# Il Problema dell'Overfitting nei Modelli di Apprendimento Automatico

DOI: 10.69109/NLD2_MA

di Alessandro Manna, *DIITET–CNR*
Roma

### *Introduzione*

Nell'ambito dell'apprendimento automatico e della statistica, l'overfitting rappresenta un problema fondamentale che limita l'efficacia dei modelli predittivi. L'overfitting si verifica quando un modello impara eccessivamente dai dati di addestramento, catturando anche il rumore e le caratteristiche irrilevanti del dataset. Di conseguenza, il modello perde la capacità di generalizzare su nuovi dati, generando così, in contesti reali, previsioni imprecise.
Questo problema è particolarmente rilevante nell'era moderna, dove l'intelligenza artificiale e il machine learning trovano applicazione in settori come la medicina, la finanza e la guida autonoma. Un modello che sovradimensiona l'importanza dei dati di training può portare a decisioni errate e poco affidabili. In questo articolo analizzeremo brevemente le cause principali dell'overfitting, i suoi effetti e le strategie più efficaci per mitigarne l'impatto.

### *Cause dell'Overfitting*

L'overfitting è il risultato di una combinazione di fattori legati alla complessità del modello e alle caratteristiche dei dati utilizzati per l'addestramento.
**Modelli eccessivamente complessi**: Un modello con troppi parametri può adattarsi perfettamente ai dati di training, memorizzando anche le oscillazioni casuali (rumore). Questo accade quando la complessità del modello supera quella dei dati a disposizione.
**Dati insufficienti o rumorosi**: La scarsità di dati è un fattore determinante nell'overfitting. Quando i dati sono limitati, il modello tende ad adattarsi in modo troppo specifico al campione di training. Inoltre, la presenza di rumore o valori anomali può ingannare il modello, portandolo a riconoscere pattern inesistenti.
**Sovra-allenamento del modello**: Il training prolungato su un dataset senza controllo può far sì che il modello ottimizzi eccessivamente le sue previsioni sui dati di addestramento, a discapito della capacità di generalizzazione.

### *Effetti dell'Overfitting*

L'overfitting ha conseguenze dirette sulla capacità predittiva e sull'affidabilità dei modelli.

**Scarsa generalizzazione**: Un modello affetto da overfitting ottiene risultati eccellenti sui dati di training, ma fallisce nell'applicazione su dati nuovi (test set). Ad esempio, un modello di riconoscimento delle immagini potrebbe memorizzare caratteristiche specifiche delle immagini di training, senza imparare le proprietà generali delle classi da riconoscere.
**Errori in scenari reali**: Nei contesti pratici, dove il modello deve prendere decisioni basate su dati non conosciuti, l'overfitting può portare a risultati errati e inefficaci. In applicazioni come la medicina, questo potrebbe significare diagnosi inaccurate, mentre nella finanza potrebbe tradursi in previsioni economiche non affidabili.

### *Soluzioni per Prevenire l'Overfitting*

Esistono diverse strategie per mitigare l'overfitting e migliorare la capacità di generalizzazione dei modelli:

1. **Regolarizzazione**: La regolarizzazione introduce una penalizzazione per la complessità del modello. Tecniche come L1 (lasso regression) e L2 (ridge regression) riducono l'importanza di parametri non significativi, impedendo al modello di adattarsi troppo ai dati di training.
2. **Cross-validation**: La validazione incrociata consiste nel suddividere il dataset in più sottoinsiemi per valutare il modello su diverse combinazioni di dati di training e test. Questo metodo fornisce una stima più robusta delle performance del modello.
3. **Early stopping**: Durante l'addestramento, il monitoraggio delle performance sul set di validazione permette di interrompere l'allenamento quando l'errore su quest'ultimo inizia ad aumentare, evitando il sovra-allenamento del modello.
4. **Raccolta di dati aggiuntivi**: Un modo semplice ma spesso efficace per mitigare l'overfitting è raccogliere più dati di qualità. Con un dataset più ampio, il modello può imparare pattern più generali.
5. **Dropout nelle reti neurali**: Nei modelli di deep learning, la tecnica del dropout consiste nel disattivare casualmente un certo numero di neuroni durante l'addestramento. Questo impedisce al modello di dipendere troppo da specifici percorsi di apprendimento, migliorando la sua generalizzazione.

### *Conclusione*

Una delle conclusioni principali riguarda la necessità di trovare un equilibrio tra la complessità del modello e la sua capacità di generalizzare su dati nuovi. L'overfitting è un promemoria che modelli

troppo sofisticati, se non adeguatamente regolati, possono portare a soluzioni inadatte per il mondo reale. La generalizzazione è infatti la chiave del successo di qualsiasi modello predittivo. Il bilanciamento tra complessità del modello e capacità di generalizzazione rimane dunque una sfida cruciale nel campo dell'intelligenza artificiale, richiedendo un approccio attento e metodologico per garantire risultati robusti e affidabili.

L'overfitting evidenzia il ruolo cruciale dei dati nella costruzione dei modelli. Quantità e qualità dei dati sono determinanti per l'efficacia delle previsioni. Per migliorare le prestazioni e mitigare l'overfitting, è essenziale lavorare su dataset più completi, diversificati e privi di rumore.

Le tecniche illustrate (regolarizzazione, early stopping, cross-validation, ecc.) dimostrano che l'overfitting può essere controllato con un approccio metodico. Tuttavia, la scelta della soluzione più adeguata dipende dal contesto specifico e dal tipo di modello utilizzato.

L'overfitting non è solo un concetto teorico ma ha implicazioni pratiche significative. Ad esempio, in applicazioni critiche come diagnosi mediche, previsioni economiche o sistemi di guida autonoma, un modello incapace di generalizzare può portare a errori gravi con conseguenze economiche, etiche e sociali. Per questo, il controllo dell'overfitting non è solo un problema tecnico, ma anche una responsabilità applicativa.

Guardando al futuro, la sfida principale è quella di sviluppare modelli che siano allo stesso tempo potenti e robusti, in grado di sfruttare la crescente disponibilità di dati senza però cadere nella trappola dell'overfitting. La combinazione di tecniche avanzate, come ad esempio regolarizzazione automatica o ensemble learning, e di strumenti tecnologici sempre più performanti continuerà a evolversi per affrontare il problema.

In sintesi, l'overfitting rappresenta dunque uno dei principali ostacoli nell'apprendimento automatico, poiché limita la capacità dei modelli di operare efficacemente su dati sconosciuti. Le cause principali includono la complessità eccessiva dei modelli, la scarsità di dati e il sovra-allenamento. Le conseguenze di un modello affetto da overfitting sono gravi, in quanto compromettono l'affidabilità delle previsioni in applicazioni pratiche.

L'overfitting ci ricorda quindi un principio fondamentale dell'apprendimento automatico: "un modello perfetto sul training set non è necessariamente un modello utile". La capacità di generalizzare è ciò che distingue un buon modello da uno inefficace. Questo è il principio che deve guidare gli sviluppatori di modelli nella ricerca di soluzioni applicabili e affidabili.

***Bibliografia***


- Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.
- Bishop, C.M. (2006). Pattern Recognition and Machine Learning. Springer.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.
- Abu-Mostafa, Y.S., Magdon-Ismail, M., & Lin, H.T. (2012). Learning from Data: A Short Course. AMLBook.

- Murphy, K. P. (2012). Machine Learning: A Probabilistic Perspective. MIT Press.
- Ng, A. (2018). Regularization Techniques for Overfitting. Coursera - Machine Learning Specialization.
- Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. Springer.
- Russell, S.J., & Norvig, P. (2020). Artificial Intelligence: A Modern Approach (4th ed.). Pearson.
- Zhang, Y., & Yang, Q. (2018). "A Survey on Multi-Task Learning." IEEE Transactions on Knowledge and Data Engineering, 34(3), 1-20.
- Scikit-learn Documentation. (2023). Overfitting and Underfitting. Retrieved from: https://scikit-learn.org


.