Data Mining e Bioimmagini

# Classificazione e istruzione del modello

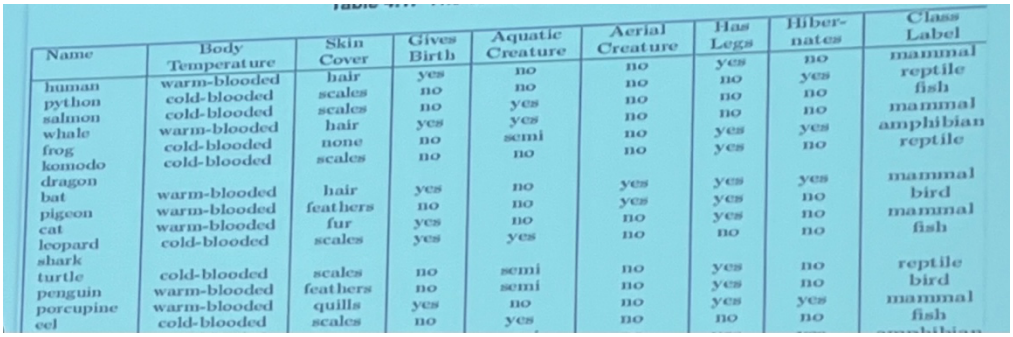
Prof. Pierangelo Veltri – 11/10/2023- Autori: Panarello, Saeedzadeh - Revisionatori: Panarello, Saeedzadeh

## Cosa significa fare classificazione?

Un modello di classificazione permette di distinguere tra oggetti appartenenti a classi differenti.
Data una raccolta di record (training set), ogni record è caratterizzato da una tupla (x,y), dove x è un set di attributi e y è la variabile di classe (class label).
X: attributo, predittore, variabile indipendente, input;
Y: classe, risposta, variabile dipendente, output.
Lo scopo è di attribuire ad ogni elemento del dataset la variabile di classe predefinita y in base al set di attributi x.

Esempio

Ad esempio, potrebbe essere utile ai biologi per classificare, all'interno di un dataset di vertebrati, le classi di appartenenza dei vari animali (mammifero, rettile, pesce, ecc.). Indichiamo con X l'insieme di attributi che descrivono i vari animali (Body Temperature, Skin Cover, ecc.) e con Y la o le caratteristiche specifiche che indicano la classe di appartenenza (Class Label). La Y è un attributo significativo che caratterizza il campione come appartenente ad una classe.



| Name | Body Temperature | Skin Cover | Gives Birth | Aquatic Creature | Aerial Creature | Has Legs | Hibernates | Class Label |
|---|---|---|---|---|---|---|---|---|
| human | warm-blooded | hair | yes | no | no | yes | no | mammal |
| python | cold-blooded | scales | no | no | no | no | yes | reptile |
| salmon | cold-blooded | scales | no | yes | no | no | no | fish |
| whale | warm-blooded | hair | yes | yes | no | no | no | mammal |
| frog | cold-blooded | none | no | semi | no | yes | yes | amphibian |
| komodo dragon | cold-blooded | scales | no | no | no | yes | no | reptile |
| bat | warm-blooded | hair | yes | no | yes | yes | yes | mammal |
| pigeon | warm-blooded | feathers | no | no | yes | yes | no | bird |
| cat | warm-blooded | fur | yes | no | no | yes | no | mammal |
| leopard shark | cold-blooded | scales | yes | yes | no | no | no | fish |
| turtle | cold-blooded | scales | no | semi | no | yes | no | reptile |
| penguin | warm-blooded | feathers | no | semi | no | yes | no | bird |
| porcupine | warm-blooded | quills | yes | no | no | yes | yes | mammal |
| eel | cold-blooded | scales | no | yes | no | no | no | fish |

Altro esempio

Supponiamo di avere una tupla di un record fatta nel seguente modo:

- X: [nome, nazionalità, gender, familiarità, problemi cardiologici pregressi, valoriColesterolo, stile di vita (sv)]
- Y= Rischio: moderato, alto, basso
- Sv: sedentaria, sportivo, normale

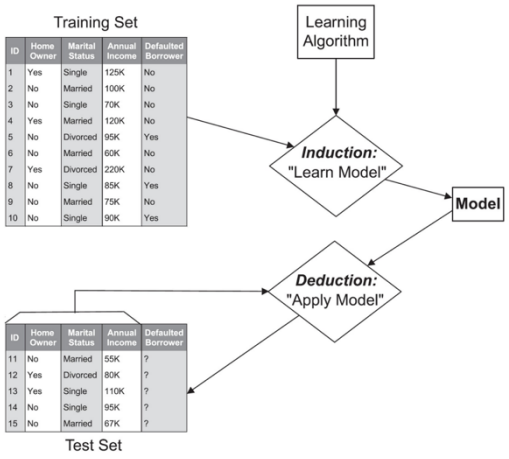
| Nome/ID | Nazionalità | Gender | Familiarità | Pb cardio | ColesLDL | sv | RISCHIO |
|---|---|---|---|---|---|---|---|
| 12 | Italiana | M | Si | No | 190 | normale | alto |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |

Supponiamo di riempire la nostra tabella con i dati dei pazienti; si ottengono dei record che permettono di classificare il rischio (per esempio d'infarto miocardico) di ogni paziente e classificarli in base al livello di tale rischio.
In tabella altri possibili task.

| Task | Attribute set, $x$ | Class label, $y$ |
|---|---|---|
| Categorizing email messages | Features extracted from email message header and content | spam or non-spam |
| Identifying tumor cells | Features extracted from x-rays or MRI scans | malignant or benign cells |
| Cataloging galaxies | Features extracted from telescope images | Elliptical, spiral, or irregular-shaped galaxies |

*L'insieme dei valori x sono le variabili utilizzate per definire gli attributi della classe. Per questo la x può essere considerata input, la y l'output. L'insieme delle x è un subset, la y il target.*

## General Approach for Building Classification Model



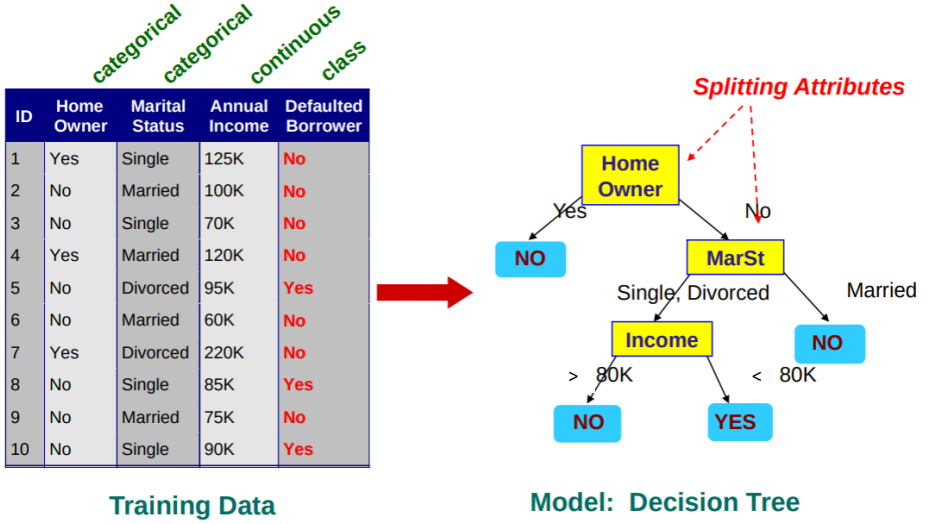

Per creare un modello di classificazione è necessario **istruire** l'algoritmo tramite un training set, cosicchè per induzione il modello diventa in grado di riconoscere le caratterstiche (X) che descrivono l'appartenenza ad una classe (Y); una volta istruito, tale modello può, per deduzione, andare a classificare anche i dati di cui non è indicata la classe.

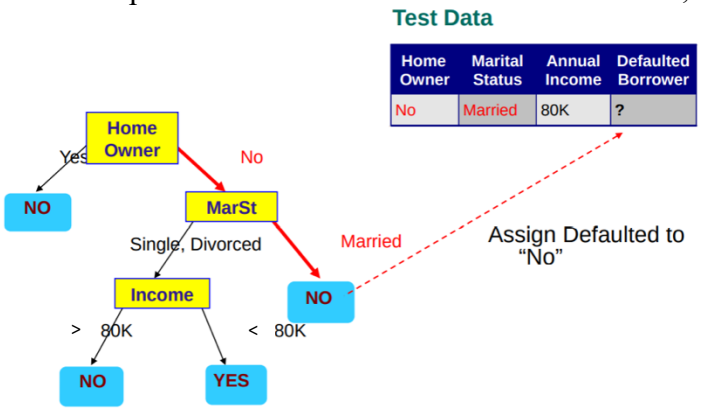
**Esempio Defaulted Borrower**

In questo esempio si hanno dati relativi allo status socioeconomico di alcuni clienti di una banca per comprendere se potenzialmente sono dei mutuatari inadempienti (quindi sarebbe meglio negargli il prestito). In particolare, come X abbiamo:

- Possessore di una casa
- Stato civile
- Reddito annuo

La nostra Y (class label) è defaulted borrower (booleano, si/no). Inizialmente, il modello viene istruito tramite un training set (riferito a clienti precedenti) e sulla base di tali dati "impara" come prendere la sua decisione di classificazione; tale scelta è schematicamente rappresentata da un **albero decisionale** che permette, in base ad una serie di caratteristiche, di prendere una scelta piuttosto che un'altra. Nell'esempio in esame, se il cliente, ad esempio, possiede una casa viene classificato come "degno di fiducia" perché tutti i casi precedenti di clienti possessori di una casa questi si sono rilevati come affidabili. Questa operazione viene effettuata per tutti gli attributi cosicché avremo un albero decisionale completo, in grado di classificare in base alle varie caratteristiche X.




Un modello di questo tipo può essere utilizzato per due casi in particolare:

- **Classificazione basata sulla predizione:** L'algoritmo è in grado di "predire" l'appartenenza di un campione non ancora classificato alla classe di appartenenza. Per esempio, un nuovo

cliente della banca vuole un prestito, con l'algoritmo ormai istruito si può predire la sua affidabilità. PREVISIONE FUTURA

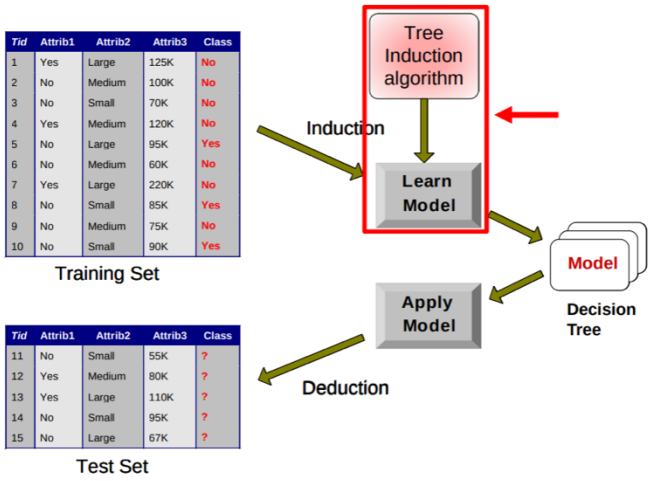
- **Classificazione di un dato missing (mancante):** L'algoritmo è in grado di "ritrovare" la classe di appartenenza di un elemento qualora tale informazione sia stata persa. Ad esempio, per qualche motivo si è persa l'informazione di Default Borrower di un dato cliente; tramite l'algoritmo si riattribuisce tale dato perso. CLASSIFICAZIONE PRESENTE/DESCRIZIONE

## ISTRUZIONE DELL'ALGORITMO

Esistono diverse tecniche di classificazione che vengono usate dall'algoritmo per imparare:

- **Classificatori di base:**
  - Metodi basati sull'albero decisionale
  - Metodi basati su regole
  - Nearest-neighbor (vicino più prossimo)
  - Naïve Bayes and Bayesian Belief Networks
  - Macchine a supporto vettoriale
  - Reti neurali (connessioni neurali profonde).
- **Classificatori d'insieme:**
  - Boosting, Bagging, Random Forests

In questa lezione abbiamo visto il metodo basato sull'albero decisionale.

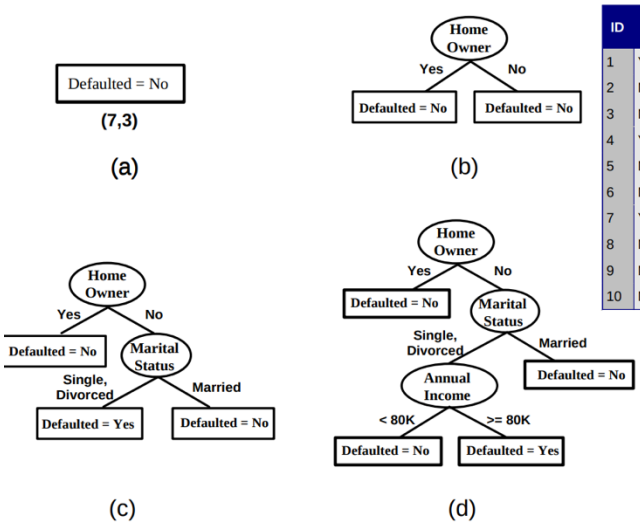



### Induzione tramite albero decisionale

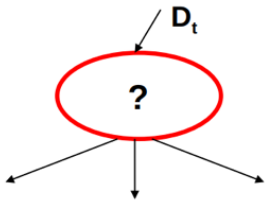
Esistono vari algoritmi in grado di classificare tramite le scelte effettuate all'interno di un albero decisionale:

- Hunt's Algorithm
- CART
- ID3, C4.5
- SLIQ, SPRINT

## ALGORITMO DI HUNT




| ID | Home Owner | Marital Status | Annual Income | Defaulted Borrower |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 95K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 75K | No |
| 10 | No | Single | 90K | Yes |

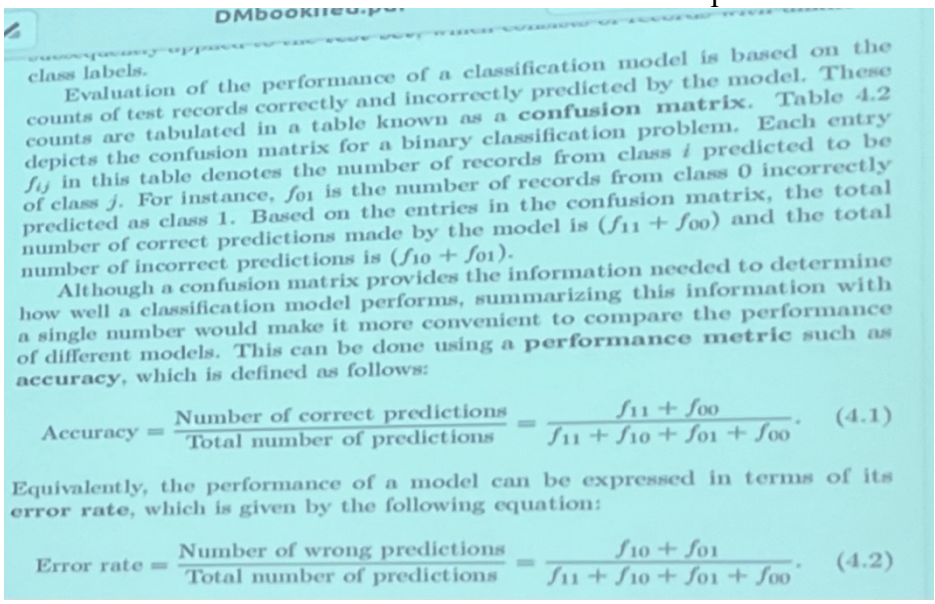



Indicato $D_t$ il set di training records (letteralmente una riga della tabella) che raggiunge un dato nodo t di un albero decisionale:
- Se $D_t$ contiene dei record (attributi) che appartengono esclusivamente alla classe $y_t$, allora t è un nodo foglia classificato come $y_t$.
- Se $D_t$ contiene dei record che appartengono a più di una classe, usa un attribute test (qui avviene la scelta) per dividere il dataset in subsets più piccoli. Applica ricorsivamente la procedura ad ogni subset.

***Misura della validità della classificazione***

La valutazione delle performance del modello di classificazione è basata sul numero di campioni correttamente classificati e non dal modello stesso. Fondamentali sono due parametri:

- *Accuracy:* numero delle predizioni corrette diviso il numero di totale di predizioni
- *Tasso di errore:* numero di predizioni errate diviso il numero totale di predizioni



class labels.

Evaluation of the performance of a classification model is based on the counts of test records correctly and incorrectly predicted by the model. These counts are tabulated in a table known as a **confusion matrix**. Table 4.2 depicts the confusion matrix for a binary classification problem. Each entry $f_{ij}$ in this table denotes the number of records from class $i$ predicted to be of class $j$. For instance, $f_{01}$ is the number of records from class 0 incorrectly predicted as class 1. Based on the entries in the confusion matrix, the total number of correct predictions made by the model is $(f_{11} + f_{00})$ and the total number of incorrect predictions is $(f_{10} + f_{01})$.

Although a confusion matrix provides the information needed to determine how well a classification model performs, summarizing this information with a single number would make it more convenient to compare the performance of different models. This can be done using a **performance metric** such as **accuracy**, which is defined as follows:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} = \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}}. \quad (4.1)$$

Equivalently, the performance of a model can be expressed in terms of its **error rate**, which is given by the following equation:

$$\text{Error rate} = \frac{\text{Number of wrong predictions}}{\text{Total number of predictions}} = \frac{f_{10} + f_{01}}{f_{11} + f_{10} + f_{01} + f_{00}}. \quad (4.2)$$

*Nota: è importante che il modello venga istruito con un training set vario e possibilmente ampio per fare in modo che "conosca" più casi possibili e riduca le possibilità di errore dovuto a mancanza di "conoscenze" (inteso come casi precedenti).*

**Human in the loop**: aggiungo informazioni nel dataset di training per arricchire il modello di nuove informazioni acquisite da una nuova predizione.
Ad esempio, chat GTP in base alle domande precedenti crea un canale indirizzato verso alcuni argomenti e cambia la propria risposta in base alla reazione dell'utente.

*Di fatto facciamo classificazione in ogni azione quotidiana che implica una scelta, percorrendo idealmente un albero decisionale.*