra Homogeneity?**).

Al tempo stesso la matrice di correlazione evidenza una forte correlazione positiva tra gli indici che misurano la **morfologia della distribuzione**, ovvero le feature di *Skewness* e *Kurtosis*. Questa dipendenza implica il fatto che più la distribuzione è leptokurtica (Kurtosis grande), ovvero la frequenza dei livelli di grigio dei pixel si concentrano interamente vicino alla media/mediana/ moda, allora più grande sarà la Skewness, ovvero maggiore sarà la tendenza ad avere frequenze di livelli di grigio più vicino al bianco (coda di destra più altra rispetto alla coda di sinistra).

La matrice della correlazione evidenzia anche una correlazione positiva tra le feature di *Contrast* e *Dissimilarity*, ovvero maggiore sarà il contrasto e maggiore sarà la complessità della texture.

In aggiunta dalla matrice si evidenza che le features di *Dissimilarity* e *Homogeneity* sono correlate negativo, ovvero maggiore sarà il contrasto allora minore è la complessità della texture.

Dalla correlazione delle features è possibile ridurre la dimensionalità del dataset considerando sono le seguenti features:

- **Mean**
- **Entropy**
- **Skewness**
- **Contrast**
- **Correlation**

A puro scopo didattico è stato pensato di eseguire i modelli non solo sul dataset semplificato eliminando le correlazioni, ma anche applicando l'algoritmo PCA.

2.2 PCA

Bibliografia

[1] Namita Aggarwal e RK Agrawal. “First and second order statistics features for classification of magnetic resonance brain images”. In: (2012).