Valutazione di Modelli di Classificazione

Programmazione di Applicazioni Data Intensive

Laurea in Ingegneria e Scienze Informatiche
DISI – Università di Bologna

Gianluca Moro
Dipartimento di Informatica – Scienza e Ingegneria
Università di Bologna
Via dell'Università, 50 – I-47522 Cesena (FC)
Gianluca.Moro@Unibo.It

(draft)

Valutazione di Modelli di Classificazione: Matrice di Confusione (Esempio con due classi)

I modelli di classificazione si valutano sul test/validation set calcolando la seguente matrice

		classe predetta	
		a	b
classe reale	a	veri positivi TP	falsi negativi FN
	b	falsi positivi FP	veri negativi TN

con k cross fold validation, la matrice riflette i risultati di k modelli

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FN+FP}$$

$$\text{Precision(a)} = \frac{TP}{TP+FP} \quad \text{Precision(b)} = \frac{TN}{TN+FN} \quad \text{Precision} = \frac{\text{Precision(a)}+\text{Precision(b)}}{2}$$

$$\text{Recall(a)} = \frac{TP}{TP+FN} \quad \text{Recall(b)} = \frac{TN}{TN+FP} \quad \text{Recall} = \frac{\text{Recall(a)}+\text{Recall(b)}}{2}$$

$$\text{F1-Measure(a)} = \frac{2 \times \text{Precision(a)} \times \text{Recall(a)}}{\text{Precision(a)} + \text{Recall(a)}} \quad \text{F1-Measure(b)} = \frac{2 \times \text{Precision(b)} \times \text{Recall(b)}}{\text{Precision(b)} + \text{Recall(b)}}$$

$$\text{F1-Measure} = \frac{\text{F1-Measure(a)} + \text{F1-Measure(b)}}{2}$$

Precision e Recall sono essenziali quando le classi sono sbilanciate

F1-Measure riassume Precision e Recall in un unico valore

Classificazione e Tasso di Errore

- il tasso di errore calcolato sul training set è inevitabilmente ottimistico rispetto all'errore atteso su nuovi dati
- i dati di training possono essere lievemente diversi da quelli di test
  - ad esempio: in un'applicazione bancaria il training set per predire l'insolvenza su prestiti può riguardare una sola regione, con la necessità di estendere il risultato all'intero paese
- I dati nei problemi reali sono suddivisi in tre subset
  - training
  - validation, usato per fare il tuning degli iperparametri
  - test, per simulare il tasso di errore atteso sui nuovi dati

Kappa Statistics: Guadagno Rispetto ad un Classificatore Casuale

- Two confusion matrices for a 3-class problem: actual predictor (left) vs. random predictor (right)

	Predicted Class					Predicted Class				
(A)		*a*	*b*	*c*	*total*	(B)		*a*	*b*	*c*	*total*
Actual class	*a*	88	10	2	*100*	**Actual Class**	*a*	60	30	10	*100*
	b	14	40	6	*60*		*b*	36	18	6	*60*
	c	18	10	12	*40*		*c*	24	12	4	*40*
	total	*120*	*60*	*20*			*total*	*120*	*60*	*20*	

- Number of successes: sum of entries in diagonal (*D*)
- *Kappa* statistic: *(success rate of actual predictor - success rate of random predictor) / (1 - success rate of random predictor)*
- Measures relative improvement on random predictor: 1 means perfect accuracy, 0 means we are doing no better than random

Intervallo di Confidenza dell'Accuratezza di un Modello di Classificazione

- Supponiamo che un classificatore abbia un tasso di successo sul test set, i.e. accuratezza, del 75%
- quanto è attendibile questa accuratezza sull'intera popolazione dei dati, compresi quelli nuovi ignoti ?
  - 75% ± ??? non un singolo valore, ma un intervallo di accuratezze
  - l'intervallo di accuratezza dipende dalle dimensioni del test set
  - quanto le dimensioni del test set influenzano l'intervallo di accuratezza ?
- Tiriamo ad indovinare con risposte soggettive ??
- No! usiamo un metodo oggettivo dalle scienze statistiche
- In ogni progetto di data science con modelli predittivi, queste misure sono essenziali per valutare l'affidabilità del risultato
  - sono determinanti anche nella contrattazione con il committente

Modellazione della Classificazione come Processo di Bernoulli

- La classificazione di N istanze è modellabile con un processo Bernoulliano di N eventi binari indipendenti: errore o successo
  - esempio: testa o croce nel lancio di una moneta
  - se con 100 lanci abbiamo 75 teste, qual è la probabilità ***p*** di ottenere testa nel prossimo lancio? e dopo 1000 lanci?
  - siano N gli esperimenti, S successi (i.e. classificazioni corrette)
  - ***f*** = S/N tasso di successo (la nostra ACCURATEZZA)
- Intervallo di Confidenza
  - data ***f***, possiamo predire la reale accuratezza ***p*** del relativo modello ?
  - ***p*** è in un intervallo, con una data probabilità nota come confidenza
  - N=100 $\Rightarrow$ ***p*** $\in$ [69.1, 80.1] con confidenza (i.e. probabilità) del 80%
  - N=1000 $\Rightarrow$ ***p*** $\in$ [73.2, 76.7] con confidenza (i.e. probabilità) del 80%
    - All'aumentare di N, l'intervallo di confidenza si restringe

Dal teorema del limite centrale.

Se errori/successi non sono sempre eventi indipendenti, questo approccio è comunque una buona approssimazione

Processo di Bernoulli

- N esperimenti: **f** = S/N (**accuratezza**)
  - **f** ha distribuzione binomiale Bin(N, p) con media **p** e varianza **p(1-p)/N**
  - **p** è la reale accuratezza che vogliamo stimare
  - per N grande (N > 30) la distribuzione di **f** è approssimabile con la **distribuzione z** (distribuzione normale standardizzata)
  - Pr[**-z** ≤ (**f** – **p**) ≤ **z**] = **confidenza** fissata a priori
    - valori pre-calcolati per deviazione standard unitaria, vediamo dopo
  - Esempio, desideriamo una confidenza *c* del 90% da cui deriva che
    - **z** = 1.65 ➔ Pr[-1.65 ≤ (**f** – **p**) ≤1.65] = **90%** --- dato **z** risolviamo per **p**

$$Pr\left[-z < \frac{f-p}{\sqrt{p(1-p)/N}} < z\right] = c$$

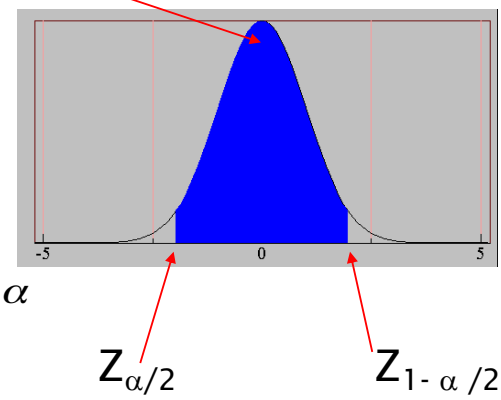
$$p = \left(f + \frac{z^2}{2N} \pm z\sqrt{\frac{f}{N} - \frac{f^2}{N} + \frac{z^2}{4N^2}}\right) / \left(1 + \frac{z^2}{N}\right)$$

Approfondimento: Analisi dell'Intervallo di Confidenza

- Quando nel test set N > 30
  - l'accuratezza approssima la distribuzione normale standard con media p e varianza p(1-p)/N (*acc = f*)

$$P(Z_{\alpha/2} < \frac{acc - p}{\sqrt{p(1-p)/N}} < Z_{1-\alpha/2}) = 1 - \alpha$$

Confidenza = Area =1− α (quando α invece è calcolato α è il p-value)



- Risolvendo per *p* abbiamo l'intervallo di confidenza

$$p = \frac{2 \times N \times acc + Z^2_{\alpha/2} \pm Z_{\alpha/2}\sqrt{Z^2_{\alpha/2} + 4 \times N \times acc - 4 \times N \times acc^2}}{2(N + Z^2_{\alpha/2})}$$

Intervallo di Confidenza dell'Accuratezza: Esempio

- Consideriamo un modello con accuratezza 80% valutata su un test set di 100 istanze:
  - N = 100, acc = 0.8
  - Sia 1-α = 0.95 (95% confidenza)
  - Dalla tabella otteniamo
    $Z_{\alpha/2} = 1.96$
  - Sostituendo questi valori nella formula precedente otteniamo

1-α	Z
0.99	2.58
0.98	2.33
0.95	1.96
0.90	1.65

N	50	100	500	1000	5000
p min	0.670	0.711	0.763	0.774	0.789
p max	0.888	0.866	0.833	0.824	0.811

Confrontare l'Accuratezza di 2 Modelli (i)

- Dati due modelli M1 ed M2, qual è il migliore ?
  - M1: testato su un dataset D1 di cardinalità n1, con errore e_1
  - M2, testato su un dataset D2 di cardinalità n2, con errore e_2
  - se n1 ed n2 sono abbastanza grandi (> 30) **il loro errore** è approssimabile da una distribuzione normale con media μ e standard deviation σ:

$$e_1 \sim N(\mu_1, \sigma_1) \qquad e_2 \sim N(\mu_2, \sigma_2)$$

  - La varianza approssimata è: $\hat{\sigma}_i^2 = \frac{e_i(1-e_i)}{n_i}$

Confrontare l'Accuratezza di 2 Modelli (ii)

- Come valutare se la differenza **d** tra le accuratezze dei due modelli è statisticamente significativa ?
- Sia **d** = e1 – e2
  - d ~ $N(d_t,\sigma_t)$ dove d_t è la reale differenza da stimare
  - la varianza σ_t^2 si ottiene come segue

$$\sigma_t^2 = \sigma_1^2 + \sigma_2^2 \cong \hat{\sigma}_1^2 + \hat{\sigma}_2^2$$
$$= \frac{e1(1-e1)}{n1} + \frac{e2(1-e2)}{n2}$$

Infine d_t (con confidenza 1-α) è $$d_t = d \pm Z_{\alpha/2}\hat{\sigma}_t$$

Confrontare due Modelli: Esempio

- Siano M1: n1 = 30, e1 = 0.15
  M2: n2 = 5000, e2 = 0.25
- d = |e2 – e1| = 0.1

$$\hat{\sigma}_d^2 = \frac{0.15(1-0.15)}{30} + \frac{0.25(1-0.25)}{5000} = 0.0043$$

- Con confidenza 1-α = 0.95, $Z_{\alpha/2}$ = 1.96

$$d_t = 0.100 \pm 1.96 \times \sqrt{0.0043} = 0.100 \pm 0.128$$

=> l'intervallo contiene 0 => la differenza tra i 2 modelli **non è statisticamente significativa**

- la differenza è solo frutto del caso, ossia ripetendo numerose volte gli esperimenti potremmo ad esempio ottenere risultati opposti

Quale Livello di Confidenza Rende Significativa la Differenza tra Due Modelli

- Sia **M1:** n1 = 30, e1 = 0.15 **M2:** n2 = 5000, e2 = 0.25
- d = |e2 – e1| = 0.1

$$\hat{\sigma}_d^2 = \frac{0.15(1-0.15)}{30} + \frac{0.25(1-0.25)}{5000} = 0.0043$$

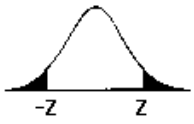
- Qual è la soglia della confidenza che rende la loro differenza statisticamente significativa ?
- Dobbiamo determinare il valore di $\mathbf{Z_{\alpha/2}}$ tale che

$$-d < Z_{\alpha/2}\hat{\sigma}_t < d \implies -\frac{d}{\hat{\sigma}_t} < Z_{\alpha/2} < \frac{d}{\hat{\sigma}_t} \text{ i.e. } 1-\alpha = P\left(-\frac{d}{\hat{\sigma}_t} < Z_{\alpha/2} < \frac{d}{\hat{\sigma}_t}\right)$$

- Sostituendo nell'esempio d e σ_t otteniamo $\mathbf{Z_{\alpha/2}} = \pm 1.527 \approx \pm 1.53$
  - che corrisponde a $\alpha = 0.126$ (p-value), $1-\alpha = 0.874$ quindi la ***differenza diventa significativa quando la confidenza è < 0.874***

Tabella per Calcolare la Confidenza (1-α) da Z

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	1.000	0.992	0.984	0.976	0.968	0.960	0.952	0.944	0.936	0.928
0,1	0.920	0.912	0.904	0.897	0.889	0.881	0.873	0.865	0.857	0.849
0,2	0.841	0.834	0.826	0.818	0.810	0.803	0.795	0.787	0.779	0.772
0,3	0.764	0.757	0.749	0.741	0.734	0.726	0.719	0.711	0.704	0.697
0,4	0.689	0.682	0.674	0.667	0.660	0.653	0.646	0.638	0.631	0.624
0,5	0.617	0.610	0.603	0.596	0.589	0.582	0.575	0.569	0.562	0.555
0,6	0.549	0.542	0.535	0.529	0.522	0.516	0.509	0.503	0.497	0.490
0,7	0.484	0.478	0.472	0.465	0.459	0.453	0.447	0.441	0.435	0.430
0,8	0.424	0.418	0.412	0.407	0.401	0.395	0.390	0.384	0.379	0.373
0,9	0.368	0.363	0.358	0.352	0.347	0.342	0.337	0.332	0.327	0.322
1,0	0.317	0.312	0.308	0.303	0.298	0.294	0.289	0.285	0.280	0.276
1,1	0.271	0.267	0.263	0.258	0.254	0.250	0.246	0.242	0.238	0.234
1,2	0.230	0.226	0.222	0.219	0.215	0.211	0.208	0.204	0.201	0.197
1,3	0.194	0.190	0.187	0.184	0.180	0.177	0.174	0.171	0.168	0.165
1,4	0.162	0.159	0.156	0.153	0.150	0.147	0.144	0.142	0.139	0.136
1,5	0.134	0.131	0.129	0.126	0.124	0.121	0.119	0.116	0.114	0.112
1,6	0.110	0.107	0.105	0.103	0.101	0.099	0.097	0.095	0.093	0.091
1,7	0.089	0.087	0.085	0.084	0.082	0.080	0.078	0.077	0.075	0.073
1,8	0.072	0.070	0.069	0.067	0.066	0.064	0.063	0.061	0.060	0.059
1,9	0.057	0.056	0.055	0.054	0.052	0.051	0.050	0.049	0.048	0.047
2,0	0.046	0.044	0.043	0.042	0.041	0.040	0.039	0.038	0.038	0.037
2,1	0.036	0.035	0.034	0.033	0.032	0.032	0.031	0.030	0.029	0.029
2,2	0.028	0.027	0.026	0.026	0.025	0.024	0.024	0.023	0.023	0.022
2,3	0.021	0.021	0.020	0.020	0.019	0.019	0.018	0.018	0.017	0.017
2,4	0.016	0.016	0.016	0.015	0.015	0.014	0.014	0.014	0.013	0.013
2,5	0.012	0.012	0.012	0.011	0.011	0.011	0.010	0.010	0.010	0.010
2,6	0.009	0.009	0.009	0.009	0.008	0.008	0.008	0.008	0.007	0.007
2,7	0.007	0.007	0.007	0.006	0.006	0.006	0.006	0.006	0.005	0.005
2,8	0.005	0.005	0.005	0.005	0.005	0.004	0.004	0.004	0.004	0.004
2,9	0.004	0.004	0.004	0.003	0.003	0.003	0.003	0.003	0.003	0.003
3,0	0.003									



Esempio con Z = ±1.53 scegliere la riga con Z = 1.5 e la colonna con 0.03 α è **0.126** **Confidenza** = 1-α = **0.874**

Calcoliamo la Soglia di Confidenza dell'Esempio Precedente

- Riprediamo M1: n1 = 30, e1 = 0.1 M2: n2 = 5000, e2 = 0.25
- d = |e2 – e1| = 0.1

$$\hat{\sigma}_d^2 = \frac{0.15(1-0.15)}{30} + \frac{0.25(1-0.25)}{5000} = 0.0043$$

- Sapendo che $Z_{\alpha/2}$ = 1.53 perché la confidenza è 0.874, otteniamo

$$d_t = 0.100 \pm 1.53 \times \sqrt{0.0043} = 0.100 \pm 0.1003$$

  - => come atteso la differenza tra i 2 modelli è ancora statisticamente non significativa perché l'intervallo contiene zero [-0.0003, 0.2003]
- ma con $Z_{\alpha/2}$ = 1.52, corrispondente a 0.871 di confidenza, abbiamo

$$d_t = 0.100 \pm 1.52 \times \sqrt{0.0043} = 0.100 \pm 0.099673$$

  - => la differenza è significativa poiché l'intervallo non contiene ZERO

Classification Models Comparison with McNemar's Statistical Hypothesis Test

Thomas Dietterich, Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms, 1998.

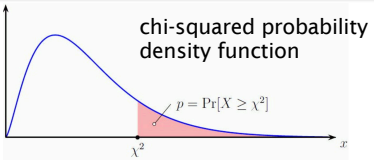
- McNemar's test is recommended for binary classifiers
  - as the Bernoulli method, it is useful when estimating the accuracy interval by multiple training copies of classifier models is too expensive
  - e.g. large deep learning models trained and evaluated on big datasets requiring often several days and weeks, some times just for 1 epoch
- It operates on a contingency table 2x2 built as follows
  - Let's consider 2 binary classifiers M1 and M2 and the contingency table

Contingency table	M2 correct	M2 incorrect
M1 correct	# same instances correctly classified by both models	# same instances correctly classified only by M1 and not M2
M1 incorrect	# same instances correctly classified only by M2	# same instances incorrectly classified by both models

  - If cells have similar counts, both models make errors in similar way → statistical hypothesis test to prove their difference is/is not significant

McNemar's Statistical Hypothesis Test

Contingency table	M2 correct	M2 incorrect
M1 correct	(a)	(c)
M1 incorrect	(b)	(d)



- It evaluates if exists a significant difference between the cells (b) and (c), i.e. on their reciprocal errors and correct predictions
  - The **null hypothesis** (H0) of marginal homogeneity states the 2 marginal prob. for each outcome are equals, i.e. $p_a+p_b=p_a+p_c$ and $p_c+p_d=p_b+p_d \rightarrow p_b=p_c$
  - where each p_j is the theoretical prob. of occurrences in *i-th* cell
  - it computes $X = \frac{(|b-c|-1)^2}{b+c}$ which follows a chi-squared distribution $\mathcal{X}^2$ of 1 freedom degree under the null hypothesis and enough large values
  - the larger the X, the more likely the difference between the two models is significant according to a confidence level
    - i.e. the more the prob. their difference is due to chance decreases
    - this prob. $p = \Pr[X \geq \chi^2]$ (see the area in the fig.) is the ***p-value*** achieved from the χ^2 chi-squared distribution and if the *p-value < 1 – confidence level*, then H0 is rejected

McNemar's Statistical Hypothesis Test: Example

- Let's suppose M1 and M2 achieved 60% and 50% of successes (accuracy) on the test set respectively, with the following values

Contingency table	M2 correct	M2 incorrect
M1 correct	80 (a)	40 (c)
M1 incorrect	20 (b)	60 (d)

- Let's set the confidence level 0.95; $X = \frac{(|b-c|-1)^2}{b+c} = 6.017$
- p-value = $\Pr[6.017 \geq \mathcal{X}^2] = 0.014 < 0.05 = 1 - 0.95$
- this means that the difference in the errors between the two models is statistically significant with 95% confidence
- → reject the **null hypothesis** (H0) of marginal homogeneity between the two models, i.e. they are not equivalent

The McNemar Test in Python

```python
from statsmodels.stats.contingency_tables import mcnemar
# example of a contingency table
contingency_table = [[80, 40],
                     [20, 60]]

# compute the mcnemar test
# with exact = False, when values are enough large, it uses the chi-squared distribution
# with exact = True it uses the exact binomial distribution
test_result = mcnemar(contingency_table, exact=False)

# print the test resuls
print('X-value=%.3f, p-value=%.3f' % (test_result.statistic, test_result.pvalue))

# accept or reject the null hypothesis using the p-value
confidence_level = 0.95
alpha = 1 - confidence_level

if test_result.pvalue > alpha:
  print("""The difference in the errors is not statistically significant (accept H0) \n\
  the two models are equivalent""")
else:
  print("""The difference in the errors is statistically significant (reject H0) \n\
  the two models are not equivalent""")
```

```
X-value=6.017, p-value=0.014
The difference in the errors is statistically significant (reject H0)
  the two models are not equivalent
```