Challenge 0

Data processing

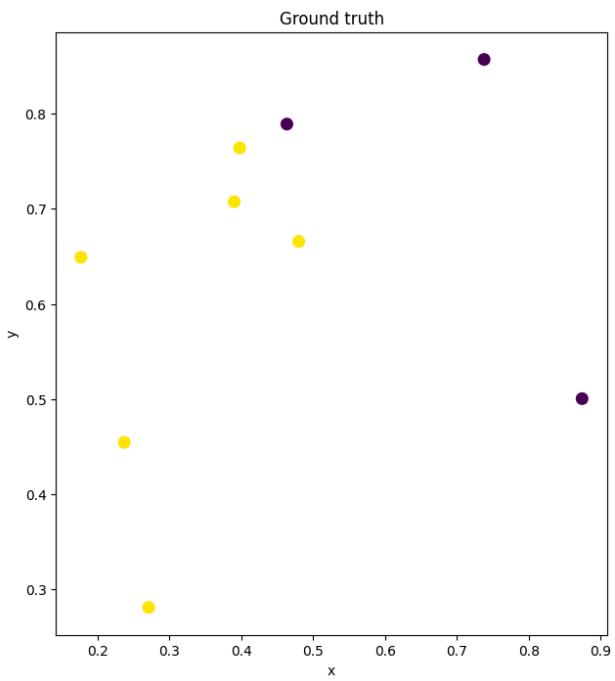
Nella prima parte del lavoro ci occupiamo della preparazione dei dati.
Usiamo il pacchetto “pandas" di python per caricare il dataset da analizzare e ripulirlo per poi procedere a farci un’idea del contenuto usando funzioni che ci permettono di esplorare il dataset.
Il dataset iniziale è composto da 5 variabili (R&D Spend, Administration, Marketing Spend, State, Profit) che hanno ognuna 50 osservazioni.
In questo caso abbiamo sostituito i valori uguali a 0 con la media della colonna a cui appartengono, per non avere valori nulli e facilitare l’analisi.
Abbiamo dunque selezionato due categorie della variabile State per la classificazione binaria e utilizzato il One-Hot Encoding per ottenere le corrispondenti variabili dummy, questo, in modo da rendere la variabile categoriale facilmente interpretabile dagli algoritmi che andremo ad utilizzare.

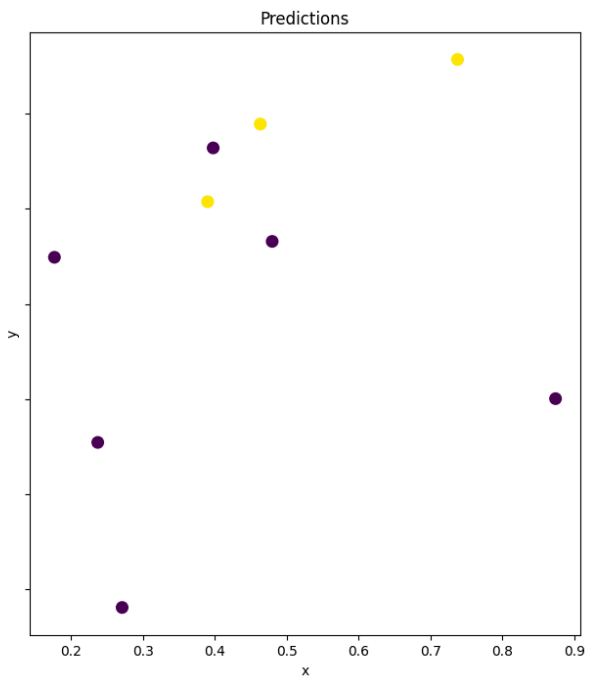
Dopo aver fatto ciò abbiamo costruito il nostro dataset finale che useremo per learning e prediction nella nostra Logistic Regression, abbiamo quindi normalizzato il dataset in modo che le features avessero tutte valori compresi tra 0 e 1.

Splitting dei dati

Una volta ottenuto il dataset su cui vogliamo lavorare è il momento di dividerlo in Training set e Test set, abbiamo scelto di avere il 75% dei dati nel Training e il rimanente 25% nel Test set.

Procedendo nel nostro lavoro usiamo la funzione built-in LogisticRegression del pacchetto sklearn.linear_model per fare il training del modello, impostando un seed a 0 per la riproducibilità dei dati, calcoliamo anche l’accuracy, misura della correttezza generale delle predizioni del nostro modello per tutte le classi, che in questo caso è molto bassa (pari a 0.2222) e i valori delle y predette.





Nei due grafici sopra riportati notiamo rispettivamente (da sinistra a destra) i punti ottenuti utilizzando la Ground truth, ossia le effettive classi che avremmo voluto predire con il nostro modello, e i valori ottenuti in predizione nella fase di training.

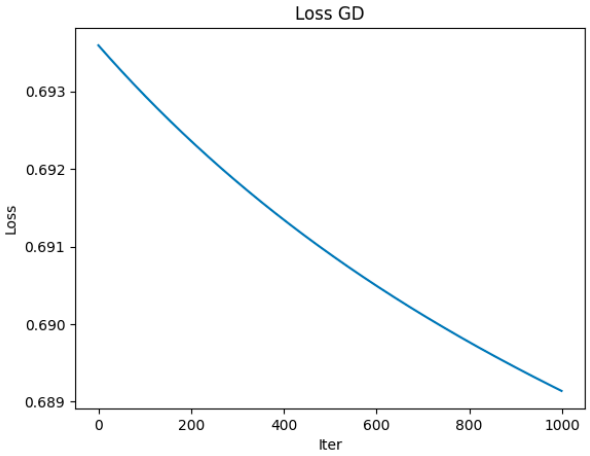
Possiamo notare come il modello scelto non faccia un buon lavoro in predizione, potrebbe non essere la scelta più opportuna per un dataset che ha poche osservazioni come quello preso in considerazione; probabilmente la mancanza di dati sarà risultata in una carenza di esempi da cui il modello può imparare pattern in modo efficiente.

Regolarizzazione

Procediamo utilizzando tecniche di regolarizzazione sulla Logistic Regression.
Le tecniche di regolarizzazione sono di solito usate per prevenire l'overfitting e migliorare il modello, solitamente introducono vincoli o penalties al processo di learning in modo tale da rendere il modello più semplice e robusto.

Prima di implementare le funzione di regressione regolarizzata definiamo la funzione sigmoide, utilizzata nella Logistic Regression per modellare la probabilità che un certo input appartenga a una determinata classe, e la Logistic Loss, che quantifica quanto sono ben stimate le predizioni del modello di Logistic Regression rispetto alle true labels in un problema di classificazione binaria.

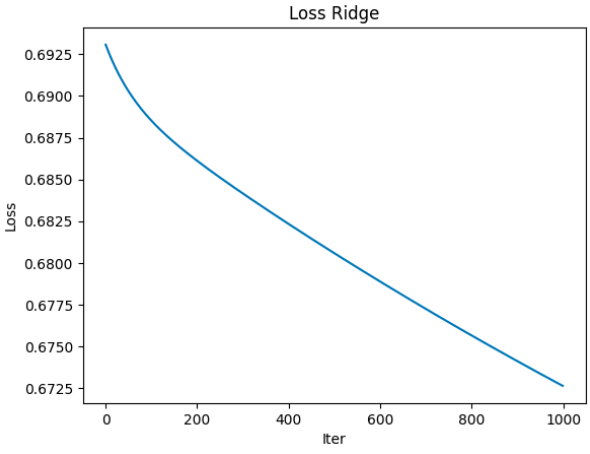
Gradient Descent



Gradient Descent è un algoritmo che viene utilizzato per minimizzare la cost function nella regressione logistica ed è utilizzato in modo iterativo per aggiornare i parametri che minimizzano la suddetta funzione.

Il grafico a sinistra è stato ottenuto con i risultati della Logistic Regression regolarizzata con la tecnica Gradient Descent.

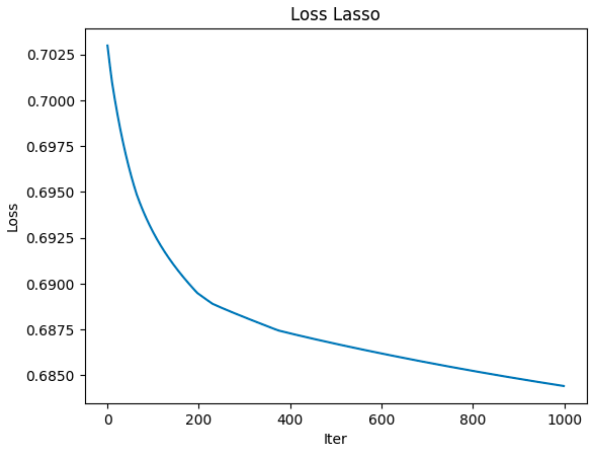
Ridge



La Ridge regularization è una tecnica che introduce un termine di regolarizzazione alla cost function della Logistic Regression. L'obiettivo di questa regolarizzazione è evitare l'overfitting aggiungendo un penalty term che tende a far evitare dei coefficienti grandi per il modello di regressione logistica.

Il grafico a sinistra è stato ottenuto con i risultati della Logistic Regression regolarizzata con la tecnica Ridge.

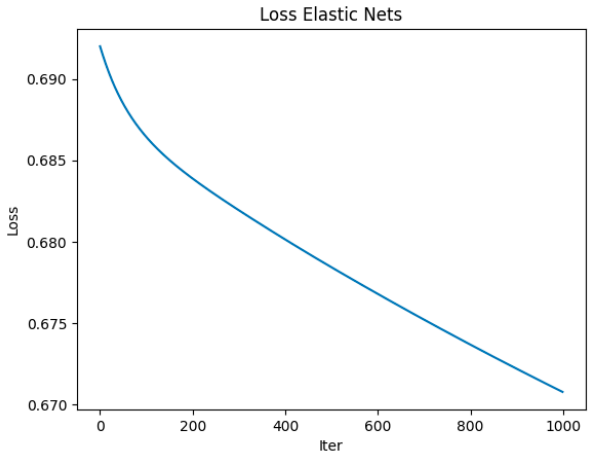
LASSO (Least Absolute Shrinkage and Selection Operator)



La regolarizzazione LASSO introduce un termine di penalty alla cost function della Logistic Regression in modo da evitare la selezione di features non necessarie.

Il grafico a sinistra è stato ottenuto con i risultati della Logistic Regression regolarizzata con la tecnica LASSO.

Elastic Nets



La regolarizzazione Elastic Nets è una combinazione della Ridge e della LASSO regularization, infatti aggiunge i termini di penalty di ambedue le precedenti alla cost function della Logistic Regression.

Il grafico a sinistra è stato ottenuto con i risultati della Logistic Regression regolarizzata con la tecnica Elastic Nets.

Generalmente i Loss plot vengono utilizzati per visualizzare la performance di un modello in training e iniziare a capire quanto bene il modello stia imparando dai dati.
Possiamo notare che in tutti i grafici la loss diminuisce man mano che si procede con le iterazioni, ossia, diminuisce mentre il modello impara a classificare le istanze in modo più accurato, in particolare notiamo che nella LASSO diminuisce più velocemente nelle prime iterazioni rispetto a quanto faccia negli altri modelli.

Model assessment

	precision	recall	f1-score	support
California	0.17	0.33	0.22	3
Florida	0.33	0.17	0.22	6
accuracy			0.22	9
macro avg	0.25	0.25	0.22	9
weighted avg	0.28	0.22	0.22	9

A sinistra troviamo le classification metrics per la Logistic Regression senza regolarizzazione.

	precision	recall	f1-score	support
California	0.33	1.00	0.50	3
Florida	0.00	0.00	0.00	6
accuracy			0.33	9
macro avg	0.17	0.50	0.25	9
weighted avg	0.11	0.33	0.17	9

A sinistra troviamo le classification metrics per la Logistic Regression con regolarizzazione Gradient Descent.

	precision	recall	f1-score	support
California	0.38	1.00	0.55	3
Florida	1.00	0.17	0.29	6
accuracy			0.44	9
macro avg	0.69	0.58	0.42	9
weighted avg	0.79	0.44	0.37	9

A sinistra troviamo le classification metrics per la Logistic Regression con regolarizzazione Ridge.

	precision	recall	f1-score	support
California	0.38	1.00	0.55	3
Florida	1.00	0.17	0.29	6
accuracy			0.44	9
macro avg	0.69	0.58	0.42	9
weighted avg	0.79	0.44	0.37	9

A sinistra troviamo le classification metrics per la Logistic Regression con regolarizzazione LASSO.

	precision	recall	f1-score	support
California	0.38	1.00	0.55	3
Florida	1.00	0.17	0.29	6
accuracy			0.44	9
macro avg	0.69	0.58	0.42	9
weighted avg	0.79	0.44	0.37	9

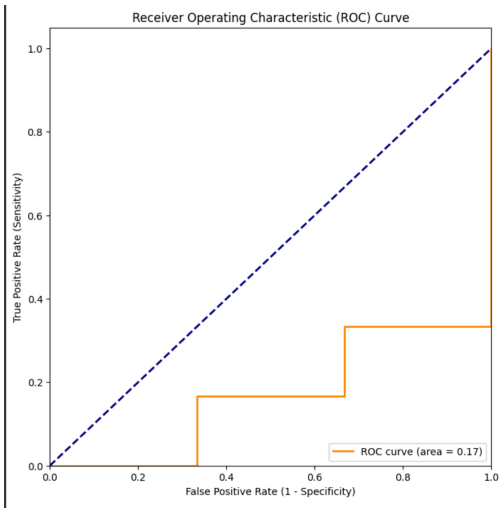
A sinistra troviamo le classification metrics per la Logistic Regression con regolarizzazione Elastic Net.

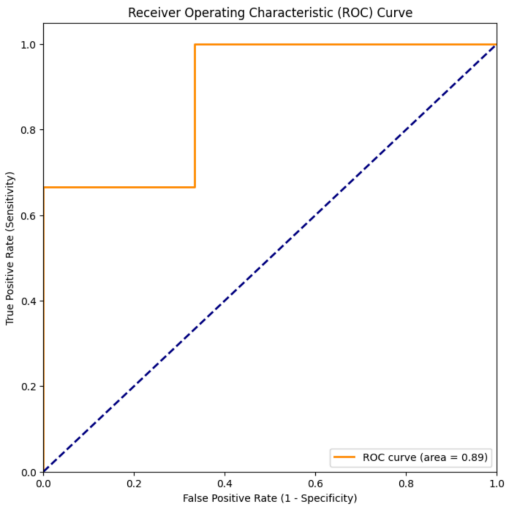
Dai dati sopra riportati notiamo che per il modello senza regolarizzazione la precision, metrica che indica la percentuale di predizioni corrette, è abbastanza bassa per entrambe le categorie e ciò accade anche nel modello regolarizzato con gradient descent, in cui addirittura abbiamo una precision di 0 per la classe Florida; nei modelli con regolarizzazione Ridge, LASSO e Elastic Net notiamo comunque una precision bassa per la classe California, ma una precision molto alta per la classe Florida, ciò si può riscontrare nella media, che infatti risulta prossima allo 0.70.

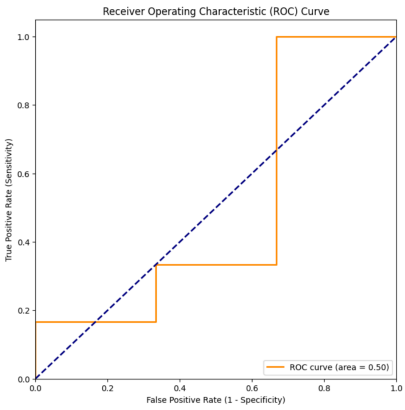
Per quanto riguarda la metrica recall, che indica la frazione di positivi che sono stati identificati correttamente, notiamo che, per il modello non regolarizzato le percentuali sono basse, ma, in questo caso, quella della classe California è più altra rispetto a quella della classe Florida, questo rimane vero anche per gli altri modelli, con un aumento significativo fino a 1 della recall per la classe California.

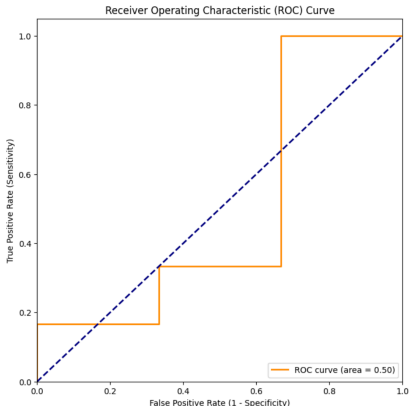
La metrica F1-score è uguale per entrambe le classi nel modello senza regolarizzazione, mentre è pari a 0.50 per la classe California e 0 per la classe Florida (prevedibile visto che precisione e recall sono entrambe pari a 0) per il modello in cui abbiamo usato la Gradient descent, per gli altri modelli i valori sono pari a 0.55 per California e 0.29 per Florida. La media di questa metrica è generalmente usata per comparare i modelli, più vicino a 1 è il valore, migliore sarà il modello. Possiamo quindi dire che i modelli con Ridge, LASSO e Elastic Net, con una macro average dell' F1-score pari a 0.44, sono in media migliori rispetto ai primi due modelli considerati.

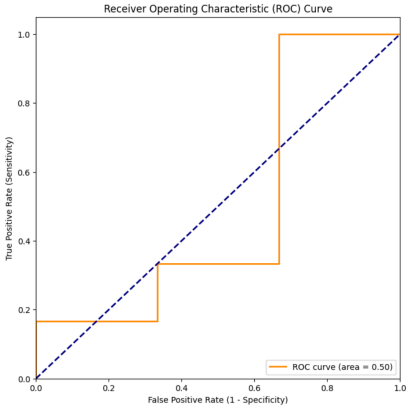
Il valore di support ci indica il numero di occorrenze di una specifica classe nel dataset utilizzato, un support poco bilanciato nel set di training dei dati potrebbe indicare una debolezza dal punto di vista strutturale negli scores del modello e potrebbe indicare il dover utilizzare sampling stratificati o rebalancing. Il valore di support non cambia tra i modelli ma serve a valutare l'evaluation process. Nel nostro caso abbiamo un support pari a 3 per la classe California e uno pari a 6 per la classe Florida, questo è un indicatore del fatto che potremmo dover tornare a lavorare sulla selezione del training set per ottenerne uno maggiormente bilanciato.











ROC curve

Risp. ROC Logistic regression, ROC Gradient Descent

Risp. ROC Ridge, ROC LASSO, ROC Elastic Nets

La ROC curve è il tracciato grafico del true positive rate e del false positive rate per vari valori di threshold e ci aiuta a capire l'abilità diagnostica di un modello di classificazione binaria per i diversi valori di soglia utilizzati. Di solito è usata per valutare la performance di classificatori, in particolare in casi in cui le classi non sono bilanciate.

Il valore AUC è una misura numerica della performance della ROC curve.
Quello che vorremmo ottenere è il valore più alto possibile di AUC ossia un valore massimizzato della area rotto la curva (un classificatore ideale avrà AUC pari a 1).
Si vuole quindi che la linea rappresentata sopra in arancione sia quanto più vicina al top-left corner quanto possibile.

Nel nostro caso notiamo che per i modelli dove è stata usata la regolarizzazione Ridge, LASSO e Elastic Nets l'area sotto la curva è pari a 0.5, abbiamo un valore molto alto, 0.89, per il modello con regolarizzazione con Gradient Descent e un valore basso 0.17 per il modello senza regolarizzazione.
Il fatto che il valore di AUC per i modelli con Ridge, LASSO e Elastic Nets sia 0.5 indica una poca abilità di discriminazione tra le due classi e il valore 0.17 per il modello semplice di Logistic

Regression indica, come ci aspettavamo anche dai grafici analizzati precedentemente che, da parte del modello, non c'è una buona abilità di distinzione tra le due classi.