Progetto di Machine Learning corso del professor Candelieri

Algoritmi per la classificazione dei raggi gamma

Lorenzo Meroni 875319

Univerisita' degli studi Milano Bicocca

June 20, 2022

Abstract

Magic gamma telescope data 2004 https://archive.ics.uci.edu/ml/datasets/MAGIC+Gamma+Telescope è un dataset che raccoglie particelle gamma ad alta energia acquisite tramite un telescopio gamma Cherenkov. Le informazioni vengono registrate direttamente da terra, inoltre sappiamo che il telescopio basa il proprio funzionamento sulla tecnica detta Imaging Atmospheric Cherenkov Technique (tecnica molto usata per la sua grande versatilità e per la possibilità di effettuare analisi in maniera non distruttiva).
Lo scopo dello studio è prevedere, in base a particolari feature, se ogni raggio presente nel dataset è raggio gamma (segnale) oppure è frutto di adronic shower, causata da raggi cosmici nell'alta atmosfera.
Verranno sperimentati vari algoritmi di classificazione al fine di verificare mediante l'utilizzo di metriche come l'Accurancy, la Sensitivity e l'Auc (area under the curve) quale sia il miglior algoritmo, e per quale motivo un determinato algoritmo classifichi la variabile risposta in maniera migliore rispetto agli altri.

Keywords
Gamma Telescope, Algoritmi di classificazione, accurancy e AUC

1 Introduzione

I raggi gamma non raggiungono mai la superficie terrestre, il telescopio gamma Cherenkovo utilizza la rivelazione della cosiddetta luce Cherenkov che si genera nell'interazione dei raggi gamma con l'atmosfera, infatti la luce Cherenkov si genera tutte le volte che una particella carica si muove in un mezzo con una velocità più alta di quella della luce.
I raggi gamma ad alta energia, entrando in contatto con l'atmosfera terrestre producono cascate di particelle subatomiche altamente energetiche, che possono viaggiare a velocità superiore a quella della luce nell'aria, dando luogo a un debole e brevissimo (nell'ordine del miliardesimo di secondo) lampo di luce blu, questo è il cosiddetto "effetto Cherenkov".

Dall'analisi delle immagini ottenute si ricavano informazioni descrittive dei raggi gamma primari. È questa in sintesi la tecnica detta "Imaging Atmospheric Cherenkov Technique" (IACT). Per questo tipo di telescopi sono state adottate finora configurazioni ottiche che prevedono un solo specchio in cui la luce viene riflessa per essere catturata direttamente dalla camera di rilevazione. In genere, dopo una preelaborazione consente di estrarre esclusivamente l'area luminosa (quest'area solitamente assume la forma di un ellisse) rappresentata da un sottoinsieme di pixel dell'immagine.
I parametri caratteristici di questa ellisse (spesso chiamati parametri di Hillas) sono i parametri caratteristici dell'area in esame che possono essere utilizzati per determinare se il raggio sia un raggio gamma o meno.
In particolare è stato condotto uno studio sull'utilizzo di tecniche di machine learning al fine di prevedere se appunto in base a queste variabili un raggio sia raggio gamma o non sia raggio gamma.

2 Data Set

Il dataset contiene 10 feature ed una variabile risposta binaria "class" (che assume "g" True quando il raggio è raggio gamma mentre assume valore "h" quando è hadron ovvero non gamma). Le 10 feature descrivono la struttura e le variazioni cromatiche dell'elisse o raggio luminoso. Una volta estratta dall'immagine la parte luminosa attraverso tecniche di elaborazione immagini vengono calcolate queste feature. Di seguito vengono riportate le feature estratte con una rapida descizione:

- **fLength**: misura dell'asse maggiore dell'ellisse espressa in millimetri [mm].
- **fWidth**: misura dell'asse minore dell'ellisse espressa in millimetri [mm].
- **fSize**: trasformazione logaritmica in base 10 della somma di tutti i pixel contenuti.
- **fConc**: rapporto della somma dei due pixel piu' alti rispetto a *fSize* [ratio].
- **fConc1**: rapporto del pixel piu' alto rispetto a *fSize* [ratio].
- **fAsym**: distanza dal pixel più alto al centro, proiettata sull'asse maggiore dell'ellisse [mm].
- **fM3Long**: radice terza del momento terzo lungo l'asse maggiore (dove il momento terzo è una particolare media dell'intensità dei pixel) [mm].
- **fM3Trans**: radice terza del momento terzo lungo l'asse minore [mm].
- **fAlpha**: angolo tra l'asse maggiore e il vettore di origine [deg].
- **fDist**: distanza dall'origine al centro dell'ellisse [mm].
- **class**: g,h variabile binaria, dove g sta per gamma (segnale) mentre h sta per hadron (sfondo).

3 Analisi esplorativa

Il dataset è composto da 10 variabili di cui una la variabile risposta *class* e da 19020 osservazioni ed è stato ripartito in training set e test set con una proporzione di 75/25.
La strategia adottata in questo progetto, di conseguenza, è stata di applicare la cross validation sul primo insieme, dividendolo ogni volta in training set e validation set, al fine di stabilire i migliori iperparametri per ogni algoritmo.
Spostiamo ora l'attenzione sulla variabile risposta invece, il rapporto tra raggi gamma e pioggia adronica è 2/3 a 1/3, si è deciso in questo caso di non ribilanciare le classi anche per non intaccare la capacità previsiva dei modelli. Per ciascuna delle variabili abbiamo esaminato i grafici delle distribuzioni di densità condizionate per le due classi della variabile *class*, e successivamente applicato il test t di Welch test noto anche come test t per varianze disuguali, è un test di localizzazione a due campioni che viene utilizzato per verificare l'ipotesi per cui le medie ottenute dalle distribuzioni condizionate alla variabile risposta fossero significativamente diverse tra loro. Solo per la variabile *fM3Trans* e *fConc1* si acetta l'ipotesi tale per cui le medie non siano statisticamente diverse in quanto p-value > 0.05, di seguito vengono riportata la tabella relativa al Welch test solo per *fConc1* e *fM3Trans*:

Variabile	T-value	df	p-value
fConc1	0.33771	9357	0.5612
fM3Trans	0.21914	6137.3	0.6397

Table 1: Risultati Welch test

Mentre per quanto riguarda le distribuzioni condizionate:

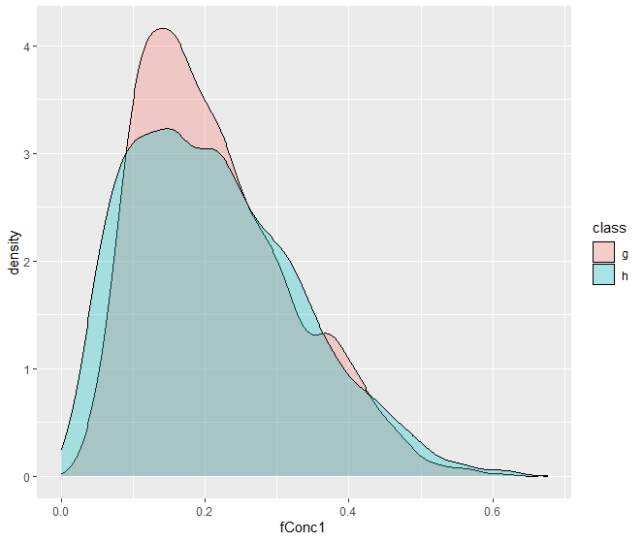

Figure 1: Distribuzione fConc1 condizionata alla variabile risposta class

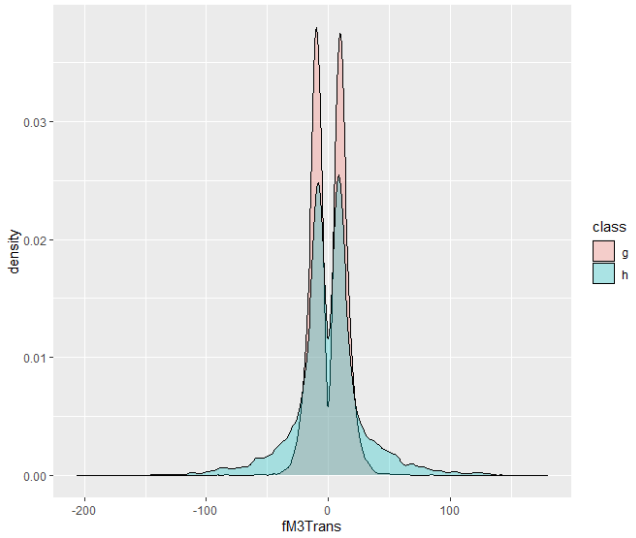

Figure 2: Distribuzione fM3Trans condizionata alla variabile risposta class

Per valutare la correlazione a livello globale, abbiamo calcolato la matrice di correlazione delle variabili esplicative, dalla quale si deduce che vi sia una forte correlazione positiva tra *fLength*,*fWidth* e *fSize*, mentre allo stesso tempo una forte correlazione tra *fConc* e *fConc1*.

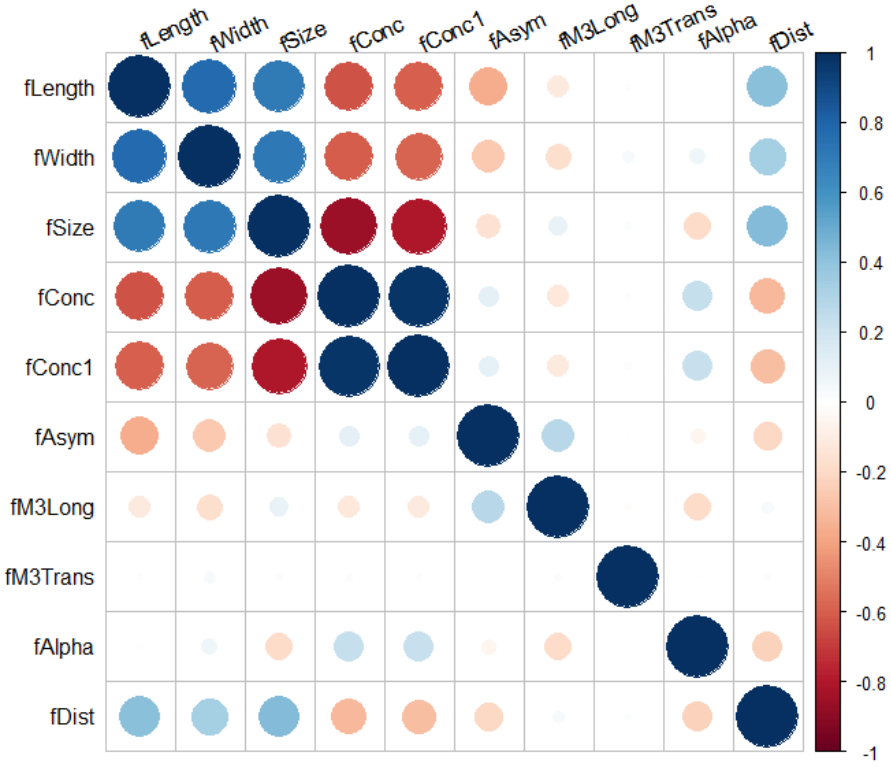

Figure 3: Correlogramma

Si è persino pensato oltre alla semplice analisi esplorativa dei dati di implementare una rapida random forest al fine di verificare e nel caso selezionare esclusivamente le varibili importanti mediante l'utilizzo di variable importance di random forest, il risultato è stato il seguente:

Tuttavia, siccome il numero di variabili non è eccessivo, non sono presenti particolari criticità nonostante alcune variabili siano fortemente correlate tra di loro e inoltre la variable importance della random forest ha effettivamente mostrato che *fAlpha* soprattutto ma anche *fLenght* e *fWidth* siano le più determinanti nella classificazione, allo stesso tempo si può notare che non esista una variabile che non contribuisca in alcun modo alla determinazione del risultato finale.
Pertanto alla luce del fatto che i migliori risultati in termini di classificazione vengono raggiunti mediante l'utilizzo di tutte le variabili e non esista in motivo evidente per eliminarne alcune si è deciso di utilizzare tutte le variabili per l'applicazione degli algoritmi di classificazione.

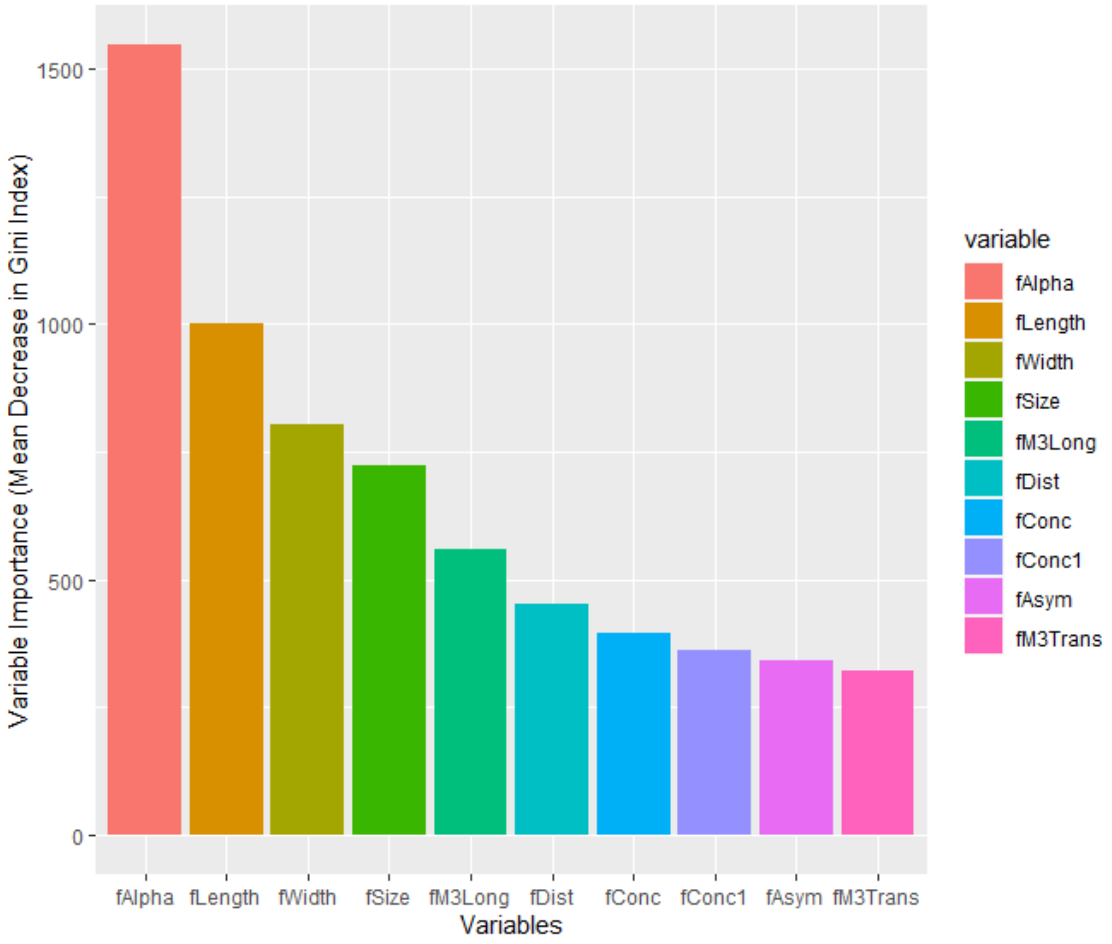

Figure 4: Variable importance di Random Forest

4 Metriche e modelli usati

Per il problema in oggetto, sono state quindi scelte, come metriche per valutare i risultati delle applicazioni dei vari algoritmi, la "Accuracy" e la "Sensitivity" e oltre a queste anche l'area under the curve, l'auc è necessaria in questo caso perche' la semplice accurancy (gamma classificati correttamente sommati a hadronic shower classificata correttamente su tutti i dati da classificare) non da una completa visione della performance di un algoritmo per questi dati, poiché classificare un evento di hadronic shower come raggio gamma è peggio che classificare un raggio gamma come hadronic shower. Ciascun modello è stato prima valutato sul dataset originale dopo la fase di pre-processing, La validazione degli iper-parametri, è stata effettuata attraverso il metodo della k-fold cross validation sul training set, attraverso una 10-folds cross validation.
I modelli usati per l'analisi sono:

- K-nearest neighbour
- Random forest (RF).
- SVM lineare.

- SVM con kernel radiale.

K-nearest neighbour

Il primo modello utilizzato è stato il K-nearest neighbour (Knn) per via della sua semplicità che tuttavia non è però sempre sinonimo di cattivi risultati. L'iperparametro di interesse riguarda il numero di punti k vicini presi in esame dall'algoritmo per l'assegnazione della classe, che nel nostro caso ha assunto valori da 1 a 40. Abbiamo dedotto che il k ottimale, considerando come metrica di riferimento il tasso di accurancy, è = 7 (individuato tramite il metodo della 10 fold cross-validation):

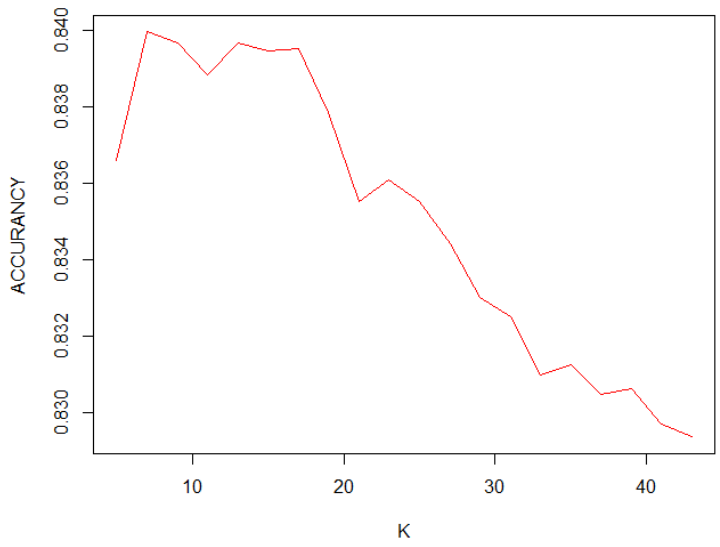

Figure 5: Accurancy al variare del numero dei k vicini

Sul test set, il modello KNN così costruito ha riportato un valore dell'Accuracy pari a 0.8355, mentre per la Sensitivity 0.9458 mentre il valore dell'auc è di 0.789, l'algoritmo funziona abbastanza bene soprattutto sul lato della Sensitivity (i risultati ottenuti sul test set replicano quelli stimati via cross validation).

Random Forest

La random forest con motore ranger prevede l'ottimizzazione di alcuni iperparametri tra cui m-try che rappresenta il numero di variabili su cui eventualmente dividere in ciascun nodo che è stato scelto considerando come metrica di riferimento il tasso di accurancy e il migliore è risultato uguale a 2, e la split-rule ovvero la regola decisionale che permette all'albero di decidere che divisione adottare, sempre secondo la metrica dell'accurancy è stato scelto il criterio di gini, mentre l'iperparametro min.node.size è stato tenuto costante e pari a 1.
Di seguito la tabella con i valori degli iperparametri e i relativi valori di accurancy per i primi 5 tra cui è presente anche il migliore:
Appunto con il primo set di iperparametri il modello random forest classificativo ha

MTRY	SPLITRULE	ACCURANCY
2	gini	0.8784444
2	extratrees	0.8723451
3	gini	0.8766217
3	extratrees	0.8727660
4	gini	0.8765519

Table 2: Iperparametri Random Forest e relativi valori di accurancy

ottenuto un valore di accurancy sul test set pari a 0.878 e un valore di Sensitivity pari a 0.9439 e un valore dell'auc pari a 0.8502. Risulta quindi abbastanza chiaro che l'algoritmo random forest classifichi in maniera più precisa di Knn (questo è tendenzialmente dovuto alla sua maggiore capacità previsiva).
Sempre relativo alla random forest e' stato testato un approccio diiferente da quello appena illustrato, utilizzando come features le prime 7 componenti risultanti da una PCA sui dati in esame, tuttavia questo metodo ha portato ad un risultato più scarso rispetto a quello appena illustrato pertanto si è deciso di scartarlo.

Support Vector Machine Lineare

L'SVM lineare è arrivato a convergenza ma ha una performance peggiore sia di KNN che della random forest, il valore dell'iperparametro C che rappresenta la penalità per i punti classificati erroneamente è stato scelto tra una griglia di valori *0.01, 0.1, 1, 10* pari a 1 in quanto minimizza l'errore di classificazione. Il valore ottenuto di accurancy è pari a 0.7918 mentre sensitivity dell' 0.8984 e un'area under the curve di 0.7468.

Support Vector Machine con kernel Radiale

L'SVM con kernel radiale ha una performace significativamente migliore rispetto a quella con il kernel lineare, il valore dell'iperparametro C è stato scelto tra una griglia di valori *0.01, 0.1, 1* pari a 1 mentre gamma che controlla la distanza dell'influenza di un singolo punto di addestramento è stato scelto tra una griglia di valori *0.1, 0.5, 1, 2* pari a 0.1. Si è ottenuto così un valore dell'Accuracy pari a 0.8677, per la Sensitivity 0.9539 mentre il valore dell'auc é di 0.8313 sintomo del fatto che i dati siano lineamente separabili ma una trasformazione kernel permetta di classificare in maniera più proficua.

5 Conclusione

Abbiamo valutato 4 algoritmi differenti tra loro, utilizzandoli sul dataset originale derivante dall'analisi esplorativa.
È stato riscontrato che gli algoritmi più efficaci sono stati gli algoritmi più complessi come Random Forest e support vector machine con kernel radiale, a discapito di modelli più semplici quali il K-Nearest Neighbour. In particolare l algoritmo che ha fornito il miglior risultato in termini di Accuracy e Auc é stata la Random Forest seguita dalla SVM radiale.

MODELLO	ACCURANCY	SENSITIVITY	AUC
Knn	0.8355	0.8784	0.789
Random Forest	0.8784	0.9439	0.8502
Svm lineare	0.7918	0.8984	0.7468
Svm radiale	0.8677	0.9539	0.8313

Table 3: Riepilogo modelli usati e risultati ottenuti