

Progetto di Metodi statistici per l'apprendimento: K-Nearest Neighbors

Studente: Marco Odore - Matricola: 868906

Abstract

Implementazione in R e C++ dell'algoritmo K-nearest neighbors e sua applicazione ad un problema di classificazione binaria, nella sua versione classica e online.

Keywords

Apprendimento Supervisionato — K-NN — Classificazione binaria

Contents

Introduction	1
0.1 Il dataset	1
0.2 Ottimizzazione delle performance	2
1 Metodi e valutazione dei modelli	2
1.1 K-NN	2
1.2 K-NN online	2
1.3 Nested Cross-Validation	2
1.4 Rischio sequenziale	2
2 Results and Discussion	2
2.1 Subsection	2
Subsubsection • Subsubsection • Subsubsection	
2.2 Subsection	3
Acknowledgments	4
References	4

Introduzione

Si è scelto di implementare in R l'algoritmo di apprendimento supervisionato K-NN e di applicarlo ad un problema di classificazione binario, dove si vuole associare ad una persona il suo possibile reddito annuale (discretizzato, tramite una soglia, a due valori possibili), date alcune sue informazioni di base.

Il dataset

Il dataset utilizzato [1] contiene 48842 istanze, ognuna delle quali è caratterizzata da 14 feature:

- **Età** - Tipo continuo
- **Workclass** - Tipo nominale (8 valori possibili)
- **Fnlwgt** - Tipo continuo
- **Education** - Tipo nominale (16 valori possibili)

- **Education-num** - Tipo continuo (trasformazione di education in tipo continuo)
- **Marital-status** - Tipo nominale (7 valori possibili)
- **Occupation** - Tipo nominale (14 valori possibili)
- **Relationship** - Tipo nominale (6 valori possibili)
- **Race** - Tipo nominale (6 valori possibili)
- **Sex** - Tipo nominale (2 valori possibili)
- **Capital-gain** - Tipo continuo
- **Capital-loss** - Tipo continuo
- **Hours-per-week** - Tipo continuo
- **Native-country** - Tipo nominale (41 valori possibili)

Per rendere l'input gestibile da K-NN, si sono trasformate le feature di tipo nominale in una serie di feature binarie, e cioè utilizzando delle variabili dummy per ogni possibile valore che la feature può assumere¹.

Per quanto riguarda l'etichettatura, ognuna delle istanze può assumere due possibili valori, e cioè

- **classe 1:** $> 50k$
- **classe 2:** $\leq 50k$

Il dataset risulta essere sbilanciato, in quanto la classe 1 rappresenta il 23.93% del dataset (la classe 2 il 76.07%). Inoltre al suo interno vi sono alcune istanze che per alcune feature non possiede specificazioni. Per questa motivazione si è deciso di escluderle dal processo di valutazione ed apprendimento, riducendo così la cardinalità complessiva del dataset a 45222 istanze².

¹Con l'introduzione delle variabili dummy si è passati a 88 feature complessive.

²Senza le istanze con valori sconosciuti le percentuali delle classi 1 e 2 si attestano rispettivamente su 24.78% e 75.22%.

Ottimizzazione delle performance

Data la lentezza dell'algoritmo K-NN, soprattutto con grandi dataset, si è deciso di implementare la funzione di ricerca per i k più vicini in C++, che rispetto ad R risulta molto più veloce. Per quanto riguarda il task di *cross-validation* invece, si è deciso di parallelizzarlo grazie ad una funzione fornita da una libreria di R³, che ha permesso di suddividere agevolmente il problema.

1. Metodi e valutazione dei modelli

Per il problema trattato si sono utilizzate due versioni di K-NN, e cioè quella classica e quella online.

K-NN

Questo algoritmo permette di generare un classificatore h_{k-NN} memorizzando l'intero training set fornito in input e calibrando un suo parametro k , il quale permette di gestire l'underfitting e l'overfitting sul problema.

L'assegnazione dell'etichetta ad un'istanza x avviene nella seguente maniera:

$h_{k-NN}(x)$ = la maggioranza delle etichette y_i delle k istanze x_i più vicine a x

K-NN online

Si tratta di una variante di K-NN che in fase di training opera nella seguente maniera:

- Prova a classificare l'istanza x_t usando K-NN sull'insieme S .
- Se $h_{K-NN}(x_t) \neq y_t$ aggiungi x_t all'insieme S .

L'algoritmo inizia da un insieme S vuoto, aggiungendovi man mano esempi quando sbaglia, lasciandolo inalterato invece quando non commette errori.

Nested Cross-Validation

Per la stima delle performance dei classificatori generati dalla versione base di K-NN e quella online si è deciso di utilizzare la *cross-validation esterna* a 5 fold, sfruttando come metrica il *test error*, calcolato come segue:

$$\tilde{er}(h_{k-NN}) = \frac{1}{n} \sum_{i=1}^n l(y_i, h_{k-NN}(x_i))$$

Dove l è la funzione di perdita *zero-uno*, che vale 1 nel caso in cui il predittore commette un errore.

Per selezionare invece il k ottimale automaticamente, si è effettuata una *cross-validation* interna, sempre a 5 fold, sia nella fase di selezione (dove viene scelto il k finale del modello), sia per la selezione del k ottimale di ogni singolo

³*parLaply*, nella libreria di R *parallel*.

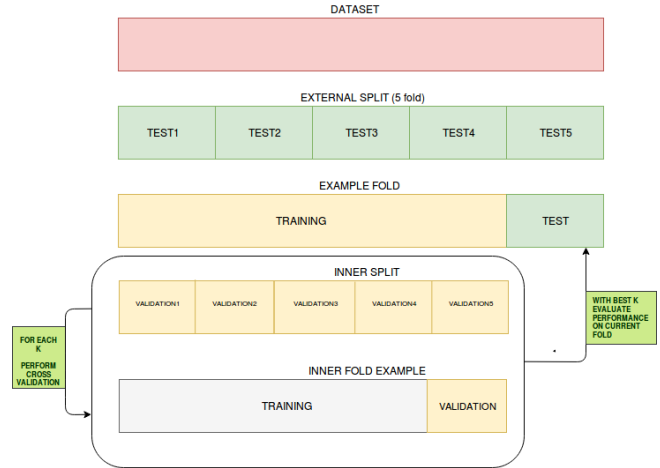


Figure 1. Il modo in cui è avvenuto lo split del dataset sia per la cross validazione esterna (split più esterno) sia per la cross validazione interna (split interno sui singoli training set di ogni rispettivo fold)

fold generato dalla cross validazione esterna. Nella figura 1 sono mostrati i passi della *nested cross-validation* (cross validazione interna + cross validazione esterna)

Rischio sequenziale

Nel caso specifico di K-NN online si è deciso inoltre di valutare l'andamento del rischio sequenziale sull'intero dataset, calcolato come segue:

$$er_{seq}(t) = \frac{1}{t} \sum_{i=1}^t l(y_i, h_{k-NN}(x_i))$$

ad ogni passo t , e cioè alla t -esima istanza fornita in pasto all'algoritmo.

2. Results and Discussion

Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetur odio sem sed wisi.

Subsection

Sed feugiat. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Ut pellentesque augue sed urna. Vestibulum diam eros, fringilla et, consectetur eu, nonummy id, sapien. Nullam at lectus. In sagittis ultrices mauris. Curabitur malesuada erat sit amet massa. Fusce blandit. Aliquam erat volutpat. Aliquam euismod. Aenean vel lectus. Nunc imperdiet justo nec dolor.

Table 1. Table of Grades

Name		
First name	Last Name	Grade
John	Doe	7.5
Richard	Miles	2

Subsubsection

Etiam euismod. Fusce facilisis lacinia dui. Suspendisse potenti. In mi erat, cursus id, nonummy sed, ullamcorper eget, sapien. Praesent pretium, magna in eleifend egestas, pede pede pretium lorem, quis consectetuer tortor sapien facilisis magna. Mauris quis magna varius nulla scelerisque imperdiet. Aliquam non quam. Aliquam porttitor quam a lacus. Praesent vel arcu ut tortor cursus volutpat. In vitae pede quis diam bibendum placerat. Fusce elementum convallis neque. Sed dolor orci, scelerisque ac, dapibus nec, ultricies ut, mi. Duis nec dui quis leo sagittis commodo.

Word Definition

Concept Explanation

Idea Text

Subsubsection

Aliquam lectus. Vivamus leo. Quisque ornare tellus ullamcorper nulla. Mauris porttitor pharetra tortor. Sed fringilla justo sed mauris. Mauris tellus. Sed non leo. Nullam elementum, magna in cursus sodales, augue est scelerisque sapien, venenatis congue nulla arcu et pede. Ut suscipit enim vel sapien. Donec congue. Maecenas urna mi, suscipit in, placerat ut, vestibulum ut, massa. Fusce ultrices nulla et nisl.

- First item in a list
- Second item in a list
- Third item in a list

Subsubsection

Etiam ac leo a risus tristique nonummy. Donec dignissim tincidunt nulla. Vestibulum rhoncus molestie odio. Sed lobortis, justo et pretium lobortis, mauris turpis condimentum augue, nec ultricies nibh arcu pretium enim. Nunc purus neque, placerat id, imperdiet sed, pellentesque nec, nisl. Vestibulum imperdiet neque non sem accumsan laoreet. In hac habitasse platea dictumst. Etiam condimentum facilisis libero. Suspendisse in elit quis nisl aliquam dapibus. Pellentesque auctor sapien. Sed egestas sapien nec lectus. Pellentesque vel dui vel neque bibendum viverra. Aliquam porttitor nisl nec pede. Proin mattis libero vel turpis. Donec rutrum mauris et libero. Proin euismod porta felis. Nam lobortis, metus quis elementum commodo, nunc lectus elementum mauris, eget vulputate ligula tellus eu neque. Vivamus eu dolor.

Subsection

Nulla in ipsum. Praesent eros nulla, congue vitae, euismod ut, commodo a, wisi. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Aenean nonummy magna non leo. Sed felis erat, ullamcorper in, dictum non, ultricies ut, lectus. Proin vel arcu a odio lobortis euismod. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Proin ut est. Aliquam odio. Pellentesque massa turpis, cursus eu, euismod nec, tempor congue, nulla. Duis viverra gravida mauris. Cras tincidunt. Curabitur eros ligula, varius ut, pulvinar in, cursus faucibus, augue.

Nulla mattis luctus nulla. Duis commodo velit at leo. Aliquam vulputate magna et leo. Nam vestibulum ullamcorper leo. Vestibulum condimentum rutrum mauris. Donec id mauris. Morbi molestie justo et pede. Vivamus eget turpis sed nisl cursus tempor. Curabitur mollis sapien condimentum nunc. In wisi nisl, malesuada at, dignissim sit amet, lobortis in, odio. Aenean consequat arcu a ante. Pellentesque porta elit sit amet orci. Etiam at turpis nec elit ultricies imperdiet. Nulla facilisi. In hac habitasse platea dictumst. Suspendisse viverra aliquam risus. Nullam pede justo, molestie nonummy, scelerisque eu, facilisis vel, arcu.

Curabitur tellus magna, porttitor a, commodo a, commodo in, tortor. Donec interdum. Praesent scelerisque. Maecenas posuere sodales odio. Vivamus metus lacus, varius quis, imperdiet quis, rhoncus a, turpis. Etiam ligula arcu, elementum a, venenatis quis, sollicitudin sed, metus. Donec nunc pede, tincidunt in, venenatis vitae, faucibus vel, nibh. Pellentesque wisi. Nullam malesuada. Morbi ut tellus ut pede tincidunt porta. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam congue neque id dolor.

Donec et nisl at wisi luctus bibendum. Nam interdum tellus ac libero. Sed sem justo, laoreet vitae, fringilla at, adipiscing ut, nibh. Maecenas non sem quis tortor eleifend fermentum. Etiam id tortor ac mauris porta vulputate. Integer porta neque vitae massa. Maecenas tempus libero a libero posuere dictum. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Aenean quis mauris sed elit commodo placerat. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Vivamus rhoncus tincidunt libero. Etiam elementum pretium justo. Vivamus est. Morbi a tellus eget pede tristique commodo. Nulla nisl. Vestibulum sed nisl eu sapien cursus rutrum.

Nulla non mauris vitae wisi posuere convallis. Sed eu nulla nec eros scelerisque pharetra. Nullam varius. Etiam dignissim elementum metus. Vestibulum faucibus, metus sit amet mattis rhoncus, sapien dui laoreet odio, nec ultricies nibh augue a enim. Fusce in ligula. Quisque at magna et nulla commodo consequat. Proin accumsan imperdiet sem. Nunc porta. Donec feugiat mi at justo. Phasellus facilisis ipsum quis ante. In ac elit eget ipsum pharetra faucibus. Maecenas viverra nulla in massa.

Nulla ac nisl. Nullam urna nulla, ullamcorper in, inter-

dum sit amet, gravis ut, risus. Aenean ac enim. In luctus. Phasellus eu quam vitae turpis viverra pellentesque. Duis feugiat felis ut enim. Phasellus pharetra, sem id porttitor sodales, magna nunc aliquet nibh, nec blandit nisl mauris at pede. Suspendisse risus risus, lobortis eget, semper at, imperdiet sit amet, quam. Quisque scelerisque dapibus nibh. Nam enim. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nunc ut metus. Ut metus justo, auctor at, ultrices eu, sagittis ut, purus. Aliquam aliquam.

Etiam pede massa, dapibus vitae, rhoncus in, placerat posuere, odio. Vestibulum luctus commodo lacus. Morbi lacus dui, tempor sed, euismod eget, condimentum at, tortor. Phasellus aliquet odio ac lacus tempor faucibus. Praesent sed sem. Praesent iaculis. Cras rhoncus tellus sed justo ullamcorper sagittis. Donec quis orci. Sed ut tortor quis tellus euismod tincidunt. Suspendisse congue nisl eu elit. Aliquam tortor diam, tempus id, tristique eget, sodales vel, nulla. Praesent tellus mi, condimentum sed, viverra at, consectetur quis, lectus. In auctor vehicula orci. Sed pede sapien, euismod in, suscipit in, pharetra placerat, metus. Vivamus commodo dui non odio. Donec et felis.

Etiam suscipit aliquam arcu. Aliquam sit amet est ac purus bibendum congue. Sed in eros. Morbi non orci. Pellentesque mattis lacinia elit. Fusce molestie velit in ligula. Nullam et orci vitae nibh vulputate auctor. Aliquam eget purus. Nulla auctor wisi sed ipsum. Morbi porttitor tellus ac enim. Fusce ornare. Proin ipsum enim, tincidunt in, ornare venenatis, molestie a, augue. Donec vel pede in lacus sagittis porta. Sed hendrerit ipsum quis nisl. Suspendisse quis massa ac nibh pretium cursus. Sed sodales. Nam eu neque quis pede dignissim ornare. Maecenas eu purus ac urna tincidunt congue.

Donec et nisl id sapien blandit mattis. Aenean dictum odio sit amet risus. Morbi purus. Nulla a est sit amet purus venenatis iaculis. Vivamus viverra purus vel magna. Donec in justo sed odio malesuada dapibus. Nunc ultrices aliquam nunc. Vivamus facilisis pellentesque velit. Nulla nunc velit, vulputate dapibus, vulputate id, mattis ac, justo. Nam mattis elit dapibus purus. Quisque enim risus, congue non, elementum ut, mattis quis, sem. Quisque elit.

Acknowledgments

References

- [1] M. Lichman. UCI machine learning repository. 2013. <https://archive.ics.uci.edu/ml/datasets/adult>