Università degli studi di Milano-Bicocca

Scuola di Economia e Statistica
Corso di Laurea Triennale in Scienze Statistiche ed Economiche

Proposta di tesi

Relatore:
Matteo Borrotti

Tesi di Laurea di:

Edoardo Olivieri
Matr. N. 880280

ANNO ACCADEMICO 2023/2024

Indice

1. Contesto

I metodi di resampling o ricampionamento si sono diffusi a partire dagli anni '60 grazie al rapido avanzamento della tecnologia in ambito di potenza di calcolo. Questi metodi rappresentano una valida alternativa alla statistica parametrica e un miglioramento rispetto alla statistica non parametrica. La necessità di sviluppare nuovi metodi per fare statistica inferenziale è emersa dagli evidenti limiti delle tecniche parametriche e non parametriche tradizionali. La statistica parametrica classica si basa sull'assunzione che la popolazione da cui è stato estratto il campione segua una distribuzione Normale nota, ad eccezione del valore dei parametri θ. Tuttavia, questa assunzione non è sempre verificata, specialmente in presenza di campioni non particolarmente ampi. In questi casi, le stime ottenute possono essere inaffidabili. D'altro canto, la statistica non parametrica, prima della diffusione dei generatori di numeri casuali e della rapida crescita della potenza di calcolo, soffriva di limiti dovuti all'eccessiva o talvolta errata approssimazione dei dati.
I metodi di ricampionamento, permettono di superare questi limiti ricampionando ripetutamente il set di dati a disposizione, al fine di stimare determinate caratteristiche della popolazione che ha dato origine al set di dati.

1.1 Plug-in principle

Il principio sulla quale si basano questi metodi è chiamato *plug-in principle* o *principio di sostituzione*. Si basa sull'idea che per risolvere problemi di statistica inferenziale, bisogna ricorrere alla stima di una distribuzione di ripartizione F sulla base di un sottocampione estratto casualmente da F. La funzione di ripartizione empirica, che chiameremo $\hat{F}$, è una stima della funzione di ripartizione originale F. La funzione $\hat{F}$ viene ottenuta costruendo una distribuzione di frequenze di tutti i valori che essa può assumere in una determinato set di dati (Efron B. e Tibshirani R.J. (1993)).

2. Problema

La stima accurata dell'errore di previsione è fondamentale sia per i modelli di regressione che per i problemi di classificazione. Questo processo è cruciale per valutare l'affidabilità dei modelli predittivi e per prendere decisioni informate basate sui risultati di tali modelli.

2.1 Loss functions

Per il calcolo dell'errore di previsione si utilizzano delle Loss functions (funzioni di perdita) che permettono di misurare la differenza tra il valore della risposta per un'osservazione futura (y) e la previsione ottenuta dal modello ($\hat{y}$). In modelli di regressione, delle possibili funzioni di perdita possono essere l'errore assoluto

$$L(y,\hat{y}) = |(y-\hat{y})|, \tag{2.1}$$

o l'errore quadratico

$$L(y,\hat{y}) = (y-\hat{y})^2, \tag{2.2}$$

Per problemi di classificazione invece, dove la risposta cade in una delle possibili K non-ordinate classi, si utilizza la funzione indicatrice

$$L(y,\hat{y}) = I[\hat{y} \neq y], \tag{2.3}$$

mentre l'errore di classificazione è definito come la probabilità di classificare incorrettamente una osservazione futura

$$PE = P(\hat{y} \neq y), \tag{2.4}$$

chiamato anche misclassification rate.

2.2 Splitting

Il motivo per cui (2.4) si calcola sulle osservazioni future è che, per calcolare l'errore di previsione reale, occorre avere delle osservazioni non precedentemente usate per allenare il classificatore. Se

si decidesse di calcolare l'errore di classificazione sugli stessi dati su cui è stato allenato il modello, si otterrebbe un valore troppo ottimista. Per ovviare a questo problema, il dataset iniziale viene comunemente diviso in training set e test set. Inoltre, il training set può essere ulteriormente suddiviso in subtraining set e validation set. Le osservazioni contenute all'interno del training set sono quelle utilizzate per allenare il classificatore, mentre le osservazioni del test set non vengono utilizzate durante la fase di training, e simulano dunque delle osservazioni future. Questo approccio permette di ottenere una stima più realistica delle performance del modello. La ricerca e l'ottimizzazione del modello viene fatta sui dati del subtraining e poi testata sui dati del validation. Una volta scelto il modello che porta a un errore di previsione sul validation più basso, si procede unendo il training set con il validation set per poi testare il modello finale sul test set e ottenere così l'errore di previsione (true error).

2.3 Dilemma

Ogni osservazione contiene al suo interno delle informazioni preziose. Nel caso di dataset grandi con centinaia o migliaia (o più) osservazioni, escluderne una parte per calcolare il vero errore di previsione può risultare ragionevole. Tuttavia, il problema/dilemma nasce quando i dati a disposizione non sono molti. In questi casi, non ci si può permettere di togliere informazioni potenzialmente utili dal training set. Non essendo quindi in possesso di osservazioni non utilizzate per allenare il modello, occorre ottenere una stima dell'errore di previsione. Questo dilemma è particolarmente rilevante nei contesti ad alta dimensionalità, come i dati genomici o di microarray, dove il numero di variabili supera di gran lunga il numero di osservazioni. In questi casi, l'esclusione di dati dal training set può ridurre significativamente la capacità del modello di generalizzare su nuovi dati, aumentando il rischio di overfitting (Cawley G.C. e Talbot N.L.C. (2010)). Per affrontare questo problema, sono stati sviluppati diversi metodi di ricampionamento, come il bootstrap e la cross-validation, che permettono di stimare l'errore di previsione utilizzando al meglio tutte le osservazioni disponibili. Questi metodi offrono una soluzione robusta per la stima dell'errore di previsione anche in situazioni di scarsità di osservazioni disponibili.

3. Stato Dell'Arte

3.1 Errore apparente

Prima della nascita degli stimatori plug-in, lo stimatore usato per l'errore di previsione era l'errore apparente, o errore di sostituzione. Questa stima, che utilizza gli stessi dati sia per allenare il modello che per calcolarne l'errore, risulta essere distorta positivamente, soprattutto quando i dati a disposizione sono pochi. Questo avviene perché il modello è testato sugli stessi dati che ha visto durante l'allenamento, portando a una stima eccessivamente ottimistica delle sue performance. Una semplice soluzione a questo problema è quella di dividere il dataset iniziale in due, assegnando un set di dati alla selezione del modello e un altro alla valutazione del modello. Questi set di dati sono comunemente chiamati training set e test set. Tuttavia, questo metodo soffre di problematiche relative alla variabilità della stima quando il dataset è di piccole dimensioni. Quando il dataset è ridotto, la parte riservata alla valutazione può non essere rappresentativa dell'intero dominio dei dati, causando una stima inaccurata delle performance del modello. Inoltre, la variabilità tra le possibili partizioni può portare a stime dell'errore di previsione molto diverse a seconda della partizione scelta. Questo problema è amplificato nei contesti ad alta dimensionalità, dove il numero di variabili supera di gran lunga il numero di osservazioni, come spesso accade nei dati di microarray (Ambroise C. e McLachlan G.J. (2002)).

Per affrontare questi problemi, tecniche più sofisticate, come la cross-validation e il bootstrap, sono state proposte e hanno guadagnato accettazione come standard nel campo della statistica e del machine learning. Queste tecniche, discusse più avanti, offrono soluzioni più robuste e affidabili per la stima dell'errore di previsione.

3.2 Cross Validation

Una delle prime proposte per uno stimatore che massimizzasse l'utilizzo delle informazioni disponibili, evitando di escludere troppe osservazioni per formare un test set ma che fosse anche unbiased,

è stato il metodo leave-one-out. Questo metodo prevede di escludere una sola osservazione dal training set, sul quale il modello viene poi valutato. Tuttavia questo metodo introduce una grande variabilità nella stima dell'errore di previsione, poiché il modello può ottenere un'accuratezza stimata o del 100% o dello 0% a seconda dell'osservazione esclusa.

Per risolvere questo problema, Lachenbruch P.A. e Mickey M.R. (1968) hanno proposto una tecnica oggi conosciuta come leave-one-out cross-validation. Questa tecnica consiste nel ripetere il metodo leave-one-out per tutte le osservazioni nel dataset, calcolando poi la media degli errori ottenuti per stimare l'errore di previsione finale. Stone M. (1974) ha ulteriormente sviluppato il concetto di cross-validation, discutendone l'uso per la valutazione delle previsioni statistiche e ponendo importanti basi per studi successivi. Per diminuire la potenza di calcolo richiesta dalla LOOCV, viene introdotta la k-fold cross-validation che consiste nel dividere il dataset in k parti uguali, per poi allenare il modello su $k-1$ parti e testato sulla parte rimanente. Questo processo viene ripetuto k volte, ogni volta con una diversa parte come set di test, e la stima finale dell'errore di previsione è la media degli errori ottenuti. Questo metodo bilancia meglio il bias e la varianza rispetto al leave-one-out, riducendo la varianza senza aumentare significativemente il bias.

Wong T.T. (2015) valuta la performance di algoritmi di classificazione utilizzando queste due tecniche di cross-validation, evidenziando i vantaggi e gli svantaggi di ciascuna tecnica a seconda dello scenario applicativo. Ambroise C. e McLachlan G.J. (2002) conducono uno studio su dati di tipo microarray ed ottengono risultati in favore della 10-fold cross-validation rispetto alla leave-one-out, dato che la prima offre un ottimo bilanciamento tra bias e varianza, mentre la seconda potrebbe presentare un basso bias ma a costo di un'alta varianza. Arlot S. e Celisse A. (2010) hanno condotto una revisione completa delle procedure di cross-validation per la selezione del modello, evidenziando i punti di forza e le debolezze di ciascun metodo. Il loro lavoro è fondamentale per comprendere l'evoluzione delle tecniche di cross-validation e le loro applicazioni pratiche.

3.3 Bootstrap

Efron B. (1983) introduce le tecniche basate sul bootstrap per migliorare l'accuratezza della stima dell'errore di previsione, ponendo le basi per sviluppi successivi nell'ambito dei metodi di ricampionamento. Il principale obiettivo dei metodi proposti da Efron è ridurre il bias ottimista associato all'errore apparente. Tra gli stimatori presentati vi sono il bootstrap semplice, il doppio bootstrap e il bootstrap .632. Quest'ultimo, in particolare, combina la stima ottenuta dal bootstrap semplice con l'errore apparente, assegnando un peso di .632 al primo e .368 al secondo, diminuendo così il bias ottimista dell'errore apparente. Successivamente, Efron B. e Tibshirani R.J. (1997) propongono il bootstrap .632+, che introduce il concetto di no-information error rate. Questo metodo assegna i pesi ai due errori in base al livello di overfitting, dando maggiore importanza all'errore apparente quando l'overfitting è minore.
I libri di Efron B. e Tibshirani R.J. (1993) e di Chernick M.R. (2007) rappresentano una guida completa sui metodi basati sul bootstrap, offrendo una panoramica dettagliata delle varie tecniche e delle loro applicazioni. Jiang W. e Simon R. (2007) confrontano diversi stimatori bootstrap per la stima dell'errore di previsione nella classificazione microarray, introducendo due nuovi stimatori: il repeated-leave-one-out bootstrap (RLOOB) e l'adjusted bootstrap (ABS), volti a correggere il bias degli stimatori bootstrap. Jiang W. e Chen B.E. (2013) modificano il bootstrap .632+ per risolvere i problemi riscontrati nei dati microarray, dove il numero di variabili supera quello delle osservazioni. L'articolo fornisce sia spunti teorici sia linee guida pratiche per applicare la loro versione del bootstrap .632+. Bischl B., Mersmann O., Trautmann H. e Weihs C. (2012) esplorano vari metodi di ricampionamento, tra cui tecniche basate sul bootstrap, per valutare metamodelli in computazione evolutiva. Gli autori dimostrano come queste tecniche possano migliorare l'affidabilità del modello e fornire una valutazione più accurata.

3.4 Metodi di stima alternativi

Fu W.J., Carroll R.J. e Wang S. (2005) esplorano l'utilizzo del bootstrap combinato con la cross-validation per la stima dell'errore di previsione. Il loro studio mostra che questa combinazione

fornisce stime dell'errore più affidabili rispetto ai metodi tradizionali, soprattutto quando il numero di osservazioni è limitato. Van Sanden S. et al. (2012) evidenziano le possibili distorsioni nella stima dell'errore di previsione in presenza di dati ad alta dimensionalità, proponendo l'uso del BCV per mitigare tali problemi. Hefny A. e Atiya A.F. (2010) introducono un nuovo stimatore dell'errore di previsione basato sul metodo Monte Carlo, fornendo un'alternativa sia alla tradizionale cross-validation sia ai metodi basati sul bootstrap. Il loro studio, che include analisi teoriche e risultati empirici, dimostra l'efficacia del nuovo stimatore che chiamano GMCP.
Cawley G.C. e Talbot N.L.C. (2010) affrontano il problema dell'overfitting nella selezione del modello e la conseguente distorsione nella valutazione della performance. Varma S. e Simon R. (2006) esaminano il bias nella stima dell'errore quando si utilizza la cross-validation per la selezione del modello, sottolineando l'importanza di considerare questo aspetto per evitare stime ottimistiche delle performance del modello.
Studi comparativi come quelli di Wong T.T. (2015) e Kim J.H. (2009) hanno dimostrato che ripetere la *k*-fold cross-validation riduce la varianza in modo più efficiente rispetto alla semplice k-fold cross-validation. Tuttavia questa maggiore efficienza richiede uno sforzo computazionale che aumenta esponenzialmente in base al numero delle osservazioni. Mostrano inoltre come in diversi casi restituisca una stima più accurata del .632+.

3.5 Applicazioni pratiche e studi comparativi

Uno dei primi studi comparativi è stato quello di Kohavi R. (1995), che ha confrontato metodi basati su cross-validation e bootstrap per la stima dell'accuratezza e la selezione del modello. I suoi risultati sono ancora rilevanti nelle tecniche di valutazione dei modelli contemporanee. Kohavi ha mostrato che, sebbene diversii metodi possano essere efficaci, la scelta del metodo appropriato dipende dal contesto specifico e dalle caratteristiche del dataset. Nell'ambito dei dati microarray, caratterizzati dalla ristretta quantità di osservazioni disponibili e dati ad alta dimensionalità, sono stati condotti diversi studi comparativi. Ambroise C. e McLachlan G.J. (2002) si concentrano sul bias di selezione nella selezione dei geni con dati di tipo microarray gene-expression. Raccomandano la 10-fold cross-validation rispetto alla leave-one-out e, tra gli stimatori bootstrap,

preferiscono il .632+, poiché questo metodo offre un miglior bilanciamento tra bias e varianza. Molinaro A.M., Simon R. e Pfeiffer R.M. (2005) confrontano vari metodi di ricampionamento per stimare l'errore di previsione. Lo studio fornisce un'analisi dettagliata dei punti di forza e di debolezza dei vari metodi, fungendo da guida per il loro utilizzo nella bioinformatica. Gli autori sottolineano che nessun metodo è universalmente migliore, ma che la scelta dipende dalle specifiche caratteristiche del dataset e del problema. Nell'ambito dell'analisi discriminante, Glele Kakaï R.L. e Palm R. (2009) e Ikechukwu E. (2016) offrono un confronto empirico di diversi stimatori dell'errore di previsione. I primi analizzano l'analisi discriminante logistica, mentre Ikechukwu esamina l'analisi discriminante con variabili multivariate binarie. I loro studi mostrano come diversi metodi di stima possano influenzare significativamente le performance del modello, suggerendo che la scelta del metodo di stima debba essere attentamente considerata in base al contesto specifico. Borra S. e Di Ciaccio A. (2008) confrontano vari metodi, inclusi quelli basati su penalità della covarianza. Il loro studio, attraverso estensive simulazioni, fornisce una valutazione delle performance relative degli stimatori. Successivamente, Borra S. e Di Ciaccio A. (2010) confrontano anche vari stimatori per l'errore extra-sample per metodi non parametrici.

3.6 Libri

Oltre agli articoli accademici, diversi libri rappresentano risorse fondamentali per comprendere a fondo le tecniche di stima dell'errore e di validazione del modello. Il libro di Hastie T., Tibshirani R.J. e Friedman J.H. (2009) è una guida estensiva che copre una vasta gamma di argomenti, tra cui la stima dell'errore di previsione. Questo testo fornisce solide fondamenta teoriche e offre numerose applicazioni pratiche, discusse in dettaglio, delle varie tecniche come la cross-validation e il bootstrap. Analogamente, James G., Witten D., Hastie T. e Tibshirani R.J. (2021) offrono un'introduzione al statistical learning con pratiche applicazioni in R. Copre i concetti fondamentali della stima dell'errore, fornendo spiegazioni chiare ed esempi concreti, utili per aiutare i lettori a comprendere e applicare queste tecniche nei loro studi. Kuhn M. e Johnson K. (2013) si concentrano sulle tecniche di predictive modeling, includendo metodi per la stima dell'errore di previsione. Gli autori forniscono una guida pratica per implementare questi metodi in scenari re-

ali, rendendola una risorsa preziosa per ricercatori e professionisti. Kuhn M. e Johnson K. (2019) discutono vari metodi per la selezione delle variabili, includendo consigli pratici su tecniche di stima dell'errore di previsione con l'obiettivo di ottenere performance del modello robuste. Infine, il libro di Efron B. e Tibshirani R.J. (1993) funge da introduzione al bootstrap, un potente strumento statistico per stimare la distribuzione di una popolazione attraverso il ricampionamento. Efron e Tibshirani trattano le fondamenta teoriche su cui si basa il bootstrap e le sue molteplici applicazioni, offrendo sia una prospettiva storica che pratica su questo metodo innovativo.

Bibliografia

Ambroise C. e McLachlan G.J. (2002). “Selection Bias in Gene Extraction on the Basis of Microarray Gene-Expression Data”. In: *Proceedings of the National Academy of Sciences of the United States of America* 99(10), pp. 6562–6566.

Arlot S. e Celisse A. (2010). “A survey of cross-validation procedures for model selection”. In: *Statistics Surveys* 4, pp. 40–79.

Bischl B., Mersmann O., Trautmann H. e Weihs C. (2012). “Resampling methods for meta-model validation with recommendations for evolutionary computation”. In: *Evolutionary Computation* 20(2), pp. 249–275.

Borra S. e Di Ciaccio A. (2008). *Estimators of extra-sample error for non-parametric methods. A comparison based on extensive simulations*. Tech. Rep. 2008/19. Dept. of Statistics, Prob. and Appl. Statistics, Univ. of Roma La Sapienza.

Borra S. e Di Ciaccio A. (2010). “Measuring the prediction error. A comparison of cross-validation, bootstrap and covariance penalty methods”. In: *Computational Statistics & Data Analysis* 54(12), pp. 2976–2989.

Cawley G.C. e Talbot N.L.C. (2010). “On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation”. In: *Journal of Machine Learning Research* 11, pp. 2079–2107.

Chernick M.R. (2007). *Bootstrap Methods: A Guide for Practitioners and Researchers*. Wiley Series in Probability and Statistics. Hoboken: John Wiley & Sons, 2nd edition.

Dimitriadou E., Hornik K., Leisch F., Meyer D. e Weingessel A. (2023). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. R package version 1.7-14.

Efron B. (1983). “Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation”. In: *Journal of the American Statistical Association* 78(382), pp. 316–331.

Efron B. e Tibshirani R.J. (1993). *An Introduction to the Bootstrap*. Monographs on Statistics and Applied Probability. London: Chapman & Hall.

Efron B. e Tibshirani R.J. (1997). “Improvements on Cross-Validation: The .632+ Bootstrap Method”. In: *Journal of the American Statistical Association* 92(438), pp. 548–560.

Fu W.J., Carroll R.J. e Wang S. (2005). “Estimating misclassification error with small samples via bootstrap cross-validation”. In: *Bioinformatics* 21(9), pp. 1979–1986.

Glele Kakaï R.L. e Palm R. (2009). “Empirical comparison of error rate-estimators in logistic discriminant analysis”. In: *Journal of Statistical Computation and Simulation* 79(2), pp. 111–120.

Hastie T., Tibshirani R.J. e Friedman J.H. (2009). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Springer Series in Statistics. New York: Springer, 2nd edition.

Hefny A. e Atiya A.F. (2010). “A New Monte Carlo-Based Error Rate Estimator”. In: Artificial Neural Networks in Pattern Recognition ANNPR 2010, pp. 37–47.

Ikechukwu E. (2016). “Evaluation of Error Rate Estimators in Discriminant Analysis with Multivariate Binary Variables”. In: *American Journal of Theoretical and Applied Statistics* 5(4), pp. 173–179.

James G., Witten D., Hastie T. e Tibshirani R.J. (2021). *An Introduction to Statistical Learning. With Applications in R*. Springer Texts in Statistics. New York: Springer, 2nd edition.

Jiang W. e Chen B.E. (2013). “Estimating prediction error in microarray classification: Modifications of the 0.632+ bootstrap when $n < p$”. In: *Canadian Journal of Statistics* 41(1), pp. 133–150.

Jiang W. e Simon R. (2007). “A comparison of bootstrap methods and an adjusted bootstrap approach for estimating the prediction error in microarray classification”. In: *Statistics in Medicine* 26(29), pp. 5320–5334.

Kim J.H. (2009). “Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap”. In: *Computational Statistics & Data Analysis* 53(11), pp. 3735–3745.

Kohavi R. (1995). “A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection”. In: IJCAI 14, pp. 1137–1143.

Kuhn M. e Johnson K. (2013). *Applied Predictive Modeling*. New York: Springer.

Kuhn M. e Johnson K. (2019). *Feature Engineering and Selection : A Practical Approach for Predictive Models*. Chapman and Hall/CRC Data Science. New York: Chapman & Hall.

Lachenbruch P.A. e Mickey M.R. (1968). “Estimation of Error Rates in Discriminant Analysis”. In: *Technometrics* 10(1), pp. 1–11.

Molinaro A.M., Simon R. e Pfeiffer R.M. (2005). “Prediction error estimation: a comparison of resampling methods”. In: *Bioinformatics* 21(15), pp. 3301–3307.

Raschka S. (2018). “Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning”. In: *CoRR* abs/1811.12808.

Stone M. (1974). “Cross-Validatory Choice and Assessment of Statistical Predictions”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 36(2), pp. 111–133.

Van Sanden S. et al. (2012). “Genomic Biomarkers for a Binary Clinical Outcome in Early Drug Development Microarray Experiments”. In: *Journal of Biopharmaceutical Statistics* 22(1), pp. 72–92.

Varma S. e Simon R. (2006). “Bias in error estimation when using cross-validation for model selection”. In: *BMC Bioinformatics* 7(1), p. 91.

Wong T.T. (2015). “Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation”. In: *Pattern Recognition* 48(9), pp. 2839–2846.