Machine Learning

Università Roma Tre
Dipartimento di Ingegneria
Anno Accademico 2021 - 2022

Classificazione:
Overfitting e Regularization

Sommario

- Introduzione
- Overfitting nella Classificazione
- Regolarizzazione
- L2 Penalty
- L1 Penalty (sparse solutions)

Metriche di Qualità

[quality metric]

- Una metrica che si usa misura la frazione delle previsioni errate fornite:

$$\text{Errore} = \frac{\#\text{previsioni\_errate}}{\#\text{esempi}}$$

miglior valore possibile: 0.0

- Un'altra metrica possibile misura la frazione delle previsioni corrette:

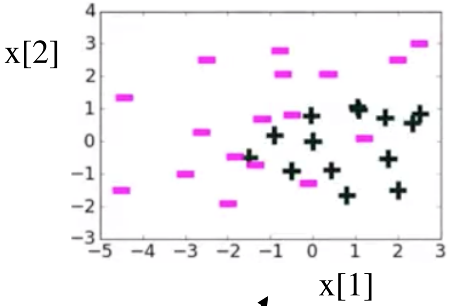
$$\text{Accuracy} = \frac{\#\text{previsioni\_corrette}}{\#\text{esempi}}$$

miglior valore possibile: 1.0

Overfitting

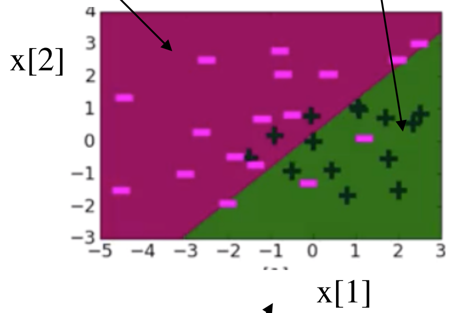
Apprendimento della Decision Boundary

j	*Φ*j	wj
0	1	0.23
1	x{1}	1.12
2	x{2}	-1.07



Data Points dell'esempio

Score(**x**) < 0 Score(**x**) > 0



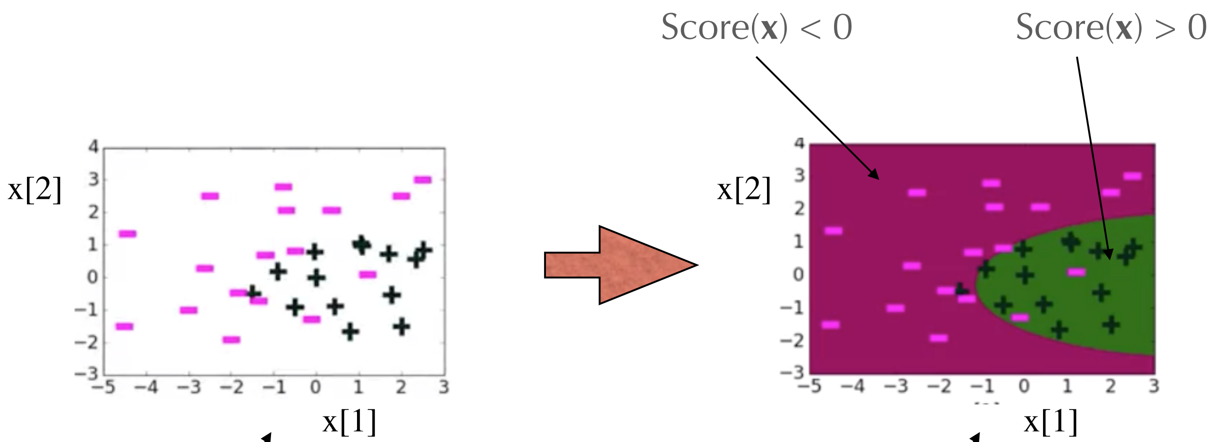
Decision Boundary:

0.23 + 1.12 x[1] - 1.07 x[2] = 0

Overfitting

Apprendimento della Decision Boundary

j	Φj	wj
0	1	1.68
1	x{1}	1.39
2	x{2}	-0.59
3	x{1}^2	-0.17
4	x{2}^2	-0.96



Data Points dell'esempio

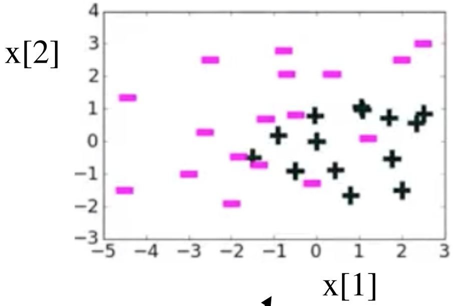
Decision Boundary:
1.68 + 1.39 x[1] - 0.59 x[2] - 0.17 x[1]^2 - 0.96 x[2]^2 = 0

Overfitting

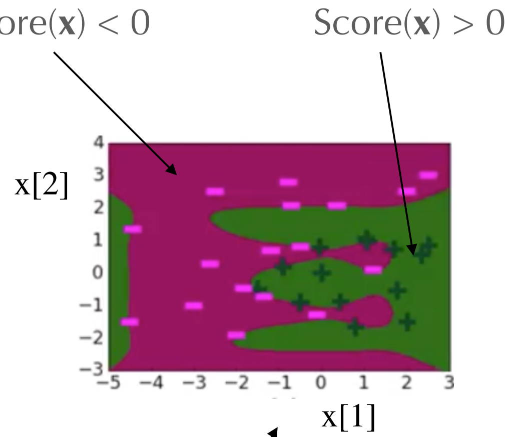
Apprendimento della Decision Boundary

j	Φj	wj
0	1	21.6
1	x{1}	5.3
2	x{2}	-42.7
3	x{1}^2	-15.9
4	x{2}^2	-48.6
5	x{1}^3	-11.0
6	x{2}^3	67.0
…	…	…
11	x[1]^6	0.8
12	x[2]^6	-8.6

I valori assoluti di vari coefficienti wj sono aumentati



Data Points dell'esempio



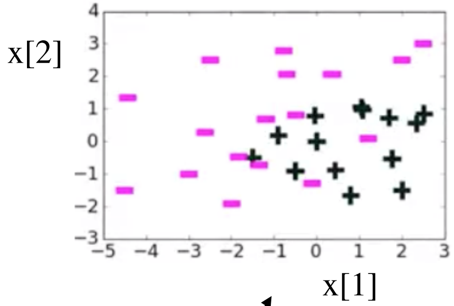
Decision Boundary (chiaro overfitting)

Overfitting

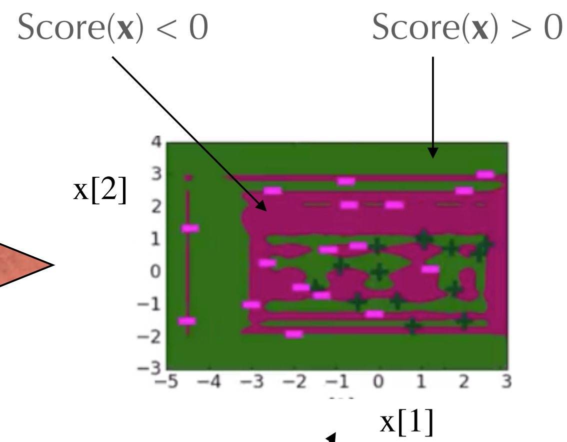
Apprendimento della Decision Boundary

j	Φj	wj
0	1	8.7
1	x{1}	5.1
2	x{2}	78.7
…	…	…
11	x{1}^6	-7.5
12	x{2}^6	**3803**
13	x{1}^7	21.1
14	x{2}^7	**-2406**
…	…	…
39	x[1]^20	-2*10^-8
40	x[2]^20	0.03

I valori assoluti di vari coefficienti wj sono aumentati ancora di più



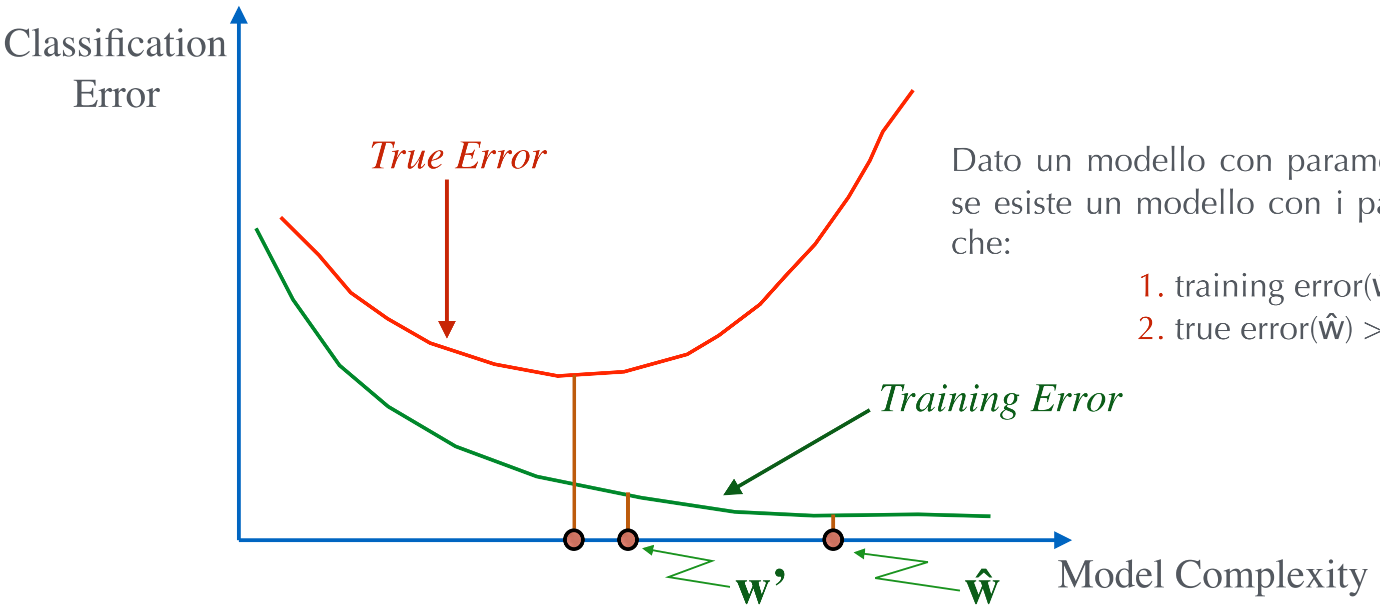
Data Points dell'esempio



Decision Boundary (overfitting ancora più evidente)

Andamento Errori e Bias-Variance Trade-off

- L'andamento del training error e del true error per la classification è in genere il seguente:



Dato un modello con parametri ŵ, si ha overfitting se esiste un modello con i parametri stimati w′ tale che:

1. training error(ŵ) < training error(w′)
2. true error(ŵ) > true error(w′)

- Dobbiamo come al solito considerare il trade-off tra bias e varianza.

Regularization nella Classificazione

- L'idea è quella di limitare il valore assoluto dei coefficienti w_i definendo come segue la funzione di qualità totale (da massimizzare nella fase di training):

  Qualità_totale = misura del "fit" - misura grandezza coefficienti

- Per misura del "fit" intendiamo una funzione come la MLE.
- La misura dei coefficienti possiamo definirla in vari modi.

Misura dei Coefficienti

👎 - Somma dei valori:

$$w_0 + w_1 + w_2 + \cdots + w_D$$

👍 - Somma dei valori assoluti (L1 norm):

$$|w_0| + |w_1| + |w_2| + \cdots + |w_D| = \sum_{j=0}^{D} |w_j| \triangleq \|\mathbf{w}\|_1$$

👍 - Somma dei quadrati (quadrato della L2 norm):

$$w_0^2 + w_1^2 + w_2^2 + \cdots + w_D^2 = \sum_{j=0}^{D} w_j^2 \triangleq \|\mathbf{w}\|_2^2$$

Funzione di Qualità nel caso L2 Penalty

- Questo è il caso in cui usiamo la somma dei quadrati (L2 Regularization).
- La funzione che rappresenta la qualità totale nel caso della logistic regression (L2 regularized logistic regression) è la seguente:

$$\text{Qualità\_totale}_{L_2} = \ln \mathcal{L}(\mathbf{w}) - \lambda \cdot \|\mathbf{w}\|_2^2$$

dove il parametro λ (tuning parameter) serve per bilanciare i due termini.

Funzione di Qualità nel caso L2 Penalty

Vediamo cosa accade a fronte di diversi valori del parametro λ:

- Se $\lambda = 0$:

  ci riconduciamo alla vecchia soluzione, ossia massimizzazione del likelihood($\mathbf{w}$) $\rightarrow$ $\hat{\mathbf{w}}_{MLE}$

- Se $\lambda \rightarrow \infty$:

  per soluzioni dove $\hat{\mathbf{w}} \neq \mathbf{0}$, il costo totale $\rightarrow - \infty$

  l'unica soluzione per massimizzare la qualità è: $\hat{\mathbf{w}} = \mathbf{0}$

- Se $0 < \lambda < \infty$:

$$0 < \|\hat{\mathbf{w}}\|_2^2 < \|\hat{\mathbf{w}}_{\mathrm{MLE}}\|_2^2$$

Scelta del Parametro di Tuning λ

Come già visto nel caso della Regressione, per la determinazione del parametro λ non usiamo mai il Test Set. Ci avvaliamo invece:

- del Validation Set, se abbiamo a disposizione un numero sufficientemente elevato di osservazioni;
- della Cross-Validation, se abbiamo a disposizione un numero limitato di osservazioni.

Bias-Variance Tradeoff

Il parametro λ controlla la complessità del modello:

- Parametro λ elevato:

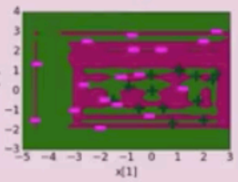
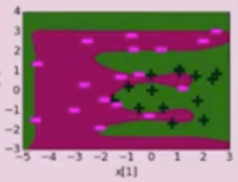
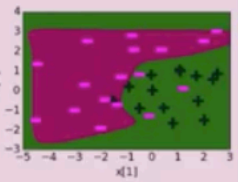
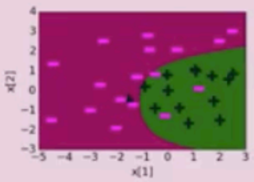
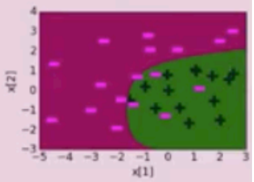
  high bias, low variance (e.g., $\hat{w} = 0$ per $\lambda = \infty$)

- Parametro λ piccolo:

  low bias, high variance (e.g., maximum likelihood (MLE) fit per polinomi di grado elevato per $\lambda = 0$)

L2 Regularization
Esempio

- Vediamo l'effetto della L2 regularization nel caso visto in precedenza (caso con 20 features):

Regularization:	$\lambda = 0$	$\lambda = 0.00001$	$\lambda = 0.001$	$\lambda = 1$	$\lambda = 10$
Range coefficienti:	-3170 to 3803	-8.04 to 12.14	-0.70 to 1.25	-0.13 to 0.57	-0.05 to 0.22
Decision boundary:					

Gradient Ascent con la L2 Regularization

- Come è noto, nell'algoritmo Gradient Ascent dobbiamo aggiornare il vettore dei pesi **w** come segue:

$$\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} + \alpha \cdot \nabla \text{Qualità\_totale}_{L_2}(\mathbf{w}^{(t)})$$

- Dobbiamo dunque calcolare il gradiente della funzione di qualità totale (L2 regularized log-likelihood):

$$\text{Qualità\_totale}_{L_2} = \ln \mathcal{L}(\mathbf{w}) - \lambda \cdot \|\mathbf{w}\|_2^2$$

Gradient Ascent con la L2 Regularization

- Nell'algoritmo l'aggiornamento dei pesi possiamo farlo per ogni componente w_j:

$$
\begin{aligned}
w_0^{(t+1)} &\leftarrow w_0^{(t)} + \alpha \cdot \frac{\partial \text{Qualità\_totale}_{L_2}(\mathbf{w}^{(t)})}{\partial w_0} \\
w_1^{(t+1)} &\leftarrow w_1^{(t)} + \alpha \cdot \frac{\partial \text{Qualità\_totale}_{L_2}(\mathbf{w}^{(t)})}{\partial w_1} \\
\cdots\cdots & \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
w_j^{(t+1)} &\leftarrow w_j^{(t)} + \alpha \cdot \frac{\partial \text{Qualità\_totale}_{L_2}(\mathbf{w}^{(t)})}{\partial w_j} \\
\cdots\cdots & \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
w_D^{(t+1)} &\leftarrow w_D^{(t)} + \alpha \cdot \frac{\partial \text{Qualità\_totale}_{L_2}(\mathbf{w}^{(t)})}{\partial w_D}
\end{aligned}
$$

Gradient Ascent con la L2 Regularization

- La derivata parziale della funzione di qualità totale rispetto al termine generico w_j è la seguente:

$$\frac{\partial \text{Qualità\_totale}_{L_2}(\mathbf{w}^{(t)})}{\partial w_j} = \underbrace{\text{derivata\_parziale}[j]}_{\text{Componente MLE}} - \underbrace{2\lambda w_j^{(t)}}_{\text{Componente L2 Penalty}}$$

Gradient Ascent con la L2 Regularization

- Questa è la versione dell'algoritmo:

$\mathbf{w}^{(1)} = 0$ (oppure lo inizializziamo in modo casuale)

$t = 1$

while $\|\nabla \text{Qualità\_totale}_{L_2}(\mathbf{w}^{(t)})\|_2 > \epsilon$

  for $j = 0, 1, ..., D$

$$\text{derivata\_parziale}[j] = \sum_{i=1}^{N} \phi_j(\mathbf{x}_i)\{I[y_i = +1] - P(y = +1 | \mathbf{x}_i, \mathbf{w}^{(t)})\}$$

$$w_j^{(t+1)} \leftarrow w_j^{(t)} + \alpha * (\text{derivata\_parziale}[j] - 2\lambda w_j^{(t)})$$

  $t \leftarrow t + 1$

Funzione di Qualità nel caso L1 Penalty

- Questo è il caso in cui usiamo la somma dei valori assoluti per la penalty (L1 Regularization). E′ in genere chiamata "sparse logistic regression".

- La funzione che rappresenta la qualità totale nel caso della logistic regression (L1 regularized logistic regression) è la seguente:

$$\text{Qualità\_totale}_{L_1} = \ln\mathcal{L}(\mathbf{w}) - \lambda \cdot \|\mathbf{w}\|_1$$

dove il parametro λ (tuning parameter) serve per bilanciare i due termini.

Funzione di Qualità nel caso L1 Penalty

Anche in questo caso vediamo cosa accade a fronte di diversi valori del parametro λ:

- Se $\lambda = 0$:

  ci riconduciamo alla soluzione standard, ossia massimizzazione del likelihood($\mathbf{w}$) $\rightarrow$ $\hat{\mathbf{w}}_{MLE}$

- Se $\lambda \rightarrow \infty$:

  per soluzioni dove $\hat{\mathbf{w}} \neq \mathbf{0}$, il costo totale $\rightarrow - \infty$

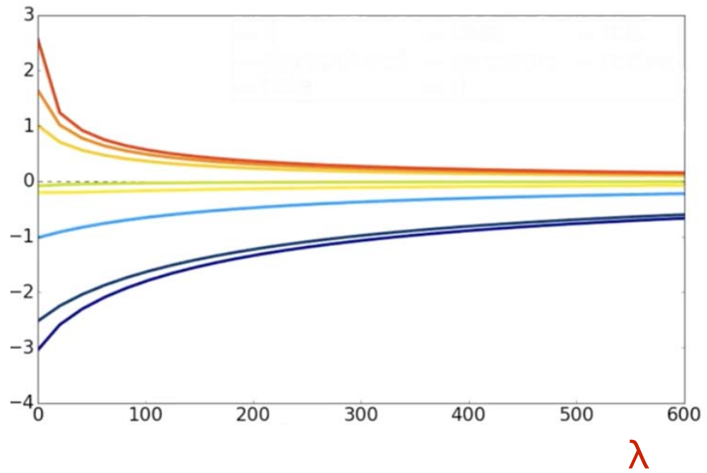
  l'unica soluzione per massimizzare la qualità è: $\hat{\mathbf{w}} = \mathbf{0}$

- Se $0 < \lambda < \infty$:

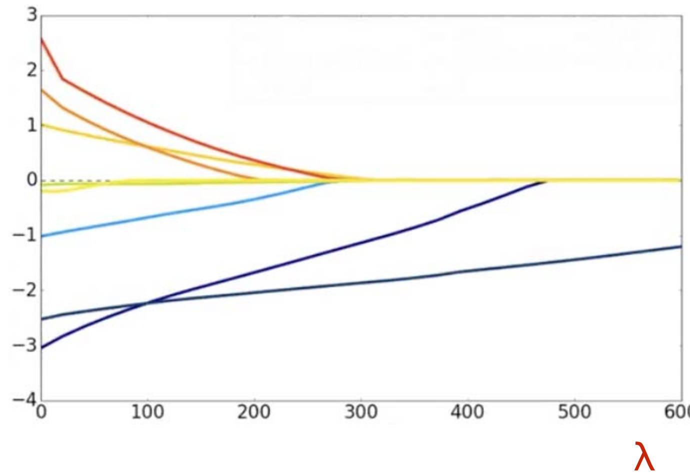
  si va verso soluzioni "sparse", in cui vari w_j sono uguali a zero.

Pesi nella regolarizzazione

- Nelle figure seguenti riportiamo un esempio di andamento dei pesi w_j al variare di λ per i due tipi di penalty:



L2 Penalty



L1 Penalty

Riferimenti

- Watt, J., Borhani, R., Katsaggelos, A.K. *Machine Learning Refined*, 2nd edition, Cambridge University Press, 2020.
- James, G., Witten, D., Hastie, T., Tibishirani, R. *An Introduction to Statistical Learning*, Springer, 2013.
- Ross, S.M. *Probabilità e Statistica per l'Ingegneria e le Scienze*, 3a edizione, Apogeo, 2015.
- *Machine Learning: Classification*, University of Washington - Coursera, 2017.
- Flach, P. *Machine Learning - The Art and Science of Algorithms that Make Sense of Data*, Cambridge University Press, 2012.