Comparazione di classificatori in Credit Card Approval

Valerio Colitta, Daniele Cominu, Alessio Fiorenza

Aprile 2017

1. Descrizione preliminare

Volendo approfondire i metodi e le tecniche di classificazione affrontate nel corso di Metodi Quantitativi per l'Informatica, abbiamo deciso di confrontarli risolvendo il problema presentato in *Credit Approval Data Set* [1]; abbiamo deciso di utilizzare modelli di diverse complessità ed abbiamo osservato che data la conformazione dei dati e la loro dimensione ridotta, si ottengono risultati migliori con modelli meno complessi.

2. Il dataset

Il dataset riguarda delle domande di approvazione di carte di credito, ed e' composto da 690 campioni, 37 dei quali con valori mancanti. Sono presenti in totale 15 feature, le quali si articolano in 6 continue e 9 categoriche. Tali dati sono poi classificati in due classi {+, -} che rappresentano rispettivamente l'approvazione o meno della carta di credito. Il significato delle 15 feature non è noto per poter mantenere la confidenzialità dei dati.

3. Manipolazione dei dati

Come prima cosa sono stati eliminati manualmente dal dataset i 37 campioni di cui non erano stati specificate alcune feature, riducendo ulteriormente la dimensione del dataset a 653 campioni.

3.1. One Hot Encoding

Per superare l'eterogeneità nella tipologia delle feature, si è deciso di applicare la tecnica del *One Hot Encoding* [2] per le feature categoriche; l'*One Hot Encoding* è una tecnica utilizzata per trattare feature categoriche in problemi di classificazione e regressione, e consiste nel tradurre una feature categorica che può assumere n valori distinti in un vettore di n feature binarie; per ogni campione la feature i-esima del vettore calcolato assume il valore 1 se e solo se il campione assume il valore i-esimo per la feature considerata; le restanti n - 1 sono dunque settate a 0 per tale campione. Se si fosse utilizzata una feature a valori reali, in cui per ogni valore assunto dalla feature viene associato un numero reale, sarebbero state introdotte delle distanze diverse tra i valori assunti dalla feature; Ad esempio, considerando una feature categorica che descrive la specie di appartenenza, mappando i valori assunti a dei numeri incrementali, si può notare come si introduca una maggiore distanza tra i campioni (2, 7) rispetto ai campioni (4, 5).

Sample	Category	Numerical
1	Human	1
2	Human	1
3	Penguin	2
4	Octopus	3
5	Alien	4
6	Octopus	3
7	Alien	4

Figura 1: mapping categorica -> reale

Introducendo il one hot encoding si può notare come le distanze tra i campioni (2,7) e (4,5) siano le stesse

Sample	Human	Penguin	Octopus	Alien
1	1	0	0	0
2	1	0	0	0
3	0	1	0	0
4	0	0	1	0
5	0	0	0	1
6	0	0	1	0
7	0	0	0	1

Figura 2: one hot encoding

3.2. Standardizzazione

Successivamente all'one hot encoding, è seguita la fase di standardizzazione delle variabili; per ogni feature sono stati calcolati la media campionaria μ e la deviazione standard campionaria σ, definite come:

$$\mu_j = \frac{1}{N}\sum_{i=1}^{N} x_{ij} \tag{1}$$

$$\sigma_j = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_{ij} - \mu_j)^2} \tag{2}$$

Una volta calcolati questi due valori, i campioni vengono standardizzati attraverso la seguente formula:

$$x_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \quad (3)$$

Questa trasformazione fa si che le nuove feature abbiano valor medio μ nullo e varianza σ^2 unitaria; ciò è necessario poiché alcuni metodi utilizzano la distanza euclidea e potrebbero essere ingannati dall'utilizzo di diverse scale per le diverse feature.

4. Principal Component Analysis

5. Visualizzazione dei dati

In genere risulta estremamente utile effettuare una visualizzazione dei dati prima di procedere all'applicazione dei modelli, in modo da individuare pattern utili per guidare il processo di selezione del modello e l'analisi dei risultati ottenuti. I grafici in Figura 3 rappresentano le distribuzioni dei campioni considerando coppie di feature continue; ossia nella cella presente alla riga *i-esima* e colonna *j-esima* è presente un grafico in cui sull'asse x sono riportati i valori assunti dalla feature *j-esima*, sull'asse y quelli assunti dalla feature *i-esima* .

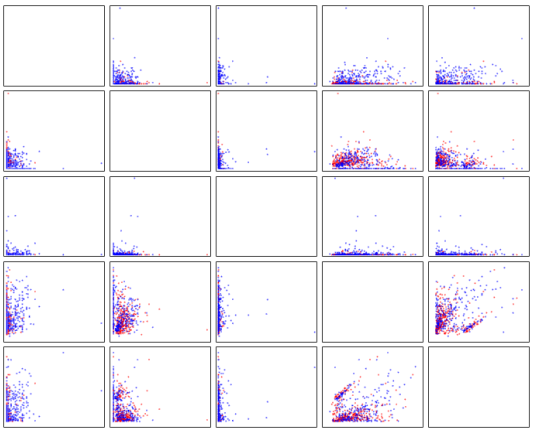

Figura 3

Dai grafici non si evincono particolari legami tra le feature. E' pertanto necessario continuare ad investigare.

6. Modelli utilizzati

Si è deciso di affrontare il problema utilizzando diversi modelli, in modo da poter approfondire la struttura del problema proposto. Si è notato come modelli più semplici riescano a generalizzare maggiormente a fronte di nuovi dati di input, e ciò è dovuto a diversi fattori, tra cui la dimensione ridotta del dataset e il numero relativamente alto di feature. Prima del training il dataset è stato inizialmente diviso in due parti: *Training set* (80%) e *Test set* (20%), utilizzati rispettivamente per il training e la selezione degli iperparametri attraverso la cross-validation e il test dei risultati ottenuti.
Nella Tabella 1 sono riportati gli error-rate medi e minimi ottenuti applicando i diversi modelli.

Modello	Min	Med
Linear Discriminant Analysis	10.12	10
Quadratic Discriminant Analysis	10	10
Diagonal Discriminant Analysis	0.04	0.224
Logistic Regression (LB)	10	10
Logistic Regression (QB)	10	10
Logistic Regression (LBR)	10	10
Logistic Regression (QBR)	10	10

Tabella 1: errori minimi e medi dei modelli utilizzati

7. Gaussian Discriminant Analysis

La Gaussian Discriminant Analysis, o GDA, è una tecnica di classificazione di tipo *generativa*, nel senso che cerca di calcolare (attraverso Maximum Likelihood Estimation) la distribuzione $p(x|y)$ di una serie di campioni, a partire dalla loro classe y di appartenenza. Con la GDA, si assume che i campioni di ogni classe c siano distribuiti come delle Gaussiane Multivariate, nel senso che

$$p(x|y = c) \sim \mathcal{N}(\mu_c, \Sigma_c) \quad \forall c \in \{1, ..., C\} \quad (4)$$

dove μ_c rappresenta il vettore dei valori medi della distribuzione, e Σ_c la matrice di covarianza delle feature. Una volta trovati i due parametri (μ, Σ), per classificare un nuovo input x_0 in una delle C classi viene calcolato

$$\arg\max_c p(y = c|x = x_0) \quad \forall c \in \{1, ..., C\} \quad (5)$$

e cioè la massima probabilità che la classe di appartenenza di x_0 sia c, per tutte le C classi.

7.1. Linear Discriminant Analysis

7.2. Quadratic Discriminant Analysis

8. Logistic Regression

8.1. Linear Boundary

8.2. Quadratic Boundary

8.3. Regolarizzazione

9. Bias vs Variance

Riferimenti bibliografici

[1] John Ross Quinlan
Credit Approval Data Set
http://archive.ics.uci.edu/ml/datasets/Credit+Approval

[2] Håkon Hapnes Strand
One Hot Encoding
https://www.quora.com/What-is-one-hot-encoding-and-when-is-it-used-in-data-science