MODULO 07 · AGOSTO 2026

Fondamenti di Machine Learning per il Forenser

Capire come apprende una macchina, quanto sbaglia e perché:
il livello di competenza necessario per usarla in giudizio

precision • recall • F1 • ROC • overfitting • XAI

Corso di Laurea Magistrale in Sicurezza Informatica

Al termine del modulo lo studente sarà in grado di

01 · DISTINGUERE

I paradigmi di apprendimento (supervisionato, non supervisionato, deep learning) e il ruolo dei dati di addestramento

02 · QUANTIFICARE

L'errore di un classificatore con precision, recall, F1 e curve ROC, e tradurlo in termini di falsi colpevoli e prove mancate

03 · SPIEGARE

Una decisione algoritmica con strumenti di explainable AI (LIME, SHAP) al livello richiesto da una relazione peritale

Traduzione forense delle metriche. In ambito investigativo un falso positivo può significare un innocente attenzionato; un falso negativo, una prova mai trovata. La scelta della soglia operativa è quindi una decisione di politica dell'errore.

SCHEMA DEL MODULO

L'anatomia di un sistema di machine learning

01 · DATI

La materia prima

- Dataset e distribuzioni
- Feature engineering
- Qualità ed etichette

02 · MODELLI

L'apprendimento

- Paradigmi
- Addestramento
- Overfitting

03 · VALUTAZIONE

La misura dell'errore

- Metriche
- Matrice di confusione
- Calibrazione

04 · COMPRENSIONE

Aprire la scatola

- Clustering
- Spiegabilità (XAI)
- LLM e transformer

Le quattro aree si susseguono nelle slide del modulo nello stesso ordine dello schema: ciascuna presuppone la precedente.

PARADIGMI DI APPRENDIMENTO

Il modello vale quanto i dati che lo hanno istruito

Supervisionato

Apprende da esempi etichettati («questo è malware, questo no»). Usi forensi: classificazione di contenuti, triage, phishing.

Dipendenza critica: qualità e rappresentatività delle etichette.

Non supervisionato

Trova strutture senza etichette: clustering di documenti, anomaly detection su log e timeline, reti relazionali.

Dipendenza critica: la definizione di «normale» (baseline).

Deep learning

Reti neurali profonde che apprendono le rappresentazioni dai dati grezzi: visione (deepfake detection), linguaggio (NLP, LLM).

Dipendenza critica: l'opacità delle decisioni.

Overfitting e generalizzazione. Un modello che memorizza il training set generalizza male sui casi reali: il perito chiede sempre *su che cosa* sono state misurate le prestazioni dichiarate.

Il dataset è il destino del modello

Tre proprietà del dataset determinano, prima di ogni scelta architetturale, ciò che il modello potrà fare:

- **Qualità delle etichette**: se gli annotatori sbagliano o divergono, il modello apprende l'errore come verità; il tetto delle prestazioni è l'accordo tra annotatori
- **Rappresentatività**: il modello generalizza solo verso ciò che somiglia a quanto ha visto; i casi reali fuori distribuzione producono errori silenziosi
- **Bilanciamento**: classi rare nel dataset restano rare nelle capacità, proprio ciò che in forensics interessa (i reperti rilevanti sono di solito una minoranza)

ESEMPIO · IL CLASSIFICATORE «OTTIMO» CHE FALLISCE

Un classificatore di phishing addestrato su email aziendali in inglese dichiara il 97% di accuratezza. Applicato a una casella italiana con gergo settoriale, segnala come phishing le newsletter legittime e lascia passare le frodi scritte bene.

Il modello opera fuori dalla distribuzione su cui è stato addestrato, e il 97% non vale per questi dati.

Domanda peritale standard, da porre a ogni fornitore e a sé stessi: su che cosa è stato addestrato, su che cosa validato, e quanto il mio caso somiglia a quei dati? Senza risposte documentate non si fa affidamento sullo strumento.

APPROFONDIMENTO · FEATURE ENGINEERING

Dal reperto al vettore: le feature forensi

Un modello elabora **vettori di numeri**. Il feature engineering traduce il reperto in vettore ed è il punto in cui la competenza forense entra nel sistema: quali proprietà catturare e come misurarle. Da un eseguibile: entropia delle sezioni, API importate, dimensioni; da una sessione di rete: durate, byte, periodicità.

Le scelte hanno conseguenze probatorie: feature interpretabili (entropia, frequenze) rendono il modello **spiegabile in aula**, mentre le rappresentazioni apprese (embedding, modulo 9) sono più difficili da verbalizzare. La **documentazione delle feature** (definizione, unità, valori mancanti) è parte della riproducibilità.

Una feature che non si sa spiegare a un giudice in una frase diventa un punto debole nel controesame.

ESEMPIO · L'ENTROPIA CHE CLASSIFICA

Per distinguere file cifrati da file compressi, un classificatore usa entropia media, varianza per blocco e test di uniformità sui byte.

Tre feature, tutte spiegabili: «il contenuto cifrato è statisticamente indistinguibile dal casuale su ogni blocco; il compresso conserva strutture». In aula la spiegazione regge, e il modello può essere ricostruito da un perito di parte con le stesse tre misure.

APPROFONDIMENTO · OVERFITTING

Overfitting: quando il modello memorizza i dati

Un modello troppo capace rispetto ai dati impara anche il rumore: aderisce perfettamente agli esempi visti e fallisce sui nuovi. Il sintomo è il **divario tra prestazione in addestramento e in validazione**, e per misurarlo correttamente serve disciplina sperimentale:

- **Train / validation / test**: tre insiemi separati, e il test si tocca una sola volta, alla fine
- **Cross-validation**: la stima si media su più suddivisioni
- **Nessuna fuga di informazione** (leakage): lo stesso caso, o suoi duplicati, mai da entrambe le parti del confine

ESEMPIO · I NUMERI CHE TRADISCONO

```
Accuratezza (train): 99,8%
Accuratezza (validazione): 71,2%
```

Ventotto punti di divario: il modello ha memorizzato il training set. Presentata senza il dato di validazione, la cifra «99,8%» è fuorviante.

In un capitolato o in una relazione, la domanda è sempre: accuratezza **su quali dati**?

Variante forense del leakage: addestrare e validare su reperti dello stesso caso. Il modello impara a riconoscere quel caso specifico e crolla sul caso successivo.

METRICHE DI ERRORE E SPIEGABILITÀ

L'errore quantificato è l'unico errore difendibile

Metrica	Lettura forense
Precision	Dei reperti segnalati, quanti sono davvero rilevanti, ossia il costo dei falsi allarmi
Recall	Dei reperti rilevanti, quanti sono stati trovati, ossia il costo delle prove mancate
F1	Sintesi dei due: utile al confronto, opaca sulle priorità
Curva ROC / AUC	Prestazione al variare della soglia: la soglia operativa è una scelta, da motivare

Accuracy da sola inganna su classi sbilanciate: con l'1% di reperti rilevanti, un modello che risponde sempre «irrilevante» è accurato al 99% e investigativamente inutile.

Explainable AI

LIME e **SHAP** attribuiscono la decisione alle feature che l'hanno determinata, con una spiegazione locale e approssimata che si può verbalizzare.

Per la relazione peritale: la spiegazione motiva perché il modello ha indicato un esito; la correttezza dell'esito si prova con la validazione.

APPROFONDIMENTO · LA MATRICE DI CONFUSIONE

Un esempio svolto su quattro numeri

Triage su 10.000 file, di cui 200 davvero rilevanti. Esito del classificatore:

	Reale: rilevante	**Reale: irrilevante**
Predetto: rilevante	170 (veri positivi)	330 (falsi positivi)
Predetto: irrilevante	30 (falsi negativi)	9.470 (veri negativi)

```
Precision = 170/(170+330) = 34%
Recall = 170/(170+30) = 85%
Accuratezza = 9.640/10.000 = 96,4%
```

Lettura forense

L'accuratezza al 96,4% suona eccellente, ma è quasi irrilevante perché dominata dai 9.470 negativi facili.

I numeri che contano: 2 file segnalati su 3 sono falsi allarmi (precision 34%), con un costo in ore-analista; e **30 reperti rilevanti sfuggono** (recall 85%), con un costo in prove mancate da coprire con la validazione su campione del declassato.

In relazione va riportata la matrice intera: ogni singola metrica omette una parte dell'errore.

APPROFONDIMENTO · APPRENDIMENTO NON SUPERVISIONATO

Clustering e riduzione dimensionale

K-means

Partiziona i dati in **k gruppi sferici** attorno a centroidi. Veloce e leggibile, ma k va scelto a priori e ogni punto finisce in un cluster, anche il rumore. Utile per raggruppare file o sessioni simili in un triage.

DBSCAN / HDBSCAN

Cluster per **densità**: trovano gruppi di forma arbitraria e, soprattutto, lasciano fuori gli outlier. In chiave forense gli outlier sono spesso proprio gli elementi che l'analista deve esaminare per primi.

PCA / UMAP

Proiettano centinaia di feature in 2-3 dimensioni **visualizzabili**. Ottime per esplorare e comunicare; ma le distanze nella proiezione distorcono quelle reali: mai trarre conclusioni probatorie dalla sola mappa.

ESEMPIO · 200.000 immagini sequestrate, prive di etichette. Gli embedding visivi, proiettati con UMAP e raggruppati con HDBSCAN, formano cluster per tipologia: l'analista esamina un campione per cluster anziché 200.000 file e dà priorità ai 300 outlier. Ogni elemento rilevante è poi verificato da una persona.

APPROFONDIMENTO · EXPLAINABLE AI

LIME e SHAP: come si spiega una scatola nera

LIME

Perturba l'input attorno al caso in esame, osserva come cambia la predizione e costruisce un **modello semplice locale** che la approssima. Risponde a: «che cosa, in *questo* caso, ha pesato?»

Approssimazione locale: spiegazioni diverse per casi vicini sono possibili.

SHAP

Fondato sui valori di Shapley della teoria dei giochi: ripartisce la predizione tra le feature come «contributi», con proprietà di coerenza che LIME non garantisce. Il calcolo è più oneroso, ma i risultati sono più stabili e aggregabili sull'intero dataset.

ESEMPIO · IN RELAZIONE

«Il classificatore segnala l'email come phishing (punteggio 0,91). Secondo SHAP pesano soprattutto il dominio del mittente, registrato da 3 giorni, e la discordanza tra testo e destinazione del link: entrambi riscontrati manualmente.»

Il riscontro manuale, orientato da SHAP, fonda la conclusione.

Limite da dichiarare: la spiegazione descrive il comportamento del modello e ne riproduce anche le distorsioni; la correttezza dell'esito si dimostra con la validazione.

APPROFONDIMENTO · CALIBRAZIONE

Quando «0,92» significa davvero 92%

Molti modelli emettono uno «score» fra 0 e 1 che **non è una probabilità**: le reti neurali moderne tendono a essere sovraconfidenti, e uno 0,92 può corrispondere a un tasso di correttezza reale del 70%. Un modello è **calibrato** quando, fra tutti i casi a cui assegna 0,9, circa il 90% è davvero positivo.

La calibrazione si misura (reliability diagram, Expected Calibration Error) e si corregge (Platt scaling, isotonic regression, temperature scaling) su un insieme di validazione separato. In ambito forense è indispensabile, perché **il numero riportato in relazione verrà letto come una probabilità** da giudici e parti.

Alternativa prudente: riportare, al posto dello score grezzo, fasce qualitative definite su dati di validazione («alta/media/bassa compatibilità») con le frequenze d'errore di ciascuna fascia.

ESEMPIO · LO 0,97 IN AULA

Un detector di manipolazione assegna 0,97 a un'immagine. Il PM lo legge come «97% di probabilità che sia falsa».

Il consulente della difesa esibisce il reliability diagram del modello: nella fascia 0,95–1,00 il tasso reale di manipolazioni è dell'81%, misurato però su un dataset che non include il modello di fotocamera in questione. Le obiezioni fondate sono quindi due: calibrazione imperfetta e dominio non coperto dalla validazione.

Come funziona un LLM e quali limiti ha

Un LLM genera la continuazione più plausibile di un testo, appresa da enormi corpora. L'architettura **transformer** (meccanismo di attenzione) ne è il motore; il criterio di generazione è la plausibilità statistica, che non garantisce la verità.

Conseguenza forense diretta: l'**allucinazione** (un'affermazione fluente e specifica ma falsa) è un comportamento atteso del meccanismo. Ogni output va trattato come ipotesi da verificare.

Lessico essenziale

- **Token e contesto**: l'unità di elaborazione e la finestra di memoria
- **Fine-tuning**: specializzazione su un dominio
- **RAG**: ancorare le risposte a documenti forniti
- **Temperatura**: quanto l'output varia a parità di input; la ripetibilità ne dipende

Da questo modulo in poi nessun reperto reale o dato personale viene inviato a servizi di AI esterni non controllati, perché ciò comprometterebbe la catena di custodia. Gli esercizi usano ambienti locali o dataset pubblici.

LABORATORIO · LETTURE · VERIFICA

Addestrare un modello e verificarne i limiti

Laboratorio

- Classificatore scikit-learn su artefatti etichettati
- Metriche e matrice di confusione
- Spiegazioni **SHAP** e analisi degli errori

`DELIVERABLE: scheda di validazione del modello`

Letture obbligatorie

- Géron, *Hands-On Machine Learning*, capitoli introduttivi
- Survey DFRWS su AI e digital forensics
- Scheda introduttiva su LIME e SHAP

Verifica

Si valuta la capacità di individuare punti deboli, bias plausibili e condizioni di validità del proprio modello.

Prossimo modulo: AI applicata alla forensics (triage e classificazione).

> Il perito deve saper interrogare un sistema di AI con le domande di chi dovrà risponderne in aula.

Grazie per l'attenzione

Vincenzo Calabrò

Domande, materiali del corso e contatti sul sito del docente

vincenzocalabro.it