Algoritmi per l'intelligenza artificiale

Simone Colli
authoremail

Appunti del corso tenuto dal **Prof. Vincenzo Bonnici**

Università degli Studi di Parma
Anno Accademico 2025/2026

Indice

1 Introduzione **3**
1.1 Apprendimento automatico supervisionato . . . 3
1.2 Apprendimento automatico semi-supervisionato . . . 4
1.2.1 Presupposti dell'apprendimento semi-supervisionato . . . 4
1.3 Apprendimento automatico non supervisionato . . . 4
1.3.1 Clustering . . . 5
1.3.2 Riduzione della dimensionalità . . . 5
1.3.3 Analisi esplorativa . . . 5
1.4 Apprendimento per rinforzo . . . 5

2 Classificazione **6**
2.1 Costruire un classificatore . . . 7
2.2 Proprietà di un classificatore . . . 7
2.3 Il problema dell'overfitting . . . 8
2.4 Validazione . . . 8
2.5 Gestione delle feature e del rumore . . . 8
2.5.1 Selezione delle feature . . . 8
2.5.2 Rumore e outlier . . . 8
2.6 Valutazione degli errori . . . 9
2.7 Fasi di un sistema di classificazione . . . 9

3 Tecniche di validazione per la classificazione **10**
3.1 Modello di validazione base: Training e Test Set . . . 10
3.2 Metriche di valutazione . . . 11
3.3 Accuracy . . . 13
3.4 Altri indici e matrice di confusione . . . 14
3.4.1 Matrice di confusione . . . 14
3.5 Area Under the Curve (AUC) e curva ROC . . . 14
3.6 Classificazione multi-classe . . . 15
3.6.1 Micro-average . . . 15
3.6.2 Macro-average . . . 16
3.6.3 Generalizzazione di AUC (Metodo Hand & Till) . . . 16
3.7 Cross-validation . . . 17
3.7.1 Cross-validazione esaustiva . . . 18
3.7.2 Cross-validazione non esaustiva . . . 18

1 Introduzione

Nel campo dell'apprendimento automatico classico, le attività sono tradizionalmente suddivise in quattro rami principali:

- Apprendimento supervisionato (supervised).
- Apprendimento semi-supervisionato (semi-supervised).
- Apprendimento non supervisionato (unsupervised).
- Apprendimento per rinforzo (reinforcement learning).

La distinzione primaria tra queste metodologie di ML risiede nel livello di disponibilità dei "dati di verità di base" (ground truth). Il **ground truth** è definito come la conoscenza preliminare dell'output che il modello dovrebbe produrre per un dato input, basata sull'osservazione diretta in contrapposizione all'inferenza.

1.1 Apprendimento automatico supervisionato

L'apprendimento automatico supervisionato ha come obiettivo l'apprendimento di una funzione che, dato un **campione di dati** e i relativi output desiderati, riesca ad approssimare la funzione sottostante che mappa gli input agli output.

Questa metodologia è comunemente applicata in due principali contesti:

- **Classificazione**: quando si desidera mappare l'input a etichette di output discrete.
- **Regressione**: quando l'obiettivo è mappare l'input a un output continuo.

In entrambi i casi, lo scopo è identificare relazioni o strutture specifiche nei dati di input che consentano di generare output corretti in modo efficace. È fondamentale notare che la correttezza dell'output è determinata interamente dai dati di addestramento, i quali costituiscono la "verità di base" che il modello apprende.

Tuttavia, l'efficacia del modello può essere significativamente ridotta dalla presenza di etichette "rumorose" o "errate" all'interno dei dati stessi. Algoritmi notevoli nell'apprendimento supervisionato includono la regressione logistica, il classificatore bayesiano naif, le macchine a vettori di supporto, le reti neurali artificiali e le foreste casuali.

Il successo di un modello di ML dipende dalla sua capacità di generalizzazione. Questo concetto è strettamente connesso alla complessità del modello, che si riferisce alla complessità della funzione che si sta cercando di apprendere. Se si dispone di una quantità limitata di dati o se questi non sono distribuiti uniformemente, è cruciale optare per un modello a bassa complessità per evitare situazioni di **overfitting** (sovradattamento). L'overfitting si verifica quando il modello apprende la funzione adattandosi troppo bene ai soli dati di addestramento, senza cogliere la tendenza o la struttura effettiva che guida l'output, e quindi non riesce a generalizzare a nuovi punti dati.

La gestione della generalizzazione è formalizzata tramite il compromesso **bias-varianza** (bias-variance tradeoff). Così facendo il modello presenterà un equilibrio tra:

- **Bias** (distorsione): l'errore sistematico dovuto a ipotesi errate nel processo di apprendimento.
- **Varianza**: la quantità in base alla quale l'errore può variare tra diversi set di dati.

La difficoltà si presenta nel creare un modello che cattura accuratamente le regolarità dei dati di addestramento e che sia in grado di generalizzare bene a dati non visti in precedenza.

Generalmente, un aumento del bias (e una conseguente riduzione della varianza) porta a modelli con livelli di prestazione più stabili e garantiti, un fattore che può essere cruciale in certe applicazioni. Per ottenere una buona generalizzazione, la varianza del modello deve essere attentamente bilanciata in base alla dimensione e alla complessità dei dati di addestramento. Nello specifico, set di dati piccoli e semplici dovrebbero essere gestiti con modelli a bassa varianza, mentre set di dati grandi e complessi richiedono modelli con una varianza più elevata per poter catturare appieno la struttura sottostante dei dati.

1.2 Apprendimento automatico semi-supervisionato

L'apprendimento automatico semi-supervisionato (semi-supervised) mira a **etichettare i punti dati senza etichetta**. Per fare ciò, utilizza le conoscenze apprese da un piccolo numero di dati già etichettati.

Questa tecnica è utile in scenari dove ottenere dati etichettati è costoso o complesso. Ad esempio, nel rilevamento di messaggi inappropriati in un social network, è impraticabile etichettare manualmente ogni messaggio. Si può, invece, etichettare manualmente un piccolo sottoinsieme e usare tecniche semi-supervisionate per comprendere e classificare il resto dei contenuti.

Metodi comuni includono le **macchine vettoriali di supporto trasversali** e i **metodi basati su grafi** (come la propagazione delle etichette).

1.2.1 Presupposti dell'apprendimento semi-supervisionato

Per poter giustificare l'uso di pochi dati etichettati per trarre conclusioni su un grande insieme di dati non etichettati, i metodi semi-supervisionati si basano su alcuni presupposti fondamentali:

- **Continuità**: Si assume che punti dati "vicini" tra loro abbiano maggiori probabilità di condividere la stessa etichetta.
- **Ipotesi del cluster**: Si presume che i dati formino naturalmente dei cluster discreti. Di conseguenza, punti nello stesso cluster hanno maggiori probabilità di condividere un'etichetta.
- **Presupposto molteplice (manifold)**: Si ipotizza che i dati si trovino approssimativamente in uno spazio di dimensioni inferiori (un *manifold*) rispetto allo spazio di input originale. Questo è rilevante quando un sistema con pochi parametri, non osservabile direttamente, produce output osservabili ad alta dimensione.

1.3 Apprendimento automatico non supervisionato

L'apprendimento automatico non supervisionato (unsupervised) opera **senza output etichettati**. Il suo obiettivo principale è quindi quello di **dedurre la struttura naturale** presente all'interno di un insieme di dati.

Questi metodi cercano di trovare modelli (pattern) intrinseci nei dati. Le attività più comuni in questo ambito sono:

- Il **clustering** (raggruppamento).
- L'apprendimento della **rappresentazione** (representation learning).

- La **stima della densità** (density estimation).

In tutti questi casi, si desidera comprendere la struttura intrinseca dei dati senza usare etichette fornite esplicitamente.

Algoritmi comuni includono il **clustering**, l'analisi dei componenti principali (**PCA**) e gli **auto-codificatori** (autoencoders).

Dato che non vengono fornite etichette, nella maggior parte dei metodi di apprendimento non supervisionato non esiste un modo specifico per confrontare le prestazioni del modello.

Le due tecniche principali per affrontare problemi di apprendimento non supervisionato sono il clustering e la riduzione della dimensionalità dei dati.

1.3.1 Clustering

Il clustering è una **tecnica esplorativa** che permette di aggregare dati in gruppi (detti *cluster*) senza avere una precedente conoscenza della loro appartenenza a tali gruppi. Si applica a dataset dove i dati al loro interno presentano elementi simili tra loro. All'interno di ogni singolo cluster si troveranno quindi dati che hanno molte **caratteristiche simili** tra loro. È un'ottima tecnica per trovare relazioni tra i dati.

1.3.2 Riduzione della dimensionalità

La riduzione della dimensionalità senza supervisione è un approccio molto usato nella **pre-elaborazione delle features**. L'obiettivo principale di questa tecnica è di **eliminare il "rumore"** dai dati.

Questa operazione può talvolta causare una minore prestazione predittiva. Tuttavia, può anche rendere lo spazio dimensionale più compatto, aiutando a **mantenere le informazioni più rilevanti**. Inoltre, è molto utile per la **rappresentazione dei dati**: dati in uno spazio delle caratteristiche ad elevata dimensionalità possono essere proiettati su uno spazio 1D, 2D o 3D per l'analisi visiva.

1.3.3 Analisi esplorativa

L'apprendimento non supervisionato è estremamente utile nell'**analisi esplorativa dei dati** (exploratory data analysis), poiché è in grado di **identificare automaticamente la struttura** nei dati. Ad esempio, se un analista volesse segmentare i consumatori, i metodi di clustering sarebbero un ottimo punto di partenza per l'analisi.

In situazioni dove è impraticabile o impossibile per un essere umano proporre tendenze nei dati, l'apprendimento non supervisionato può fornire **informazioni iniziali** che possono poi essere usate per testare o verificare singole ipotesi.

1.4 Apprendimento per rinforzo

L'apprendimento con rinforzo (reinforcement learning) ha l'obiettivo di realizzare **agenti autonomi**. Questi agenti devono essere in grado di scegliere azioni da compiere per conseguire determinati obiettivi. Questo avviene tramite l'interazione con l'ambiente in cui sono immersi, con lo scopo di massimizzare una nozione di **premio cumulativo**.

2 Classificazione

La classificazione è un'attività dell'apprendimento supervisionato che consiste nell'assegnare un'etichetta (o classe) a un dato sulla base di sue caratteristiche osservabili.

Nell'ambito della classificazione si parla di:

- **Feature** (caratteristiche): un aspetto direttamente osservabile di un fenomeno per il quale si può registrare una misura, che sia quantitativa (numerica) o categoriale (come vero/falso, rosso/verde, ecc.).
- **Classe**: un concetto astratto e generale che "spiega" le osservazioni. L'assegnazione a una classe costituisce una sintesi delle feature osservate.
- **Label** (o etichetta): il nome specifico di una classe.

Tuttavia, alcuni dati possono rendere più complesso l'assegnazione delle classi; questi esempi sono tecnicamente noti come **outlier statistici**.

Definizione 2.1: Classificazione

ata una **collezione di dati**, definita come un insieme P di M-uple del tipo:

$$m_i = (x_{1i}, \ldots, x_{Mi}) \in D_1 \times \ldots \times D_M$$

dove ogni x_{ji} rappresenta una feature ed appartiene ad un possibile dominio di valori D_j.
L'insieme P è partizionato in k classi, le cui etichette compongono l'insieme $L = (A_1, \ldots, A_k)$.
Un **algoritmo di classificazione** è una funzione compputabile $f : P \mapsto L$, tale che:

$$f(m \in P) = f(x_1, \ldots, x_m) \in L$$

Tale funzione $f(m)$ assegna a ogni dato m un'etichetta A_i scelta tra quelle presenti in L cercando di stimare l'etichetta reale del stesso.
Lo schema di classificazione può produrre due tipi di risultati:

- **Successo** (hit) se l'etichetta stimata $f(m)$ coincide con l'etichetta reale del dato.
- **Fallimento** (miss) se l'etichetta stimata è errata.

È generalmente impossibile creare classificatori *error free*. È quindi fondamentale fornire stime sul tasso percentuale di hit/miss che lo schema può ottenere. Il livello tollerabile di errore dipende dalla **criticità dell'applicazione**: per applicazioni industriali si può richiedere un tasso $< 5\%$, mentre per applicazioni mediche un tasso $> 0.5\%$ potrebbe essere già inaccettabile.

Esempio 2.1: Problema di classificazione: salmoni e branzino

Si consideri il problema di distinguere tra salmoni e branzini (sea bass) basandosi su alcune caratteristiche osservabili. Le **feature** utilizzate potrebbero essere la lunghezza, il peso in grammi e il colore dominante (un attributo qualitativo scelto da un insieme predefinito come {blu, grigio, verde}). I dati vengono tipicamente organizzati in una tabella, dove ogni riga corrisponde a un pesce e le colonne ne descrivono le feature.

L'obiettivo è costruire un classificatore che, per ogni nuovo pesce osservato, sia in grado di riempire la colonna "specie" con l'etichetta corretta ("salmone" o "branzino"). È importante notare che gli errori non hanno lo stesso costo: confondere un salmone (pesce pregiato) con un branzino (meno pregiato) è un errore più grave del contrario.

Esempio 2.2: Problema di classificazione: studenti e carriera

Si consideri il problema di predire il futuro successo economico degli studenti universitari. Le **feature** raccolte per ogni studente includono dati anagrafici, il censo familiare e i voti conseguiti durante la carriera universitaria. L'obiettivo è costruire un classificatore che predica in quale **classe** di reddito si troverà lo studente dieci anni dopo la laurea. Le etichette (o **label**) potrebbero essere {"reddito basso", "reddito medio", "reddito alto"}.
È importante notare che, a causa dell'elevato numero di fattori non misurabili che influenzano la vita di un individuo, una predizione del genere ha un valore limitato se applicata al singolo studente, che ha un'alta probabilità di essere classificato erroneamente.
Tuttavia, questo tipo di analisi è estremamente utile a livello statistico e aggregato, per comprendere le tendenze generali di un'intera popolazione studentesca e informare politiche educative o economiche.

2.1 Costruire un classificatore

Il processo di costruzione di un classificatore automatico simula il fenomeno dell'apprendimento umano o animale, noto come **training** (addestramento). L'idea è **dedurre regole generali**, applicabili a record non ancora classificati, partendo dall'osservazione di esempi già noti e ben classificati.

Si definisce **universo delle osservazioni** l'insieme complessivo dei record (passati, presenti e futuri) relativi ad un fenomeno. Molti algoritmi iniziano esaminando un sottoinsieme di questo universo, già classificato e ben compreso.

Questo insieme di "allenamento", chiamato **Training Set (TS)**, è il deposito di informazioni iniziali da cui l'algoritmo ricava le "regole" di classificazione. Le regole ricavate saranno di vario tipo: statistiche, probabilistiche, fuzzy, funzioni discendenti, ecc.

2.2 Proprietà di un classificatore

Un buon insieme di regole di classificazione deve avere tre importanti proprietà:

- **Semplicità**: Le regole non devono essere troppo complicate, per garantire efficienza e basso costo computazionale in fase di classificazione.
- **Correttezza sul TS**: Le regole devono essere statisticamente sufficientemente corrette quando applicate al medesimo Training Set che le ha generate.
- **Generalizzabilità**: Le regole devono essere statisticamente corrette anche quando applicate al resto dei record dell'universo (dati nuovi, non visti).

Statisticamente corretto è un termine che indica che il tasso dei miss non deve superare certe soglie di tolleranza che dipendono dalla criticità delle applicazioni.

2.3 Il problema dell'overfitting

Le proprietà di correttezza sul TS e di generalizzabilità sono spesso in conflitto tra loro. Questo paradosso è noto come **overfitting** (sovradattamento).

L'overfitting si verifica quando un modello si adatta "troppo bene" ai dati del Training Set. Un modello molto complesso può imparare a memoria le peculiarità e persino il rumore casuale presente nel TS, ottenendo una correttezza perfetta su di esso. Tuttavia, tale modello non avrà appreso la "tendenza" generale dei dati e fallirà nel generalizzare a nuovi record, poiché la frontiera di decisione che ha appreso è eccessivamente complessa e specifica per il campione di training.

L'obiettivo non è quindi minimizzare l'errore sul TS (che porterebbe a un modello complesso e in overfitting), ma trovare un equilibrio: un modello (es. una retta o una curva semplice) che, pur commettendo qualche errore sul TS, catturi la struttura di fondo dei dati e possa quindi generalizzare meglio.

2.4 Validazione

Per "convalidare" la proprietà di generalizzazione di un insieme di regole, si utilizza un metodo che prevede, oltre al TS, un altro insieme di record già etichettati, detto **Control Set (CS)** o **Test Set**.

Il CS **non** viene utilizzato durante la fase di training (cioè per la sintesi delle regole). Viene usato solo dopo che le regole sono state definite. Se le regole mostrano sul CS un tasso di errore (miss) simile a quello ottenuto sul TS, allora si ritiene che le regole siano **generalizzabili**.

Poiché anche il CS è un campione casuale, per una stima più precisa è buona norma ripetere i test con diversi CS, spesso creati tramite strategie di randomizzazione nella selezione del TS e del CS dall'universo disponibile.

2.5 Gestione delle feature e del rumore

Nella costruzione di un classificatore è cruciale gestire sia la selezione delle feature che la presenza di rumore.

2.5.1 Selezione delle feature

Spesso si rilevano molte feature, ma non tutte sono utili; alcune possono essere sovrabbondanti o addirittura dannose. Combinare più feature (es. lunghezza e luminosità dei pesci) è spesso una strategia conveniente, ma non è detto che sia sempre la migliore. L'inclusione di troppe feature, specialmente se irrilevanti, può amplificare il r̈umoreë confondere il classificatore.

Una buona pratica è scegliere feature che siano **invarianti** alle trasformazioni tipiche della situazione sperimentale (es. il peso di un pesce è invariante alle condizioni di luce, la luminanza no). Inoltre, deve esistere una probabile relazione tra le feature misurate e la classe da predire.

2.5.2 Rumore e outlier

I dati del mondo reale contengono inevitabilmente **rumore**, ovvero perturbazioni dovute a fenomeni non controllabili o non noti. Le cause di tale spostamento dai valori "ideali" possono essere:

- **Endogene**: Interne al fenomeno (es. un pesce con una dieta o storia anomala).
- **Esogene**: Dovute all'osservatore o allo strumento utilizzato (es. macchina fotografica starata, etichettatore distratto).

I dati molto "fuori norma" rispetto ai valori tipici di una classe sono definiti **outlier**. Un buon algoritmo di classificazione deve essere **robusto**, ovvero deve avere una forma di "protezione" o resistenza alle deviazioni che il rumore impone al processo decisionale.

2.6 Valutazione degli errori

Contare gli errori è essenziale, ma una singola percentuale di errore non è sufficientemente descrittiva. Questo perché non tutti gli errori sono uguali: i **costi degli errori** spesso **non sono uniformi** o simmetrici.

Ad esempio, in una diagnosi medica:

- Classificare un sano come malato (Falso Positivo) è un errore con un costo relativamente basso (paura, un test aggiuntivo).
- Classificare un malato come sano (Falso Negativo) è un errore gravissimo, che ritarda la diagnosi e può costare la vita al paziente.

Per analizzare questa asimmetria si usa la **matrice di confusione**. È una griglia quadrata che riporta quante istanze della classe "reale" (sulle colonne) sono state assegnate alla classe "prevista" (sulle righe).

Un classificatore perfetto ha come matrice di confusione la matrice identica (tutti i valori sulla diagonale principale, zero altrove). Un buon classificatore avrà valori percentuali bassi al di fuori della diagonale principale.

2.7 Fasi di un sistema di classificazione

Il processo di classificazione automatica si articola in diverse fasi:

1. **Sensing (o sampling)**: Raccolta dei dati dal mondo fisico e loro digitalizzazione.
2. **Segmentazione**: Partizione dei dati in unità significative, pulizia ed eliminazione di dati irrilevanti.
3. **Estrazione delle feature**: Misurazione delle caratteristiche (quantitative o qualitative). È cruciale scegliere feature invarianti e rilevanti.
4. **Classificazione**: Esecuzione dell'algoritmo scelto per l'assegnazione delle etichette.
5. **Post-processing**: Valutazione della qualità della classificazione e dei costi associati agli errori.
6. **Decisione**: Utilizzo effettivo del classificatore per risolvere il problema reale.

La costruzione di un classificatore è un **ciclo iterativo** che prevede la raccolta dei dati, la selezione delle feature, la scelta di un modello matematico, il training dell'algoritmo e infine la sua valutazione, ripetendo i passi per migliorare le prestazioni.

3 Tecniche di validazione per la classificazione

Dopo aver costruito un modello di classificazione, è fondamentale valutarne le performance. A differenza della regressione, dove si cerca di predire un valore numerico di output dato uno o più valori di input, nella classificazione si vuole predire la classe di un oggetto dato uno o più dati di input di qualsiasi tipo (numerici, categorici, testuali, ecc.).

Per questo motivo gli strumenti che si possono applicare per valutare un modello di classificazione sono sostanzialmente diversi rispetto alle metriche utilizzate per valutare i modelli di regressione.

3.1 Modello di validazione base: Training e Test Set

L'approccio base per la validazione consiste nel dividere, secondo una certa **percentuale** i dati disponibili in due insiemi:

- **Training set (TS)**: Utilizzato per addestrare il classificatore. Le etichette (label) di questi dati sono usate per addestrare il classificatore.
- **Test set**: Utilizzato per valutare la bontà del modello. Le etichette di questo set vengono usate solo per verificare se il classificatore ha predetto correttamente.

L'obiettivo della validazione non è solo misurare l'errore, ma ha lo scopo di rispondere a domande più complesse, come:

- Il classificatore performa in modo bilanciato su tutte le classi?
- Ha delle preferenze?
- Tali preferenze da cosa dipendono?

Per questo motivo per la valutazione dei classificatori si utilizzano indici matematici che permettono sia di avere stime oggettive delle performance, sia di automatizzare anche altre fasi del processo della progettazione o sviluppo del classificatore.

Gli indici principali relativamente all'aspetto computazionale utilizzati per valutare un classificatore sono:

- **Accuratezza**: La bontà nel predire correttamente le etichette.
- **Robustezza**: La capacità di gestire dati con rumore o valori mancanti.
- **Velocità**: Include sia il tempo per costruire il modello (training time) sia il tempo per usarlo (classification/prediction time).
- **Scalabilità**: L'efficienza del modello su grandi dataset, specialmente se in memoria secondaria.
- **Interpretabilità**: La facilità con cui i risultati del modello possono essere compresi.

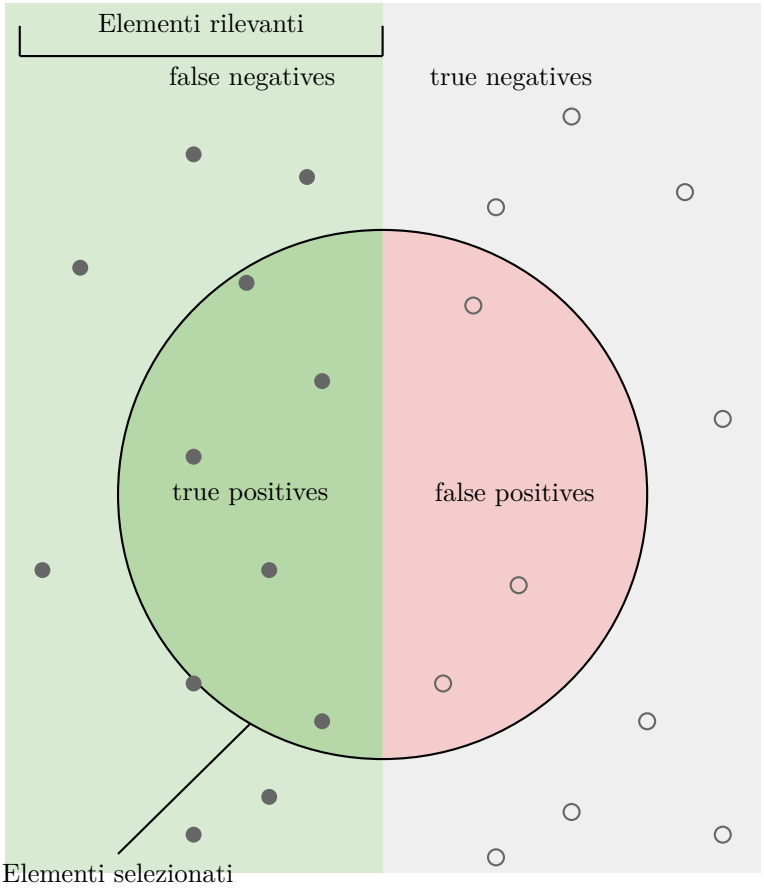
3.2 Metriche di valutazione

Per definire le metriche più comuni, si assume un problema di classificazione binaria. Si assume che l'insieme delle classi $\mathbb{C}$ dell'esperimento sia composto da due classi: $\mathbb{C} = \{A, B\}$.

Relativamente ad una delle classi è possibile definire alcune misure per calcolare la bontà dell'algoritmo in valutare tale classe.

Data una classe di interesse (es. A, la classe "positiva"), i risultati della classificazione sul test set vengono divisi in quattro categorie:

- True positive (TP).
- True negative (TN).
- False positive (FP).
- False negative (FN).



Definizione 3.1: TP, TN, FP, FN

Sia $c : CS \mapsto \mathbb{C}$ la funzione che mappa ogni record $x \in CS$ nella sua classe reale e sia $\tilde{c} : CS \mapsto \mathbb{C}$ il classificatore che assegna una classe ad A.
Sia $C = \{A, B\}$ l'insieme delle classi, composto dalle classi A e B. Prendendo come riferimento la classe A è possibile dividere CS in 4 insiemi:

- **True positive (TP)**: I record $x \in CS$ classificati **correttamente**, ovvero la cui classe reale è A, quindi $\tilde{c}(x) = c(x) = A$
- **True negative (TN)**: I record $x \in CS$ classificati **correttamente**, ovvero la cui classe reale è B, quindi $\tilde{c}(x) = c(x) = B$
- **False positive (FP)**: I record $x \in CS$ classificati **erroneamente**, ovvero la cui classe reale è B, quindi $\tilde{c}(x) = B \neq A = c(x)$
- **False negative (FN)**: I record $x \in CS$ classificati **erroneamente**, ovvero la cui classe reale è B, quindi $\tilde{c}(x) = A \neq B = c(x)$

Basandosi su queste quattro categorie, è possibile definire le metriche di performance più utilizzate.

Definizione 3.2: Precision

Sia $C = \{A, B\}$ l'insieme delle classi, composto dalle classi A e B. La precision (precisione) è la frazione di elementi rilevanti per una classe di riferimento, A, tra tutti gli elementi che il classificatore ha identificato come A. La precision misura quanto è "affidabile" la predizione positiva, ed è definita come:

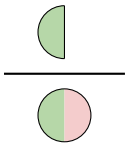
$$Precision = \frac{TP}{TP + FP}$$

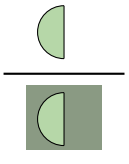
Definizione 3.3: Recall

Sia $C = \{A, B\}$ l'insieme delle classi, composto dalle classi A e B. La recall (richiamo o sensitività) è la frazione di elementi rilevanti (classi A) che sono stati correttamente classificati come A. Misura la capacità del classificatore di "trovare" tutti i positivi.

$$Recall = \frac{TP}{TP + FN}$$

Rappresentazione grafica delle metriche

Precision = 

Recall = 

Nota 3.1: Valori ottenuti

Valori alti per entrambe le metriche indicano un buon classificatore. Spesso, però, si preferisce utilizzare un indice unico che le combini.

Definizione 3.4: F_1-Score

Il F_1-Score rappresenta la media armonica di precision e recall. Fornisce un equilibrio tra le due metriche.

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$

Come precision e recall, anche l'F_1-Score ha un valore compreso tra 0 e 1. Maggiore è il valore, maggiore è la bontà del classificatore.

Nota 3.2: Overfitting

Sebbene un F_1-Score alto sia desiderabile, valori molto prossimi a 1 possono essere un campanello d'allarme per l'overfitting.

3.3 Accuracy

Definizione 3.5: Accuracy

La accuracy (accuratezza) misura la quantità totale di oggetti classificati correttamente (sia positivi che negativi) rispetto al totale degli oggetti.

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

Nota 3.3: Accuratezza per dataset sbilanciati

L'accuracy standard è poco indicata se le classi non sono bilanciate. Ad esempio, in un dataset con 95 campioni negativi e 5 positivi, un classificatore "pigro" che predice sempre "negativo" otterrebbe un'accuratezza del 95%, pur essendo totalmente inutile nel riconoscere i positivi.

In situazione di sbilanciamento delle classi, si preferisce la Balanced accuracy.

Definizione 3.6: Balanced accuracy

La Balanced accuracy (accuratezza bilanciata) è la media tra la sensitività (per i positivi) e la specificità (per i negativi).

$$Balanced\ accuracy = \frac{TPR + TNR}{2}$$

dove:

- **TPR (True Positive Rate)**: È la Recall/Sensitività: $TPR = \frac{TP}{TP+FN}$.
- **TNR (True Negative Rate)**: È la Specificità: $TNR = \frac{TN}{TN+FP}$.

3.4 Altri indici e matrice di confusione

Definizione 3.7: False Discovery Rate (FDR)

Misura il tasso di errori di tipo I ("false scoperte" o Falsi positivi) rispetto a tutte le predizioni positive.

$$FDR = \frac{FP}{FP + TP} = 1 - Precision$$

3.4.1 Matrice di confusione

La matrice di confusione è una tabella che riassume le performance di un classificatore binario, incrociando le classi reali con quelle predette e mostrando i conteggi di TP, TN, FP e FN. È fondamentale perché non tutti gli errori hanno lo stesso costo, come discusso in precedenza (es. diagnosi medica errata).

		Actual class: Positive	Actual class: Negative	
Predicted class: Positive		**TP:** True Positive	**FP:** False Positive (Type I Error)	**Precision:** TP / (TP + FP)
Predicted class: Negative		**FN:** False Negative (Type II Error)	**TN:** True Negative	**Negative Predictive Value:** TN / (TN+FN)
		Recall or Sensitivity: TP / (TP + FN)	**Specificity:** TN / (TN + FP)	**Accuracy:** TP + TN / (TP + TN + FP + FN)

Figura 1: Esempio di matrice di confusione per classificazione binaria.

È importante notare che metriche come sensitività, precisione e specificità dipendono dalla classe presa in considerazione, mentre l'accuratezza è un indice globale.

3.5 Area Under the Curve (AUC) e curva ROC

L'AUC (Area Under the Curve) è una misura basata sulla curva ROC (Receiver Operating Characteristics).

Definizione 3.8: Curva ROC

Una curva ROC è un grafico che mostra le performance di un classificatore al variare di un suo parametro (es. una soglia). Mette in relazione il **True Positive Rate (TPR)** (sull'asse Y) con il **False Positive Rate (FPR)** (sull'asse X).
($FPR = 1 - Specificità = \frac{FP}{FP+TN}$).
Le curve ROC passano sempre per i punti $(0,0)$ e $(1,1)$. Esistono inoltre due condizioni limite che rappresentano due curve di riferimento:

- Una retta che taglia il grafico a 45 gradi passando per l'origine. Questa retta rappresenta il caso del **classificatore casuale** e l'area sottesa (AUC) è pari a 0.5.
- Una curva rappresentata dal segmento che dall'origine sale verticalmente al punto $(0,1)$ e dal segmento che congiunge il punto $(0,1)$ a $(1,1)$. Questa curva ha un'area sottesa di valore pari a 1 e rappresenta il **classificatore perfetto**.

L'AUC, ha un valore compreso tra 0 e 1, e misura l'intera area bidimensionale sotto la curva ROC.

- **AUC = 1**: Rappresenta il classificatore perfetto, che passa per il punto (0,1)
- **AUC = 0.5**: Rappresenta il classificatore casuale (la linea diagonale).
- **AUC = 0**: Rappresenta il classificatore "perfettamente sbagliato" (che inverte tutte le predizioni).

Il valore di AUC (tra 0 e 1) può essere interpretato come la probabilità che il classificatore assegni un punteggio più alto a un individuo positivo scelto a caso, rispetto a un individuo negativo scelto a caso.

3.6 Classificazione multi-classe

Le misure viste finora (Precision, Recall, F_1-score) sono definite per la classificazione binaria. Per applicarle a problemi con $K > 2$ classi, si perde la visione di performance globale. Per l'F_1-score, si possono calcolare delle medie. L'approccio comune è "one-vs-rest": per ogni classe $g_i \in G = \{1, \ldots, K\}$, si costruisce una matrice di confusione dove g_i è il "caso positivo" e tutte le altre classi formano il "caso negativo". Si calcolano così TP_i, FP_i e FN_i per ogni classe i.

3.6.1 Micro-average

La micro-average (micro-media) aggrega i contributi di tutte le classi "sull'unità più piccola" (i singoli campioni) prima di calcolare le metriche. Queste metriche sono:

$$P_{micro} = \frac{\sum_{i=1}^{|G|} TP_i}{\sum_{i=1}^{|G|} (TP_i + FP_i)}$$

$$R_{micro} = \frac{\sum_{i=1}^{|G|} TP_i}{\sum_{i=1}^{|G|} (TP_i + FN_i)}$$

Da cui si può derivare il F_1-score micro-averaged, $F_{1_{micro}}$ che rappresenta la media armonica di P_{micro} e R_{micro}.

$$F_{1_{micro}} = 2 \cdot \frac{P_{micro} \cdot R_{micro}}{P_{micro} + R_{micro}}$$

Nota 3.4: Micro-average e classi sbilanciate

Questa misura non è sensibile alle prestazioni sulle singole classi e può essere fuorviante se la distribuzione delle classi è sbilanciata.

3.6.2 Macro-average

La macro-average (macro-media) calcola la media su gruppi più vasti.

$$P_{macro} = \frac{\sum_{i=1}^{|G|} P_i}{|G|}$$

$$R_{macro} = \frac{\sum_{i=1}^{|G|} R_i}{|G|}$$

Da cui si può derivare il F_1-score macro-averaged, $F_{1_{macro}}$ che rappresenta la media armonica di P_{macro} e R_{macro}.

$$F_{1_{macro}} = 2 \cdot \frac{P_{macro} \cdot R_{macro}}{P_{macro} + R_{macro}}$$

Nota 3.5: Macro-average per dati sbilanciati

Se questo valore è grande, indica che il classificatore funziona bene (in media) per ogni singola classe. Per questo motivo è più adatto per dati con distribuzione sbilanciata.

3.6.3 Generalizzazione di AUC (Metodo Hand & Till)

Esiste anche una generalizzazione dell'AUC per $k > 2$ classi (Metodo Hand & Till, 2001). L'idea è calcolare una misura di separabilità $\hat{A}(i|j)$ per ogni possibile coppia di classi (i, j).

Definizione 3.9: Generalizzazione AUC

Sia $\hat{A}(i|j)$ la probabilità che dato un elemento a caso della classe j abbia probabilità inferiore di attribuire quell'elemento alla classe i, rispetto al valore di probabilità che attribuirebbe ad un elemento a caso della classe i. È possibile calcolare $\hat{A}(i|j)$ utilizzando le seguetnti definizioni:

- $\hat{p}(X_l)$ è la probabilità stimata che l'osservazione l sia originata dalla classe i.
- per tutte le osservazioni x_l della classe i, sia $f_l = \hat{p}(X_l)$. la probabilità stimata di appertenere alla classe i.
- per tutte le osservazioni x_k della classe j, sia $g_k = \hat{p}(X_k)$. la probabilità stimata di appertenere alla classe i.

Allora i valori ottenuti ordinati in modo crescente sono: $\{g_1, \ldots, g_n, f_1, \ldots, f_n\}$. Sia r_l il rango della l-esima osservazione della classe i.

Il numero totale di coppie di punti in cui il punto della classe j ha un valore di probabilità stimato di appartenenza alla classe i inferiore a quello della classe i è:

$$\sum_{l=1}^{N_i}(r_l - l) = \sum_{l=1}^{N_i} r_l - \sum_{l=1}^{N_i} l = S_i - \frac{N_i(N_i+1)}{2}$$

Dove N_i e N_j sono il numero di osservazioni delle classi i e j e S_i è la somma dei ranghi delle osservazioni della classe i.
La probabilità che un punto scelto a caso della classe j abbia una probabilità stimata di appartenenza alla classe i inferiore a quella di un punto scelto a caso della classe i è quindi:

$$\hat{A}(i|j) = \frac{S_i - \frac{N_i(N_i+1)}{2}}{N_i \cdot N_j}$$

Inoltre considerando che non è possibile distinguere $\hat{A}(i|j)$ da $\hat{A}(j|i)$, si ha che la misura di separabilità tra le classi i e j è data dalla media tra $\hat{A}(i|j)$ e $\hat{A}(j|i)$, ovvero:

$$\hat{A}(i|j) = \frac{\hat{A}(i|j) + \hat{A}(j|i)}{2}$$

Il valore di AUC globale (M) di un classificatore multi-classe è quindi dato dalla media di tutti i valori $\hat{A}(i|j)$ calcolati è definito come:

$$M = \frac{2}{c(c-1)} \sum_{i<j} \hat{A}(i|j)$$

Dove c è il numero totale di classi, $\frac{2}{c(c-1)}$ è un fattore che viene applicato perchè sono presenti $c(c-1)$ modi differenti con cui costruire coppie distinte di classi.

3.7 Cross-validation

Definizione 3.10: Cross-Validazione

È una tecnica statistica usata per validare un modello e valutare come i suoi risultati si generalizzeranno a un insieme di dati indipendente. L'obiettivo primario è testare la capacità del modello di prevedere su nuovi dati, non usati durante l'addestramento. Serve principalmente a stimare problemi di **overfitting** o di **selection bias**.

Il *selection bias* si verifica quando la scelta del training set è viziata (da fattori esterni) e non rispecchia un campionamento uniforme dell'universo delle osservazioni.

Nota 3.6: Selection bias

l *selection bias* può portare a stime distorte delle performance del modello, poiché il training set non rappresenta adeguatamente la popolazione generale. È importante essere consapevoli di questo bias durante la fase di progettazione dello studio e nella raccolta dei dati.
Training set e test set dovrebbero essere prodotti tramite campionamento uniforme dell'universo delle possibili osservazioni.

La cross-validazione si divide in due tipi principali:

- Cross-validazione esaustiva, che testa tutte le possibili divisioni del dataset in TS e CS.
- Cross-validazione non esaustiva, che testa solo un sottoinsieme delle possibili divisioni.

3.7.1 Cross-validazione esaustiva

La cross-validazione esaustiva testa tutti i modi possibili di dividere il dataset in TS e CS.

- **Leave-p-out (LPO)**: Utilizza p osservazioni come CS e $N - p$ come TS. Questo processo viene ripetuto per tutti le $\binom{n}{p}$ possibili combinazioni.
- **Leave-one-out (LOOCV)**: È un caso particolare di LPO dove $p = 1$. È appropriata per dataset molto piccoli, dove il costo computazionale è secondario rispetto all'accuratezza della stima.

3.7.2 Cross-validazione non esaustiva

La cross-validazione non esaustiva testa solo un sottoinsieme delle possibili divisioni. La tecnica più comune è la **k-fold cross-validation**, dove il dataset viene diviso casualmente in k parti (fold) di eguale dimensione. A turno, ogni "fold" viene usato come Test Set (CS) e i restanti $k-1$ fold vengono usati come Training Set (TS). Il processo si ripete k volte e le metriche vengono mediate.