

# Apprendimento supervisionato: classificazione

## Algoritmi per l'intelligenza artificiale

Vincenzo Bonnici

Corso di Laurea Magistrale in Scienze Informatiche

Dipartimento di Scienze Matematiche, Fisiche e Informatiche

Università degli Studi di Parma

2025-2026

## Un esempio quotidiano di classificazione

Per facilitare la riflessione illustreremo di seguito un esempio "giocattolo": la discriminazione tra esseri umani di sesso maschile e di sesso femminile.

Consideriamo questo problema in un contesto in cui esso è banale: una società nudista. In tale contesto il problema è risolubile con la diretta osservazione delle “**caratteristiche** sessuali primarie” (cioè presenza di specifici organi). In inglese la parola “caratteristica” si traduce in questo contesto con il termine **feature**.

Una feature è un aspetto **direttamente osservabile** relativo ad un fenomeno per il quale si può registrare una precisa **misura quantitativa** (un valore numerico intero o double) o **categoriale** (vero/falso, dolce/salato, festivo/feriale, rosso/verde/giallo eccetera).

Le feature che i cittadini della società nudista terrebbero in considerazione sarebbero: presenza/assenza di seni sviluppati, presenza/assenza degli organi maschili, presenza/assenza di peluria. Essi potrebbero anche considerare altezza, peso, proporzioni tra le varie parti del corpo eccetera.

Ma perché i nudisti dovrebbero complicarsi la vita con **feature complesse** da osservare e misurare se il problema di decidere il genere (la **classe**) di un altro individuo può essere risolto in base ad alcune feature più semplici?

A partire dalla osservazione di varie feature il classificatore giunge alla decisione di **etichettare** (in inglese: “to **label**”) il dato sotto osservazione in una categoria più astratta e generale.

Nel nostro esempio giocattolo le etichette (o classi o popolazioni o gruppi) sono il “genere maschile” e il “genere femminile”.

La **classe** di un oggetto è quindi un concetto astratto e generale che “spiega” le osservazioni. In un certo senso l’assegnazione ad una classe costituisce la **sintesi** delle osservazioni.

Complichiamo il problema e consideriamo il caso realistico della attuale società occidentale.

Noi siamo quasi sempre in grado di decidere il genere ("classe") di un passante ("dato osservato") senza necessariamente chiedergli di denudarsi di fronte a noi.

Le feature a nostra disposizione sono: altezza, massa apparente, capelli, abiti, peluria, stile nel movimento, timbro vocale eccetera.

In base all'osservazione di ciò che è manifesto e palese siamo in grado (anche in questo caso in maniera praticamente inconscia) di **riconoscere** il "genere", cioè la etichetta con cui etichettare l'altro individuo.

Tuttavia, potremmo avere difficoltà nel vedere passare un uomo con i capelli molto lunghi e una donna con la testa rasata... questi esempi (più in là impareremo che si chiamano tecnicamente "**outliers** statistici") mettono in crisi l'algoritmo con cui noi discriminiamo il genere.

Sia data una **collezione di dati**. Essa sarà definita come insieme  $P$  di  $M$ -uple del tipo :

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

Ciascuna delle **feature**  $x_{ji}$  appartiene ad un possibile **dominio di valori**  $D_j$ .  
I domini dei valori sono insiemi numerici oppure insiemi di "categorie".

L'insieme  $P$  sia partizionato in  $k$ -classi le cui etichette compongano l'insieme  $L = (A_1, \dots, A_k)$ .

Ciascun dato appartiene ad una ed una sola classe: per ciascun dato esiste dunque una ed una sola etichetta.

Un **algoritmo di classificazione** è una funzione computabile  $f : P \mapsto L$ :

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

tale funzione assegna ad ogni dato  $m$  una etichetta  $A_i$  scelta tra quelle presenti in  $L$  cercando di **stimare** l'etichetta **reale** del dato stesso.

## Classificazione: definizione

Lo schema di classificazione ha **successo** (inglese: hit) su un dato se  $f(m)$  coincide con l'etichetta della classe cui  $m$  realmente appartiene.

Se la funzione  $f(m)$  assegna l'etichetta errata si ha un **fallimento** (inglese: miss) dello schema.

In generale è impossibile perfezionare schemi di classificazione **error free** ed è importante per ogni schema proposto fornire stime (ottenute matematicamente o mediante una sufficiente sperimentazione) sul **tasso percentuale di hit/miss** che lo schema può statisticamente ottenere.

Il **livello tollerabile di errori** per un classificatore dipende dalla criticità della applicazione cui esso viene applicato. Per le applicazioni industriali si richiede in genere un tasso di errori inferiori al 5%, Per le applicazioni mediche un classificatore con tassi di errore superiori al 0.5% è da considerarsi inaccettabile. Tali soglie sono da considerarsi indicative e variano per ogni specifico problema adottato.

## Esempi di problemi di classificazione: salmoni e sea bass



Lunghezza e colore sono gli attributi, classificare significa trovare la specie o “l’etichetta” (label) per ciascun nuovo pesce osservato.

Una possibile presentazione dei dati in questo caso sarebbe una tabella che per ciascun pesce riporta: peso in grammi, lunghezza, colore dominante (un attributo o “feature” qualitativa da selezionarsi tra un insieme finito predeterminato tipo [BLU, AZZURRO, GRIGIO, VERDE]), eccetera.

La tabella ha una ultima colonna “specie di pesce” che il classificatore dovrà cercare di riempire a partire dai dati che sono forniti in modo da non confondere troppi salmoni (pesci pregiati) con “sea bass” pesci meno pregiati.

Si raccolgono per ciascuno studente i dati anagrafici, il censo della sua famiglia e i voti conseguiti negli esami dei corsi universitari. Queste informazioni sono le “osservazioni” o feature osservate.

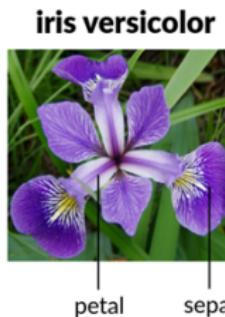
Si vuole “predire” (a volte classificare è una forma di predizione) in quale delle classi si troverà lo studente dieci anni dopo la Laurea: reddito basso, reddito medio, reddito alto.

Predizioni di questo tipo non hanno generalmente un grande valore sul singolo “studente” che ha una alta probabilità di essere erroneamente classificato dato l’elevato numero di fattori che intervengono e la impossibilità di tenere conto di tutti i fattori in un modello matematico che rimanga maneggevole. Esse sono però assai importanti per predire il futuro “statistico” (cioè in “aggregato”) della intera popolazione.

## Esempi di problemi di classificazione: Iris di Fisher

Nel 1916 il naturalista e matematico Fisher raccolse 150 iris. 50 di essi di specie "setosa", 50 di specie "virginica" e 50 di specie "speciosa". Egli era in grado di classificare ciascun iris nella sua corretta specie. In questo egli si aiutò certamente anche dalla osservazione della pianta da cui il fiore era stato spiccato.

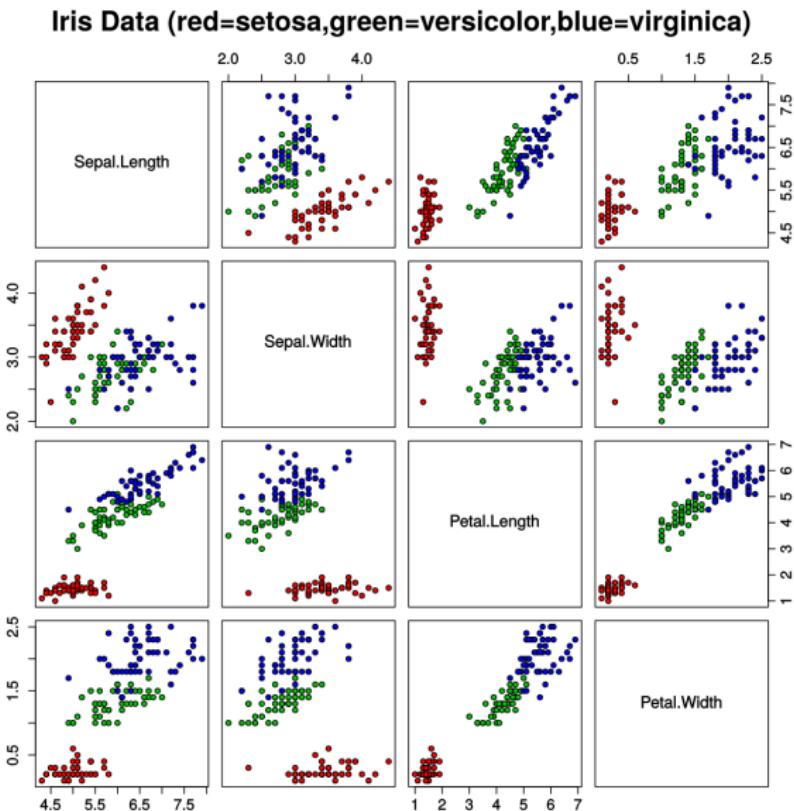
Per ciascuno dei 150 fiori Fisher rilevò le seguenti 4 misure (feature quantitative): lunghezza petalo, larghezza petalo, lunghezza sepalo, larghezza sepalo ed ovviamente annotò la osservazione qualitativa della specie.



Il problema che Fisher intendeva studiare e verificare era:

- è possibile assegnare correttamente la specie conoscendo solo le 4 misure del fiore senza sapere nulla altro sulla pianta da cui il fiore era stato colto?
- erano necessarie tutte e 4 le misure?
- le 4 misure erano “indipendenti” tra loro (grande petalo implica spesso grande sepalo)?

# Esempi di problemi di classificazione: Iris di Fisher



Queste domande originarono numerose ricerche e Fisher usò il suo data set per fornire evidenze che alcune sue idee matematiche erano corrette.

La dimensione del data set (150 record) è piccola e maneggevole (lo era anche al tempo dei calcolatori tascabili e del calcolo manuale!).

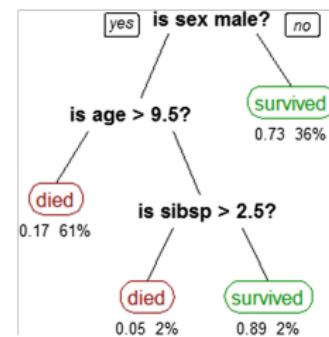
Questo data set è spesso inadeguato per verificare la validità degli schemi più complessi di classificazione automatica ma è di notevole valore come strumento di comprensione nella didattica.

# Esempi di problemi di classificazione: passeggeri del Titanic

E' stato compilata una tabella di 1046 record ciascuno relativo ad un passeggero del tragico ultimo viaggio del Titanic.

Per ogni record si registra

- classe di imbarco (1, 2 o 3)
- sesso (m=0, f=1)
- età in anni
- sopravvivenza (morto=0, salvato=1) : da predirre



## Cosa è feature e cosa è classe?

In generale la **classe** è vista come la **causa** delle feature.

Se un fiore è un iris setosa allora le misure dei sepali e dei petali rispettano certe proporzioni e misure.

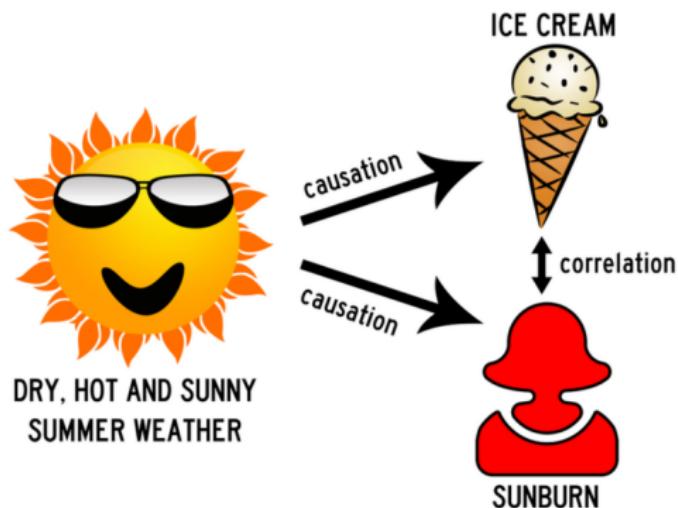
In questo senso classificare significa **cercare evidenza di cause nascoste** per fenomeni osservati. La “classe” di un dato però non è altro che l’attributo che si assegna a quel record.

Da un punto di vista teorico si potrebbe dire che ogni “feature” non nota costituisce una “classe”.

Per esempio nel caso degli iris di Fisher se si ignorasse la misura del petalo è possibile usare le altre 4 informazioni (larghezza petalo, larghezza e lunghezza sepalo e specie) per predire la lunghezza del petalo?

Nella realtà le situazioni sono spesso assai più sfumate di così, soprattutto per fenomeni complessi dove le relazioni di causa-effetto non sono sempre ben chiare. Per tali ragioni è utile studiare come un attributo varia rispetto agli altri.

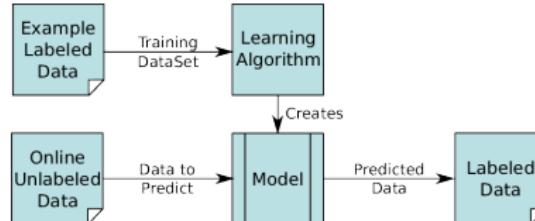
# Correlation does not imply causation



La classificazione basata sulla **osservazione di dati** è un compito ideale da affidare ad un computer.

Esso può esaminare enormi quantità di **record e dati** ed estrarne **regolarità** e particolari che guidano il processo di classificazione.

Gli algoritmi e le tecniche di classificazione automatica sono numerosissimi. Tutti i metodi noti però usano uno **schema generale** ben testato e riconosciuto dalla comunità scientifica.



# Apprendere le regole di classificazione dagli esempi

Sebbene gli studiosi delle scienze cognitive e i psicologi della età evolutiva siano ancora abbastanza incerti sui meccanismi che regolano la "maturazione" dei giovani umani esistono pochi dubbi sul fatto che un grande numero delle procedure di decisione/classificazione che utilizziamo nella nostra vita e sopravvivenza quotidiana dipendano da una fase di **apprendimento** (in inglese **training**) che si realizza nella massima parte alla fase iniziale della nostra vita.

Anche altri animali mediamente "intelligenti" condividono questo meccanismo di **apprendere dagli esempi**.





Un topino da laboratorio può essere coinvolto in alcune situazioni “esempio” e in base a questa fase di apprendimento regolerà il proprio comportamento per il resto della vita.

Se avvicinarsi ad una mattonella rossa comporterà per il topino una sensazione sgradevole esso apprenderà la regola “rosso implica dolore” e tale regola gli permetterà di classificare le cellette di un labirinto come “sicure” o “pericolose”.

E' quindi naturale per la classificazione automatica simulare tale fenomeno di apprendimento o, in inglese, **training**.

Dalla osservazione di un insieme ben classificato di record si tenta di **dedurre regole** applicabili a record **non ancora classificati** o che si presenteranno solo nel futuro.

In termini tecnici si parla di **universo delle osservazioni** per definire l'insieme complessivo dei record relativi ad un fenomeno in esame.

L'universo includerà sia i record già osservati e classificati, sia i record osservati e da classificare e anche i record che saranno osservati in futuro.

# Apprendere le regole di classificazione dagli esempi

Molti algoritmi iniziano esaminando un **sottoinsieme già classificato** dell'universo delle osservazioni.

Tale insieme di “allenamento” (in inglese **Training Set**, brevemente **TS**) è il deposito di informazioni iniziali da cui ricavare le “regole” di classificazione.



Le regole saranno di vario tipo (a seconda degli algoritmi e degli approcci adottati per trattare lo specifico problema di classificazione): statistiche, probabilistiche, fuzzy, funzioni discriminanti eccetera.

Un buon insieme di regole deve avere alcune importanti proprietà:

- **semplicità**: non deve essere troppo grande né troppo complicato (la classificazione dei record che si presenteranno nel futuro potrebbe essere soggetta a vincoli di efficienza o di **costo computazionale** per record classificato).
- **correttezza sul TS**: deve essere statisticamente sufficientemente corretto quando viene applicato al medesimo TS che ha portato alla sua creazione. **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.
- **generalizzabilità**: deve essere statisticamente sufficientemente corretto quando viene applicato al resto dei record dell'universo delle osservazioni. L'insieme delle regole deve quindi essere valido oltre che nella situazione già vista nel TS anche nel caso generale. Questa proprietà è nota come proprietà di “generalizzazione” dell'insieme di regole.

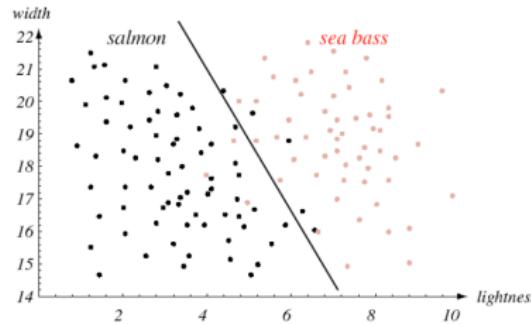
# Il problema dell'overfitting

Correttezza e generalizzabilità sono spesso in conflitto tra loro.  
Questo apparente paradosso è noto come **overfitting**.

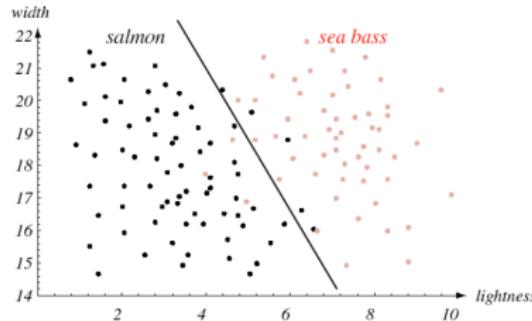
Si supponga che si misurino sia le dimensioni che il tono medio di grigio di un pesce che si vuole classificare nelle classi "Salmone" e "Sea bass".

Un pescatore esperto **ha già etichettato** un certo numero di pesci fornendoci un Training Set.

In un grafico bidimensionale rappresentiamo ciascun pesce con un punto all'incrocio tra il proprio valore di luminosità e il proprio valore di lunghezza.  
La distribuzione ipotetica che potrebbe venire fuori è quella che segue:



# Il problema dell'overfitting



Sul diagramma è stata tracciata una retta di separazione tra la “zona” dei salmoni e quella degli altri pesci.

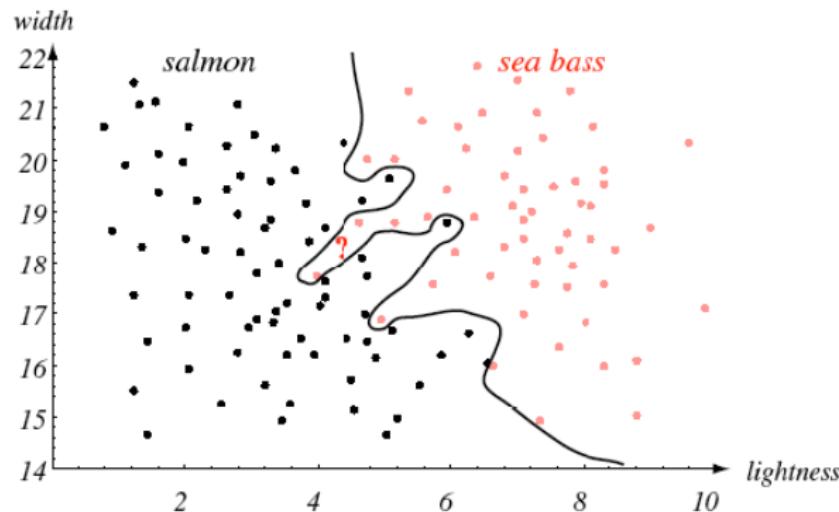
La retta è un esempio di semplice regola di classificazione. Essa è molto semplice e immediata.

E' corretta? Lo è abbastanza ma lascia alcuni errori. Alcuni preziosi salmoni (punti) neri sono lasciati nella zona dei sea bass. Alcuni sea bass sono invece trattati come salmoni.

# Il problema dell'overfitting

E' possibile trovare una regola di classificazione migliore? Sì, certamente, se si tralascia la richiesta di semplicità.

Una "frontiera" di separazione potrebbe essere quella nel diagramma che segue:



## Il problema dell'overfitting

Questa nuova “regola” è certo più corretta ma: è preferibile a quella precedente?

La risposta è no per due ragioni.

Una ovvia è la complicazione della regola.

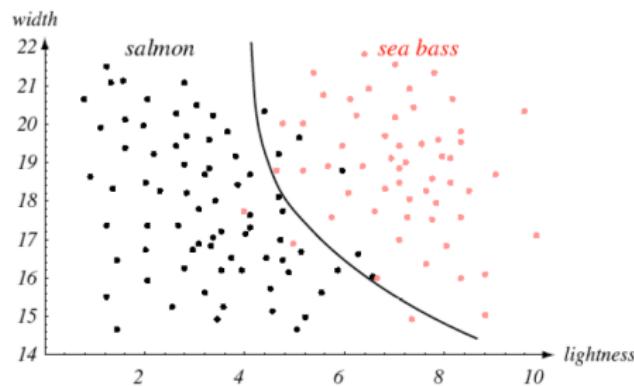
L'altra meno ovvia ma assai più importante è la seguente: la complessa frontiera nasce per “adattarsi” ai dati del training set. Tale insieme **campione** (in inglese **sample**) è **piccolo** rispetto all'universo delle informazioni ed è assolutamente **casuale**.

Tutto lascia presagire che punti neri (salmoni) potranno apparire nelle “penisole” della zona sea-bass e che sea bass appariranno nelle penisole della regione “salmone”. La nuova regola quindi spiega benissimo (**trop** **bene!**) il TS ma **non si generalizza** affatto al resto dei pesci del mare!

# Il problema dell'overfitting

Non si deve però pensare che non ci siano sistemi di classificazione intermedi tra il primo e il secondo riportati sopra.

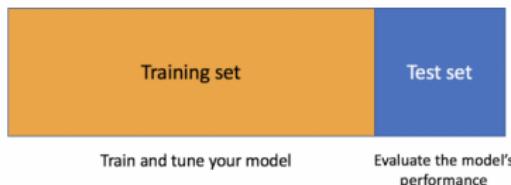
Per esempio se si accetta come linea di separazione una conica (iperbole) si complica la regola (da una equazione di primo grado ad una di secondo grado) ma la regola commette meno errori e fa meglio sperare sulle proprie proprietà di generalizzazione



# Validazione

Come “convalidare” le proprietà di “generalizzazione” di un insieme di regole? Si tratta di un problema complesso.

Uno dei metodi più usati è quello di avere a disposizione oltre al TS un altro insieme di record già etichettati detto **insieme di controllo**, anche detto di **verifica (Control Set o Test Set**, brevemente **CS**).



Il CS non viene utilizzato al momento della sintesi delle regole ma solo dopo che esse sono state ben definite a partire dal TS.  
Se le regole danno sul CS lo stesso tasso di errore che esse hanno sul TS si tratta di regole generalizzabili.

Esistono molte strategie e variazioni su questo tema

# Quali e quante feature?

Spesso rileviamo per i record molte feature.

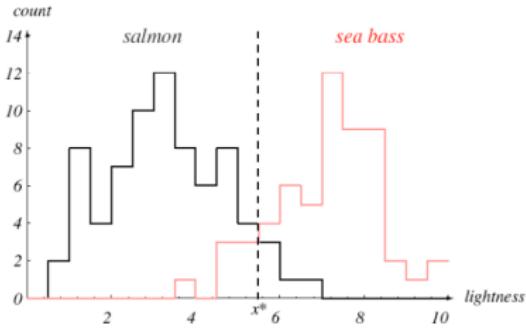
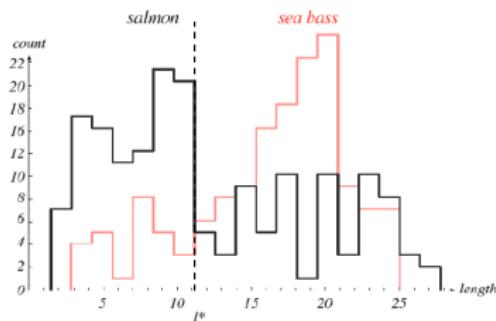
Sono tutte utili o alcune sono sovrabbondanti o addirittura dannose?

E se esse sono utili ce ne sono alcune più importanti delle altre?

“Combinare” più di una feature come abbiamo fatto sopra con peso e luminanza dei pesci è una strategia sempre conveniente?

Non esiste una risposta che vada sempre bene per tutti i casi.

Nel caso dell'esempio ecco i grafici (istogrammi) dei salmoni e dei sea bass quando le caratteristiche luminanza e lunghezza vengono esaminate in maniera disgiunta:



Gli algoritmi e gli schemi di classificazione che intendiamo usare debbono essere applicati al mondo reale. Ciò comporta una sfida particolarmente difficile perché nella realtà non si ha quasi mai il completo controllo di tutti i fattori sperimentali e di osservazione.

La ricerca delle regole di classificazione è quindi resa più complicata dalla presenza di **rumore** nei dati.

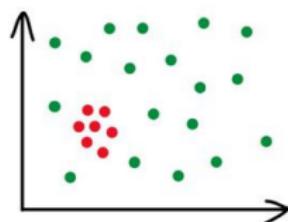
Si indica una serie di **perturbazioni** dei dati ad opera di fenomeni non controllabili o non noti.

Un “salmone **ideale**” sarebbe sempre al di qua della frontiera di separazione della sua regione nel diagramma visto sopra.

Eppure a volte un salmone si situerà in piena regione sea-bass.  
Con una radicale semplificazione si potrebbero classificare le **cause** di tale "spostamento" in due grandi tipi:

- **endogene** al fenomeno "salmone": il pesce in questione ha avuto una dieta, o una storia personale, molto diversa dal tipico salmone e come risultato finisce con l'assomigliare di più ad un sea-bass. Le cause che fanno deviare il pesce dalla sua "salmonitudine" possono essere varie, non note e impredicibili.
- **esogene** al fenomeno "salmone" ma dovute all'osservatore: forse la macchina fotografica che misura la dimensione e la luminosità della pelle del salmone si è starata (errore sistematico dello strumento), forse il vecchio pescatore che abbiamo impiegato per "etichettare" ciascun pesce nella sua specie si è distratto per una lunga fumata di pipa, eccetera.

A volta il dato è così alterato da confondere e rendere inefficace la fase di apprendimento di un algoritmo quando viene inserito nel TS. Dati molto "fuori norma" rispetto ai valori tipici di una classe vengono definiti dagli statistici con il termine tecnico di **outlier** (letteralmente: "fuori livello").



Alcuni algoritmi hanno fortunatamente una struttura che riduce molto l'influenza di tali eccezioni, altri invece sono ad esse molto sensibili.

# Il rumore e le eccezioni

Il rumore è un ostacolo serio ma inevitabile al processo di classificazione: ogni algoritmo che intende processare dati reali deve avere una qualche forma di “protezione” o resistenza alle deviazioni che il rumore impone al processo decisionale. L'algoritmo deve quindi essere **robusto alle perturbazioni**.

Si osservi infine che il fenomeno “rumore”, cioè l'inserimento di informazioni non affidabili o fuorvianti nel processo di classificazione viene amplificato se vengono prese in considerazione **troppe feature**, soprattutto se alcune di esse sono irrilevanti.



Caco



CacoMela



Mela

Contare gli errori nella messa in opera di un classificatore una fase essenziale è la determinazione della sua qualità in termini di “errori” commessi nella classificazione.

Una prima grezza stima che si può costruire se si ha disposizione un CS consiste nel riportare quanti record in percentuale non vengono correttamente assegnati alla propria classe.

Questo singolo parametro pur assai rilevante non è però completamente descrittivo delle situazioni che si possono verificare.

**Non tutti gli errori sono eguali!** Ne' essi hanno tutto il medesimo costo!

Per esempio trattare un salmone come sea bass comporta la perdita economica del denaro proveniente dalla vendita del pesce come pregiato. Mentre se un sea bass finisce in una scatoletta di salmone il cliente potrebbe anche chiudere un occhio (e ha ormai pagato la scatoletta quando la apre!).

Tecnicamente si dice che i costi degli errori non sono **uniformi** o non sono simmetrici.

Altri esempi tipici provengono dai casi di diagnosi mediche. Classificare un sano come malato è un errore che però un secondo controllo scopre. L'unico effetto che tale errore provoca, nella maggior parte dei casi, è solo una piccola immotivata paura nel paziente. Trattare da sano un malato ritardando la diagnosi che potrebbe salvargli la vita è invece un errore gravissimo.

Per tali ragione uno strumento molto diffuso nella determinazione di qualità di un classificatore è la cosiddetta **matrice di confusione** cioè una griglia quadrata. Essa riporta (in percentuale) sulle righe quanti elementi della classe corrispondente a quella riga sono stati assegnati alle varie classi presenti nella popolazione.

		Real	
			
Predicted		1	2
		0	7

# Contare gli errori

Un classificatore “perfetto” ha come matrice di confusione la matrice identica. In generale un buon classificatore non dovrebbe avere percentuali eccessive negli elementi fuori dalla diagonale principale.



Si osservi inoltre come in genere non sia sufficiente stimare l'errore su un unico CS. Esso è pur sempre un campione casuale dell'intero universo delle osservazioni. Ripetere i test e i controlli con diversi CS garantisce una migliore precisione nella stima degli errori.

Purtroppo ottenere dei buoni CS (e TS) è spesso costoso o impossibile. Ciò ha portato alla introduzione di "strategie" di randomizzazione con ripetizione nella selezione del TS e del CS dall'universo delle osservazioni.

- ① **sensing** (o **sampling**): raccolta dei dati dal mondo fisico e loro traduzione in informazioni digitali
- ② **segmentazione**: partizione dei dati in unità significative, eliminazione di particolari irrilevanti, miglioramento della qualità dei dati, isolamento delle informazioni che generano un “data item”

- ③ **estrazione delle feature:** a partire dai dati segmentati ottenere misure (quantitative o qualitative) per ciascuna caratteristica. Le misure popoleranno le colonne della tabella dei dati. Le feature sono soggette a una grande variabilità e misurarle con precisione non è sempre possibile. E' una buona idea scegliere feature che siano **INVARIANTI alle trasformazioni** tipiche della situazione sperimentale in esame.

Per esempio se i pesci vengono fotografati nella luce naturale mentre sono trasportati su un nastro, la feature luminanza risulterà non particolarmente comoda: essa non è invariante alle condizioni di illuminazione e il classificatore darebbe esiti diversi se usato la sera, la mattina, in giornate di sole o in giornate coperte. Il peso invece è una proprietà invariante alle condizioni di illuminazione e se la unica variabilità proviene dalle diverse luci essa sarà molto più affidabile come feature per la classificazione finale. Infine deve esserci almeno una generale probabile relazione tra classi e feature misurate: classificare la carriera di uno studente discriminando sulla feature "colore degli occhi" non è una grande idea.

- ④ **classificazione**: esecuzione dell'algoritmo di assegnazione delle label/classi.
- ⑤ **post-processing**: valutazione della qualità della classificazione, dei costi connessi con l'errore di classificazione.
- ⑥ decisione utilizzo del classificatore per la risoluzione di un problema reale

Un “designer” di un sistema di classificazione opera tipicamente in cicli. Un prototipo di classificatore viene creato, valutato, ridisegnato eccetera. Ciascun ciclo richiede attenzione ai seguenti aspetti (in ordine):

- ① **Raccolta dati.** In questa fase le attività importanti sono: selezione dei dati su cui lavorare per “allenare” il classificatore e conoscenza dei dati su cui il classificatore dovrà generalizzarsi.
- ② **Selezione delle feature** che saranno prese in considerazione dal nostro classificatore. Troppo poche? Troppe? Tutte rilevanti? Si possono ottenere nuove feature più significative combinando matematicamente quelle che si osservano più facilmente?
- ③ **Scelta del modello matematico** degli oggetti che si intende classificare. Si tratta di formulare ipotesi su come interagiscono le feature tra loro e con l'esito della classificazione e di comprendere come si distribuiscono (statisticamente) le misure delle feature stesse. Molto collegato con il punto precedente.

- ④ **training del classificatore**: usando un sottoinsieme di osservazioni ben comprese e ben classificate l'algoritmo di classificazione scelto può essere “accordato” in modo da offrire le risposte richieste e minimizzare gli errori (e i costi).
- ⑤ **valutazione**: abbiamo risolto il problema di classificazione o si poteva fare meglio con gli algoritmi e il software a disposizione? Solo un esperimento su un data set di valutazione può assicurarci a riguardo.