Classificazione di diagnosi relative a casi di tumore alla prostata

Secondo Progetto di Statistica

Anno Scolastico 2019-2020

Nannini Alice
Mat. 533887
Artificial Intelligence and Data Engineering

Indice

1 Introduzione

Il cancro della prostata è uno dei tumori più diffusi nella popolazione maschile e rappresenta circa il 20% di tutti i tumori diagnosticati nell'uomo: le stime, relative all'anno 2017, parlano di 34.800 nuovi casi l'anno in Italia, ma il rischio che la malattia abbia un esito infausto è basso, soprattutto se si interviene in tempo. Lo scopo di questo studio è quello di cercare un modello per poter prevedere la diagnosi di un tumore di questo tipo, partendo dai dati fisici del nodulo prostatico, ottenibili da esami poco invasivi, così da poter offrire la possibilità di velocizzare l'intervento in caso di tumore maligno.

2 Tabella dei dati

Il dataset è stato recuperato dal sito *https://www.kaggle.com/sajidsaifi/prostate-cancer*. È costituito da 100 campioni relativi a casi di cancri rilevati, per ognuno dei quali si hanno a disposizione le seguenti caratteristiche:

- **Diagnosis Result**: indica se il cancro sia stato diagnosticato come benigno (0) o maligno (1)
- **Radius**: raggio approssimativo del nodulo
- **Texture**: struttura del nodulo (più è disomogeneo più è alta la probabilità che sia maligno)
- **Perimeter**: perimetro approssimativo del nodulo
- **Area**: area approssimativa del nodulo
- **Smoothness**: regolarità del nodulo, indice di quanto il tessuto sia liscio
- **Compactness**: compattezza del nodulo, calcolata in funzione del volume e della superficie
- **Simmetry**: simmetria del nodulo (più il tessuto è asimmetrico più è alto il rischio di malignità)
- **Fractal Dimension**: dimensione frattale relativa al nodulo, rapporto che fornisce un indice statistico di complessità

```
> p = read.csv("tabella.csv")
> head(p)
  diagnosis_result radius texture perimeter area smoothness compactness symmetry fractal_dimension
1                0     23      12       151  954     143.00      278.00   242.00             79.00
2                1      9      13       133 1326     143.00       79.00   181.00             57.00
3                0     21      27       130 1203     125.00        0.16   207.00              0.06
4                0     14      16        78  386       0.07      284.00     0.26             97.00
5                0      9      19       135 1297     141.00      133.00   181.00             59.00
6                1     25      25        83  477     128.00        0.17   209.00             76.00
```

Figura 1: Prime righe del file *tabella.csv*

3 Analisi dei dati

Iniziamo l'analisi dei dati, con l'obiettivo di ottenere un modello in grado, se possibile, di classificare ogni caso di tumore come benigno o maligno, basandosi sugli attributi a disposizione.

3.1 Regressione Logistica

Proviamo, per prima cosa, una classificazione tramite regressione logistica. Questo metodo si adatta bene al nostro studio perché offre un output dicotomico. Si calcola quindi il modello lineare generalizzato, ponendo come valore di uscita il risultato della diagnosi, per poi procedere con la predizione delle probabilità di appartenenza alla classe 1 (cancro maligno) e 0 (cancro benigno).

```
> p.glm=glm(diagnosis_result~.,family=binomial,data=p)
> p.glm.pre=predict(p.glm,type="response")
```

L'accuratezza del modello è molto soddisfacente:

```
> sum((p.glm.pre>0.5)==(p$diagnosis_result>0.5))/length(p$diagnosis_result)
[1] 0.89
```

e questo significa che la nostra classificazione ha senso di esistere. Da notare anche il valore AIC, che è molto alto (87%).

```
> summary(p.glm)

Call:
glm(formula = diagnosis_result ~ ., family = binomial, data = p)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.6643  -0.3687   0.2323   0.5334   1.5409

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -25.266400   7.677495  -3.291 0.000998 ***
radius             0.020772   0.069778   0.298 0.765938
texture            0.062588   0.063703   0.982 0.325855
perimeter          0.379906   0.155760   2.439 0.014726 *
area              -0.019884   0.010499  -1.894 0.058237 .
smoothness         0.004750   0.009831   0.483 0.629012
compactness        0.003159   0.006145   0.514 0.607276
symmetry          -0.000783   0.004628  -0.169 0.865653
fractal_dimension  0.021707   0.017048   1.273 0.202909
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 132.813  on 99  degrees of freedom
Residual deviance:  69.264  on 91  degrees of freedom
AIC: 87.264

Number of Fisher Scoring iterations: 6
```

Figura 2: Sommario della regressione logistica sui dati

Procediamo, per cui, con l'analisi della matrice di confusione, della curva ROC e dell'area sotto alla curva:

```
> mconfmat(p$diagnosis_result,p.glm.pre)
            actual 1 actual 0
predicted 1       59         8
predicted 0        3        30
> p.glm.roc=mroc(p$diagnosis_result,p.glm.pre)
> mroc.plot(p.glm.roc)
> mauc(p.glm.roc)
[1] 0.9210526
```

Vediamo dalla matrice che i veri positivi, cioè le osservazioni che appartengono alla classe 1 e sono stati effettivamente classificati come casi maligni, sono 59 su un totale di 62, e questo risultato è molto buono per il nostro scopo. È infatti fondamentale che i casi gravi, per cui è necessario effettuare un rapido intervento, vengano individuati correttamente. Le osservazione relative, invece, a tumori benigni sono state individuate con una percentuale più bassa, dal momento che ne sono state classificate correttamente solo 30 su 38. Si è ottenuto, perciò, un TNR (True Negative Rate) uguale al 79%, a fronte di un TPR (True Positive Rate) del 95%. Andando infine a studiare la curva ROC (fig. 3 nella pagina seguente), possiamo confermare quanto trovato nella matrice di confusione: la curva non è perfetta, ma è di molto sopra al tirare a caso (linea rossa tratteggiata). L'area sotto la curva, uguale a 0.92, è ulteriore conferma della bontà del modello costruito.

3.2 Analisi Discriminante Lineare e Quadratica

Proviamo a effettuare anche un'analisi discriminante, per poter poi confrontare i vari modelli ottenuti e vedere quale è da ritenere migliore per il nostro problema di classificazione. Effettuiamo, inizialmente, il calcolo del modello lineare e ne valutiamo l'efficacia tramite la predizione:

```
> p.lda=lda(diagnosis_result~.,data=p,CV=F)
> p.lda.pre=predict(p.lda)
> p.lda.post=p.lda.pre$posterior[,2] #prob. a posteriori di ottenere 1
> sum((p.lda.post>0.5)==(p$diagnosis_result>0.5))/length(p$diagnosis_result)
[1] 0.86
```

Otteniamo un'accuratezza dell'86%, più bassa di tre punti rispetto al modello precedente, ma comunque alta. Procediamo con la valutazione, visualizzando la matrice di confusione e la curva *ROC* (fig. 3):

```
> mconfmat(p$diagnosis_result,p.lda.post)
            actual 1 actual 0
predicted 1       57        9
predicted 0        5       29
> p.lda.roc=mroc(p$diagnosis_result,p.lda.post)
> mroc.plot(p.lda.roc)
> mauc(p.lda.roc)
[1] 0.9089559
```

Possiamo notare che anche nella matrice di confusione viene sottolineata una perdita di accuratezza, poiché si ha un maggior numero di predizioni incorrette rispetto al modello costruito tramite regressione logistica, fatto confermato dalla curva *ROC*, la cui area risulta leggermente minore.

Eseguiamo l'analisi discriminante quadratica, calcolando il modello ed effettuandone la valutazione attraverso gli stessi mezzi:

```
> p.qda=qda(diagnosis_result~.,data=p,CV=F)
> p.qda.pre=predict(p.qda)
> p.qda.post=p.qda.pre$posterior[,2]
> sum((p.qda.post>0.5)==(p$diagnosis_result>0.5))/length(p$diagnosis_result)
[1] 0.87
> mconfmat(p$diagnosis_result,p.qda.post)
            actual 1 actual 0
predicted 1       56        7
predicted 0        6       31
> p.qda.roc=mroc(p$diagnosis_result,p.qda.post)
> mroc.plot(p.qda.roc,col="blue")
> mauc(p.qda.roc)
[1] 0.890745
```

In questo caso, otteniamo un guadagno nell'accuratezza (87%) in confronto al modello lineare, mentre la curva evidenzia invece un peggioramento: le predizioni sono, infatti, migliori in generale perché il numero di tumori classificati come maligni è più vicino alla realtà (63 rispetto al totale effettivo 62), ma nel particolare si riscontra un maggior numero di falsi positivi e falsi negativi.

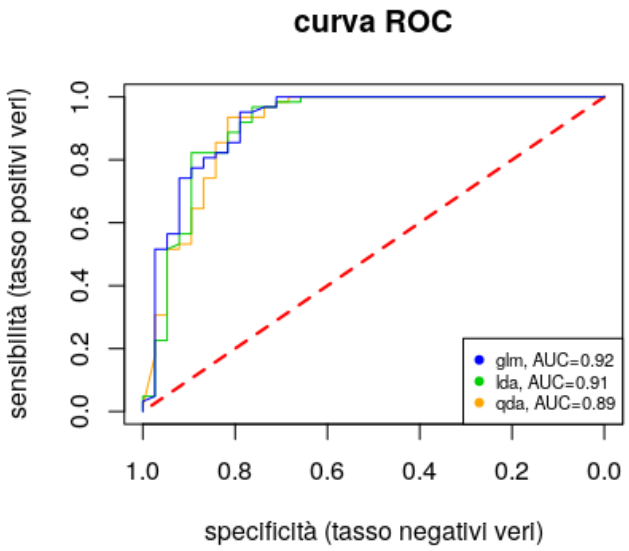


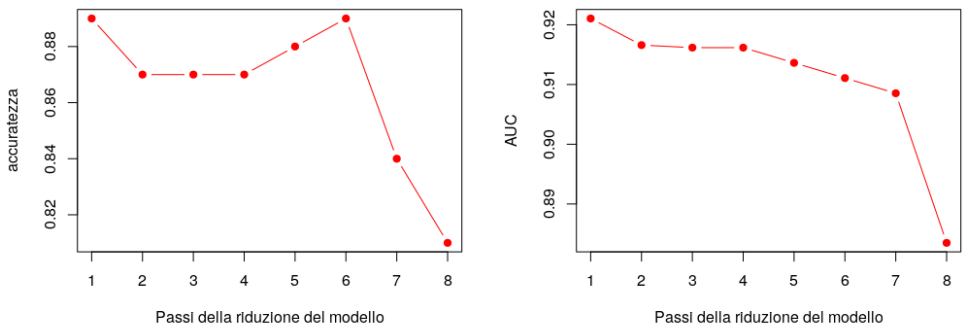
Figura 3: Curve ROC relative ai tre modelli analizzati

3.3 Conclusioni

Possiamo concludere che il primo modello, basato sulla regressione logistica, è il più soddisfacente: in particolare, è quello che riesce a classificare il maggior numero di tumori maligni, considerato il caso più critico e quindi più importante da rilevare. Grazie a questa classificazione, possiamo dire che ci sono buone probabilità di poter intervenire in supporto degli esami medici, e velocizzare il processo diagnostico del tumore preso in considerazione, riuscendo a segnalare tempestivamente i casi più gravi considerando le caratteristiche fisiche del tessuto istologico che sono identificabili tramite esami veloci e poco invasivi. È da sottolineare che quest'analisi viene effettuata su un numero relativamente ristretto di osservazioni, per cui sarà interessante portarla avanti aggiungendo dati di altri casi per vedere se l'accuratezza venga mantenuta su larga scala.

4 Appendice

Prendendo in considerazione il modello di figura 2 a pagina 2, analizziamo le correlazioni tra gli attributi per cercare di ridurlo mantenendo risultati il più alti possibili. Si trova una forte relazione tra *area* e *perimeter*, per cui procediamo a eliminare uno dei due fattori (*perimetro*), per vedere se l'altro riesce a spiegare meglio i dati: l'accuratezza scende all'87%, in confronto all'originale del 89%, e infatti si vede dalla matrice di confusione che il TPR è diminuito, mentre l'area sotto alla curva ROC è pressoché uguale. Il p-value di *area*, inoltre, è diminuito molto, come ci aspettavamo eliminando l'attributo a essa correlato. Procediamo con l'eliminazione del fattore *radius*, che presenta il *p-value* minore: la situazione rimane regolare, non abbiamo cioè né un calo di accuratezza né un calo del valore AUC. La stessa cosa succede all'eliminazione di *symmetry*, sempre scelto in base al *p-value*. Quando, però si procede a rimuovere *smoothness*, otteniamo un aumento di accuratezza, spiegato nella matrice di confusione dal fatto che ci sia stata una classificazione corretta di tumore benigno in più rispetto ai casi precedenti. Si esclude, ora, il fattore *texture*: si riscontra un ulteriore aumento dell'accuratezza, risultante da un'aggiuntiva corretta classificazione di caso di tumore maligno. Tentiamo, infine, l'eliminazione degli attributi *compactness* e $fractalDimension$, mantenendo *area* a spiegare il modello, che da sola porta un'accuratezza dell'81%. I grafici riassuntivi del processo di riduzione del modello si trovano in figura 4: vediamo che l'opzione migliore è, appunto, quella del passo 6, in cui si hanno solo 4 fattori mantenendo valori quasi uguali al passo 1 (modello con tutti i fattori).



(a) *Andamento dell'accuratezza* (b) *Andamento dell'area sotto la curva ROC*

Figura 4: Grafici riassuntivi della riduzione del modello

La regressione logistica effettuata con tutti gli attributi resta, per noi, più soddisfacente, perché riesce a classificare meglio i casi di cancro maligno, che come abbiamo già detto sono considerati i più importanti da individuare.