UNIVERSITÀ DEGLI STUDI DI MILANO
FACOLTÀ DI SCIENZE E TECNOLOGIE

DIPARTIMENTO DI INFORMATICA
GIOVANNI DEGLI ANTONI

Corso di Laurea magistrale in Informatica

AGGIUNGERE

Relatore: Prof. Elena Casiraghi
Correlatore: Prof. Dario Malchiodi

Tesi di Laurea di:
Alessandro Beranti
Matr. Nr. 977702

ANNO ACCADEMICO 2021-2022

to do

Ringraziamenti

to do

Indice

Introduzione

Durante il periodo di tesi mi sono concentrato su

Capitolo 1

Stato dell'arte

1.1 Apprendimento automatico

Il termine apprendimento automatico, comunemente chiamato *Machine learning* in inglese, sta a indicare sta a indicare la capacità dei computer di apprendere e adattarsi agli input forniti.

1.1.1 Intelligenza artificiale vs Apprendimento automatico

Molto spesso il termine *Machine learning*, o apprendimento automatico in italiano, viene usato per indicare l'intelligenza artificiale e viceversa ma non sono la stessa cosa, il *Machine learning* è un sottoinsieme della categoria più ampia chiamata appunto Intelligenza artificiale.

L'intelligenza artificiale è il campo di computer, sistemi e robot in grado di simulare il comportamento umano in modi che imitano e spesso vanno oltre le capacità umane. I programmi di IA sono in grado di analizzare e fornire dati o attivare automaticamente azioni senza il bisogno dell'uomo. Alcuni esempi pratici sono l'elaborazione del linguaggio naturale e della visione artificiale per automatizzare e velocizzare alcune attività. (quali?)

Il *Machine learning* utilizza algoritmi per apprendere in maniera automatica intuizioni e riconoscere modelli a partire da dati forniti in input. Gli algoritmi di apprendimento automatico vengono divisi in quattro categorie distinte a seconda del tipo di dato usato per eseguire la fase di apprendimento, queste categorie sono le seguenti:

- apprendimento supervisionato,
- apprendimento non supervisionato,
- apprendimento non supervisionato,

- apprendimento per rinforzo.

1.1.1.1 Apprendimento supervisionato

Nell'apprendimento supervisionato abbiamo dei si ha un insieme x di N osservazioni $x_1, x_2, ..., x_N$ di un vettore avente p dimensioni che contengono esempi di come x sia in relazione con la variabile di output y, anche chiamata etichetta. Usando modelli matematici e statistici adattati ai dati di addestramento, x in questo caso, si vuole cercare di predire l'output y, per dati "nuovi", ovvero dati che il modello non ha usato nella fase di addestramento. L'approccio usato dall'apprendimento supervisionato per "imparare" è quello di estrapolare la relazione che sussiste tra x e y, ovvero imparare usando osservazioni reali. Esistono diversi modi per valutare quanto il modello è riuscito a "imparare bene" e ci danno una stima di quando sia stato in grado di generalizzare, l'argomento verrà approfondito in 3.6.

Il tipo di output che si ricerca influenza il tipo di problema che si sta affrontando, se abbiamo un output numerico siamo di fronte a un problema di regressione mentre se abbiamo un output categorico siamo di fronte a un problema di classificazione. Una variabile numerica possiede un ordine naturale, se prendiamo un'istanza della variabile siamo in grado di dire se sia più grande o piccola di un'altra istanza della stessa variabile. Una variabile numerica può essere rappresentata da un numero reale continuo come da un numero discreto. Le variabile categoriche invece sono sempre discrete e sono prive di un ordine, un esempio è il seguente:

1.1.1.2 Apprendimento non supervisionato

Nell'apprendimento non supervisionato si ha un insieme di N osservazioni $x_1, x_2, ..., x_N$ di un vettore avente p dimensioni ma, al contrario dell'apprendimento supervisionato, l'insieme di dati non ha un'etichetta, i dati sono quindi non annotati. Nel contesto dell'apprendimento supervisionato ci sono diverse metriche quanto bene il modello è stato in grado di imparare. Nel contesto dell'apprendimento non supervisionato non esiste una vera e propria misura diretta del successo di un algoritmo. È molto difficile accertare la validità delle inferenze tratte dall'output. A questo scopo si deve ricorrere a euristiche per valutare la qualità dei risultati ottenuti. Alcune delle tecniche più famose sono calcolare le regole di associazione, usare l'algoritmo "Apriori"

1.1.1.3 Apprendimento per rinforzo

1.1.2 Apprendimento semi-supervisionato

1.1.2.1 Machine learning in Bioinformatica

Per dati omici si intendono i dati provenienti da esperimenti di geno- mica, trascrittomica, epigenomica, metage- nomica, metabolomica e proteomica

1.1.3 Decision Tree Classifier

Gli alberi decisionali sono classificatori che predicono l'etichetta della classe di appartenenza per un insieme di dati. Il vantaggio di questo tipo di classificatori sta nella loro semplicità. Essi sono costruiti analizzando un insieme di dati di addestramento

1.1.4 Random Forest Classifier

Capitolo 2

Dataset

- CNV (Copy Number Variations): per ogni gene viene indicato se lo stesso ha subito delezione (i.e. perdita di una o più copie), amplificazione (acquisizione di una o più copie), neutro (nessuna modifica). Si possono trovare al suo interno i seguenti valori: 0 (neutro), 1 o 2 (amplificazioni), -1 o -2 (delezioni) - miRNA (micro-RNA): livelli di espressione di una particolare tipologia di RNA nota come micro-RNA. Si tratta di piccoli RNA (non codificanti per proteine) che svolgono funzioni di regolazione all'interno della cellula. I livelli di espressione sono ottenuti tramite una tecnica nota come RNA-sequencing. - mRNA (RNA messaggeri): livelli di espressione di un altro tipo di RNA noti come RNA messaggeri. Gli mRNA sono fondamentali all'interno della cellula poichè codificano per le proteine, permettendo alle informazioni contenute nel DNA (che non può uscire dal nucleo della cellula) di arrivare nel citoplasma dove ci sono i ribosomi che leggono i trascritti di mRNA e producono le proteine. Anche questi dati sono ottenuti tramite RNA-sequencing. - proteine: livelli di espressione delle proteine. Le proteine sono le macromolecole che svolgono essenzialmente ogni compito all'interno della cellula (formano gli enzimi che catalizzano ogni reazione nella cellula, trasportano le macromolecole all'interno della cellula, sintetizzano le macromolecole, etc). Questi dati sono ottenuti con una tecnica nota come RPPA (Reverse Phase Protein Array).

- il file un dataset noto come TCGA-CDR (https://www.sciencedirect.com/science/article/pii/S0
senzialmente un dataset curato manualmente per avere dati clinici e di sopravvivenza il più affidabili possibile). Nello specifico, le etichette fanno riferimento ad una misura nota come PFI (Progression Free Interval) dove:

1 stands for patient having new tumor event whether it was a progression of disease, local recurrence, distant metastasis, new primary tumors all sites , or died with the cancer without new tumor event, including cases with a new tumor event whose type is N/A. 0 otherwise.

2.1 The Cancer Genome Atlas (TCGA)

2.1.1 Proteins

2.1.2 mRNA

2.1.3 miRNA

2.1.4 Cnv

Capitolo 3

Esperimenti

3.1 Preprocessing

3.1.1 Scalare i dati

3.2 Feature selection

Nel machine learning e in statistica, con il termine *feature selection* si intende il processo di selezione di un sottoinsieme di *feature*, anche chiamate caratteristiche o dimensioni rimuovendo *feature* irrilevanti, ridondanti o che producono solo rumore. Questa pratica di solito porta a una migliore capacità di addestramento, accuratezza più elevata, minore costo di computazione e aumento dell'interpretabilità del modello. La feature selection aiuta anche a non incappare nel *curse of dimensionality*.

Negli ultimi anni i dati disponibili per applicazioni di machine learning in ambiti come mining di testo, computer vision e biomedico stanno aumentando esponenzialmente sia in termini di campioni sia in termini di numero di dimensioni. L'enorme numero di feature dei dataset attualmente disponibili porta a diversi svantaggi: rallentamento significativo degli algoritmi di *learning*, peggiorare la performance dei suddetti algoritmi ma anche portare a una difficile interpretazione del modello. Le tecniche di feature selection possono essere classificate in tre famiglie: metodi supervisionati, metodi semi-supervisionati e metodi non supervisionati.

3.2.1 Tecniche univariate

3.2.1.1 Bassa variabilità

3.2.1.2 Mann-Whitney

The Mann-Whitney U test is a nonparametric test of the null hypothesis that the distribution underlying sample x is the same as the distribution underlying sample y. It is often used as a test of difference in location between distributions.

3.2.2 Tecniche multivariate

3.2.2.1 Minimum Redundancy Maximum Relevance: mrmr

3.2.2.2 Boruta

3.2.3 Dimensionalità intrinseca

3.2.3.1 ID_twoNN

3.2.4 Maximal information-based nonparametric exploration (MINE)

3.2.4.1 The maximal information coefficient (MIC)

3.2.5 Spearman

3.3 Feature extraction

Il termine di *feature extraction*, o estrazione delle caratteristiche, si riferisce al processo di trasformazione dei dati grezzi in caratteristiche numeriche che possono essere elaborate preservando le informazioni nel set di dati originale. Produce risultati migliori rispetto all'applicazione dell'apprendimento automatico direttamente ai dati grezzi.

3.3.1 Uniform Manifold Approximation: umap

3.3.2 t-SNE

3.4 Model selection

3.4.1 Tuning degli iperparametri

3.5 Cross Validation

3.6 Metrica di performance

Le metriche di performance sono molto importanti in un processo di *machine learning*. Esse ci indicano se stiamo facendo progressi nella creazione del modello che meglio si adatta ai dati in input. Esistono diverse metriche che possono essere usate a seconda dei problemi cui siamo davanti. Se stiamo trattando un problema di regressione, avente quindi output continuo, dobbiamo calcolare in qualche modo la distanza tra il dato predetto e quello originale; per fare ciò possiamo usare diverse metriche: *Mean absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE),* R^2 *(R-Squared).*

Siccome il problema affrontato è un problema di classificazione entreremo più nel dettaglio in questo argomento. I modelli di classificazione hanno un output discreto quindi abbiamo bisogno di metriche che comparino classi discrete. Le metriche di classificazione valutano le prestazioni di un modello e ti dicono quanto è buona o cattiva la classificazione, ma ognuna di esse la valuta in modo diverso. Esistono diverse metriche:

- accuratezza,
- matrice di confusione,
- precision e recall
- f1-score,
- au-roc.

3.6.0.0.1 Accuratezza L'accuratezza è la metrica più semplice da usare e implementare. Essa non è altro che il numero di predizioni che il modello ha fatto correttamente diviso per il totale di predizioni, moltiplicato per 100 per avere la percentuale.

3.6.0.0.2 Marice di confusione La matrice di confusione non è propriamente una metriche ma è molto utile per definire le altre metriche.

3.6.1 Dati sbilanciati

3.6.2 Area sotto la curva precision-recall

L'area sotto la curva precision-recall è un singolo numero che riassume l'informazione della curva precision-recall a diverse soglie. La curva PR viene sempre più usata nel *machine learning*, nei problemi di classificazione, sopratutto quando si ha a che fare con *datasets* sbilanciati, ovvero dove una classe è molto più frequente dell'altra. (aggiungi riferimento a come è composto il mio dataset). In questi contesti la curva PR è da preferire alla curva ROC

3.7 Risultati

3.8 Tecnologie usate

Capitolo 4

Conclusioni e sviluppi futuri

Bibliografia

[1] Andreas Lindholm, Niklas Wahlström, Fredrik Lindsten, and Thomas B. Schön. *Machine Learning - A First Course for Engineers and Scientists*. Cambridge University Press, 2022.

[2] Carl Kingsford and Steven L. Salzberg. What are decision trees? *Nature Biotechnology*, 26(9):1011–1013, Sep 2008.