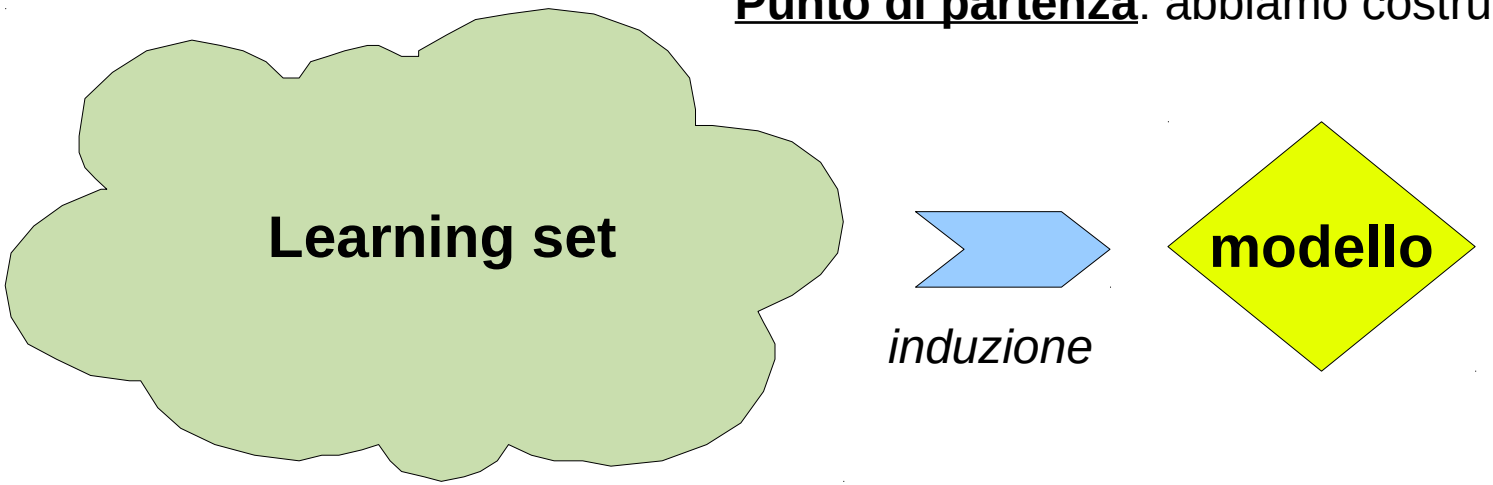
Valutazione dei classificatori appresi: bontà di generalizzazione

Per approfondimenti, di Ron Kohavi:

http://robotics.stanford.edu/%7Eronnyk/accEst.pdf
http://robotics.stanford.edu/%7Eronnyk/accEst-talk.ps

Problema: valutazione

Punto di partenza: abbiamo costruito un modello



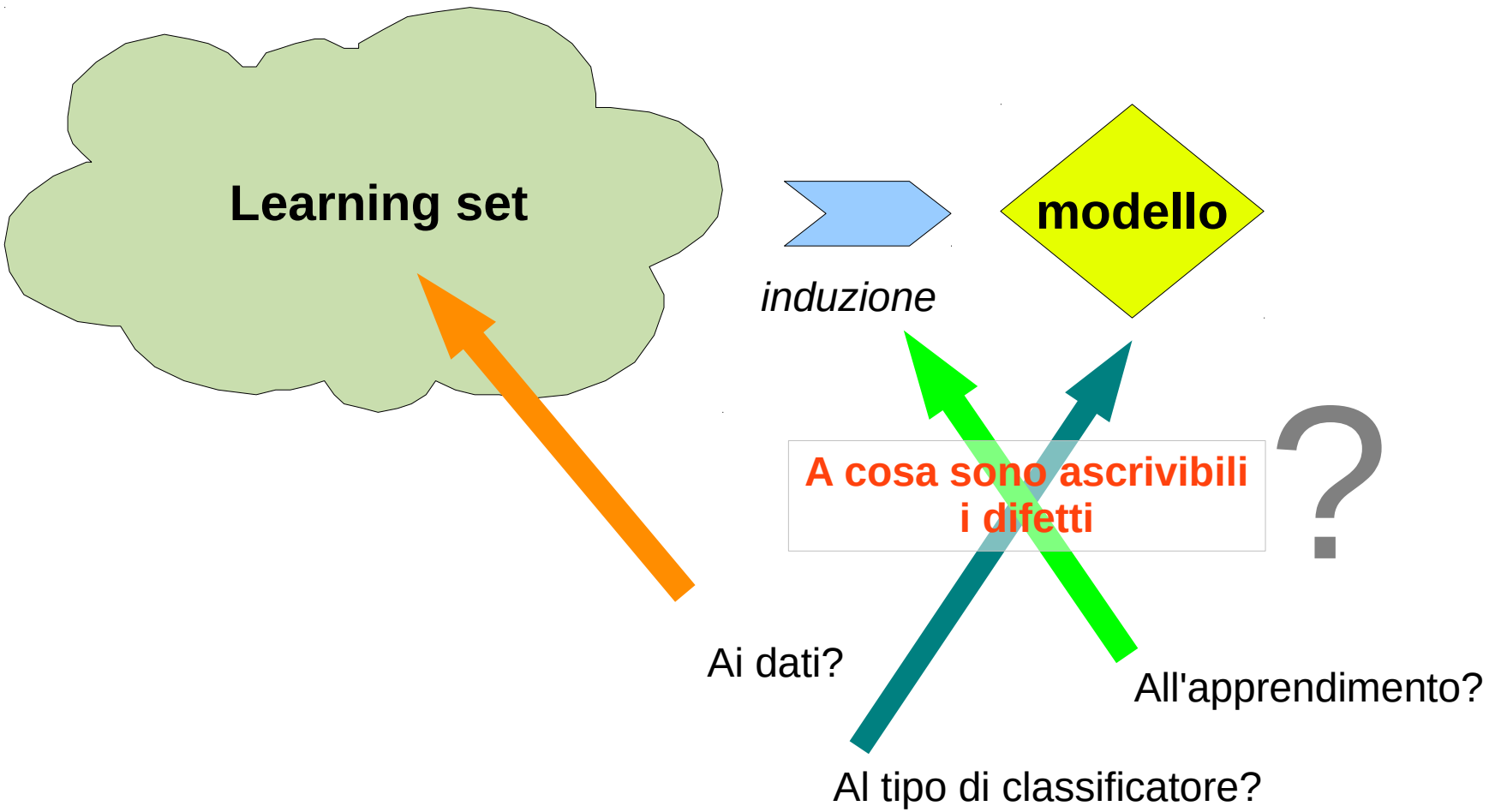
È un buon modello

Troppo generale?

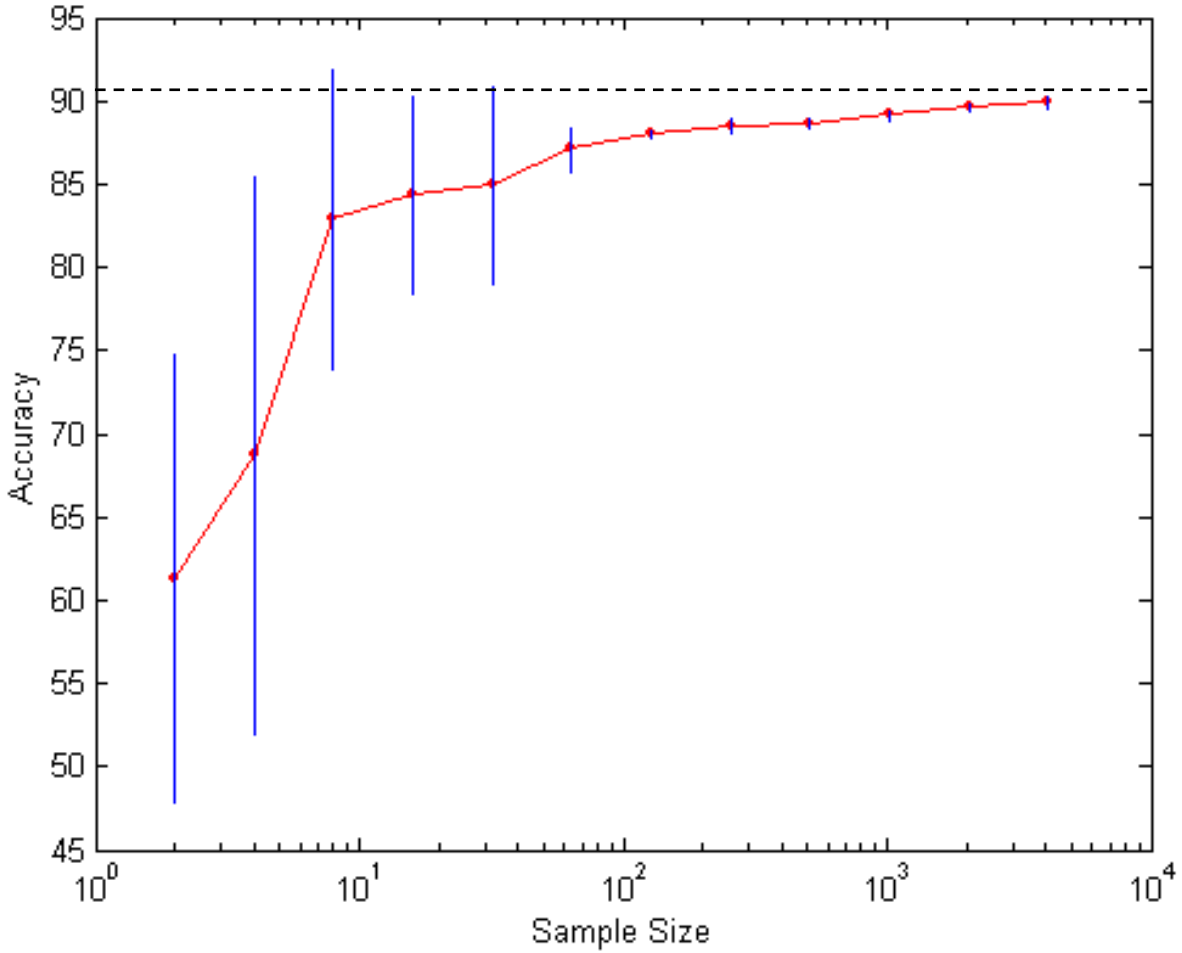
Manca qualche caso?

Troppo specifico?

Problema: valutazione



Learning curve



La **cardinalità del learning set** influenza l'accuratezza del modello appreso

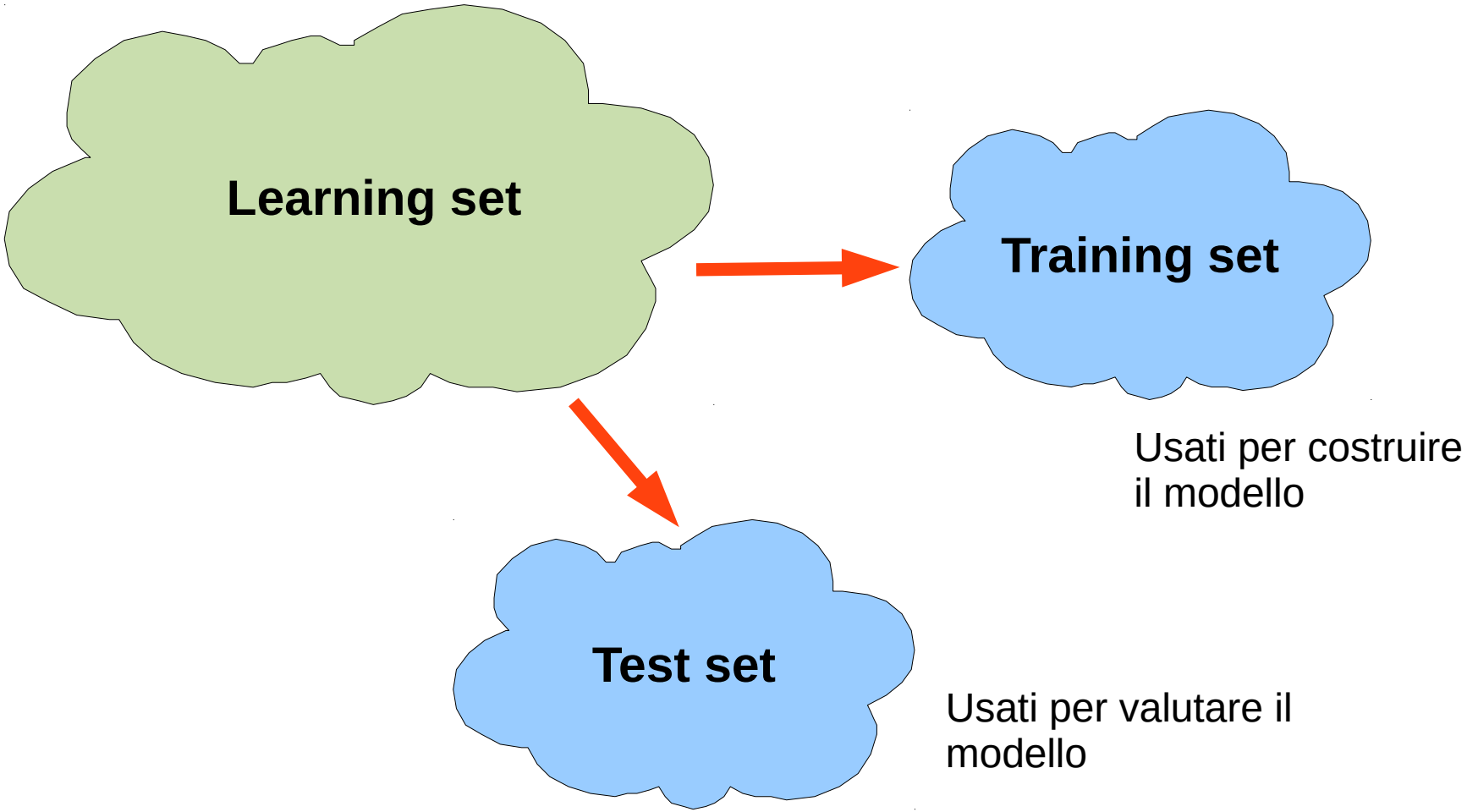
Se il campione è piccolo:

- modello focalizzato
- risultati poco affidabili

Metodi per la valutazione

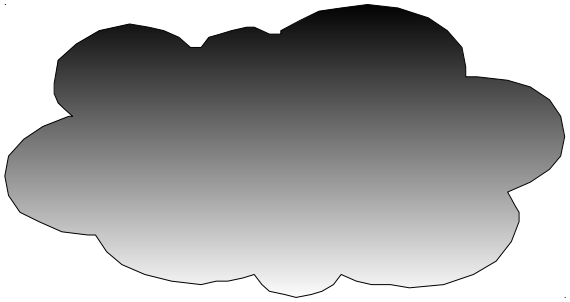
- Sono tutte valutazioni fatte su test set:
  - **Holdout**: partiziono i dati disponibili in learning/test set
  - **Random subsampling**: le prestazioni potrebbero dipendere dalla partizione effettuata, eseguiamone diverse e facciamo la media
  - **Cross-validation**: come sopra ma cerchiamo di usare i dati in modo omogeneo
  - **Bootstrap**: se ci sono pochi dati partizionare può produrre insiemi troppo piccoli per essere significativi

Holdout: divido i dati disponibili



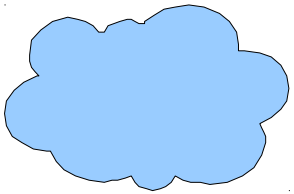
Possibili problemi

sottorappresentazione



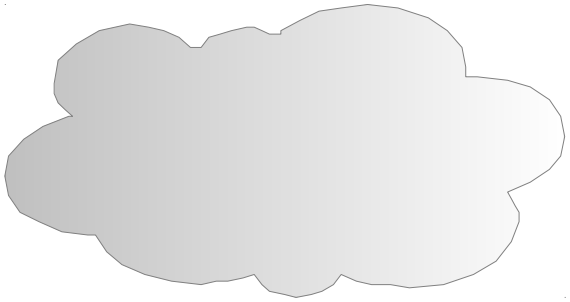
Il learning set non rappresenta bene tutte le classi

sovrarappresentazione

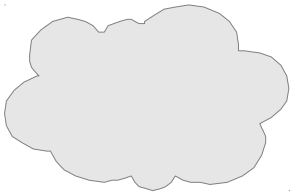


Il test set contiene prevalentemente esempi non rappresentati dal learning set

Possibili problemi

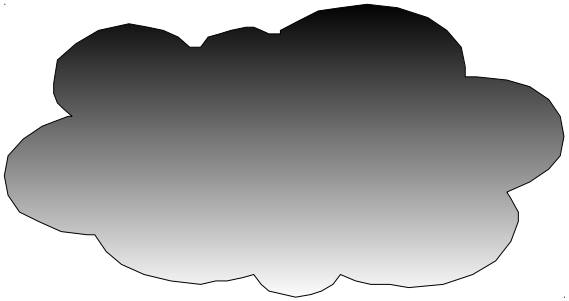


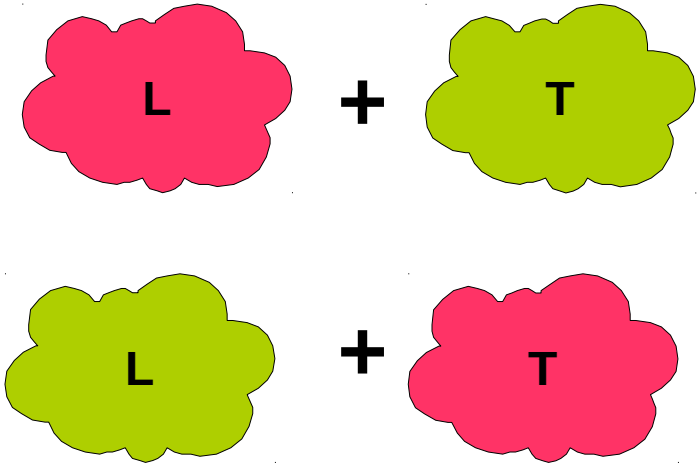
Il learning set non rappresenta bene tutte le classi



Il test set contiene prevalentemente esempi non rappresentati dal learning set

---





Variabilità della composizione (e quindi del modello appreso) dal taglio per learning set piccoli

Metodi per la valutazione

- Sono tutte valutazioni fatte su test set:
  - **Holdout**: partiziono i dati disponibili in learning/test set
  - **Random subsampling**: le prestazioni potrebbero dipendere dalla partizione effettuata, eseguiamone diverse e facciamo la media
  - **Cross-validation**: come sopra ma cerchiamo di usare i dati in modo omogeneo
  - **Bootstrap**: se ci sono pochi dati partizionare può produrre insiemi troppo piccoli per essere significativi

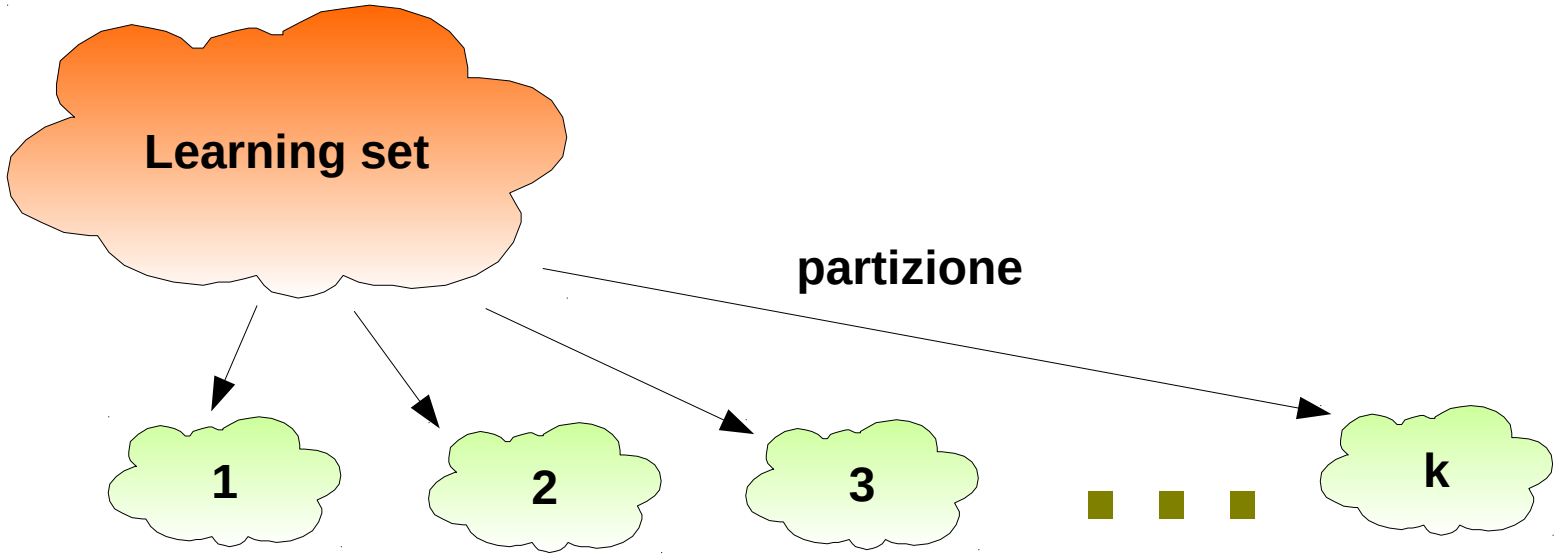
Metodi per la valutazione

- Sono tutte valutazioni fatte su test set:
  - **Holdout**: partiziono i dati disponibili in learning/test set
  - **Random subsampling**: le prestazioni potrebbero dipendere dalla partizione effettuata, eseguiamone diverse e facciamo la media
  - **Cross-validation**: come sopra ma cerchiamo di usare i dati in modo omogeneo
  - **Bootstrap**: se ci sono pochi dati partizionare può produrre insiemi troppo piccoli per essere significativi

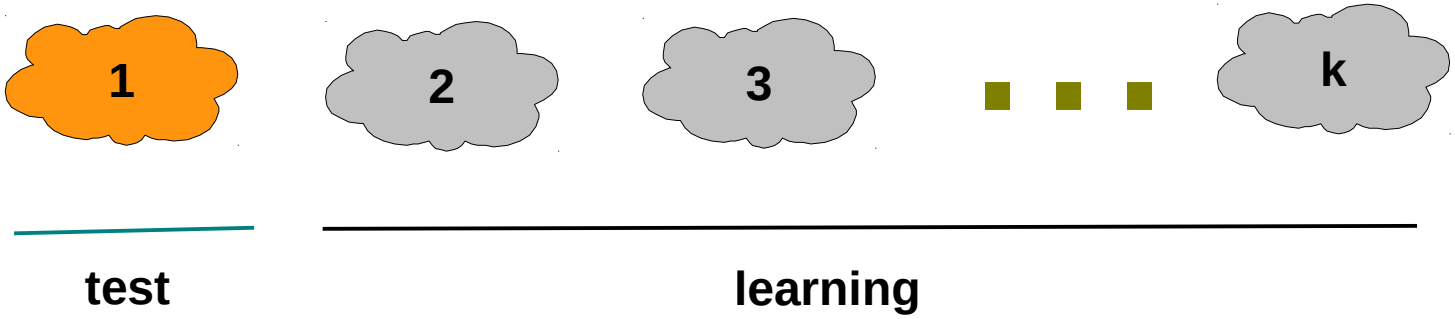
Cross-validation

Analogo al metodo precedente, caratteristica principale: tutti gli esempi sono usati lo stesso numero di volte per il training ed una volta sola per il test

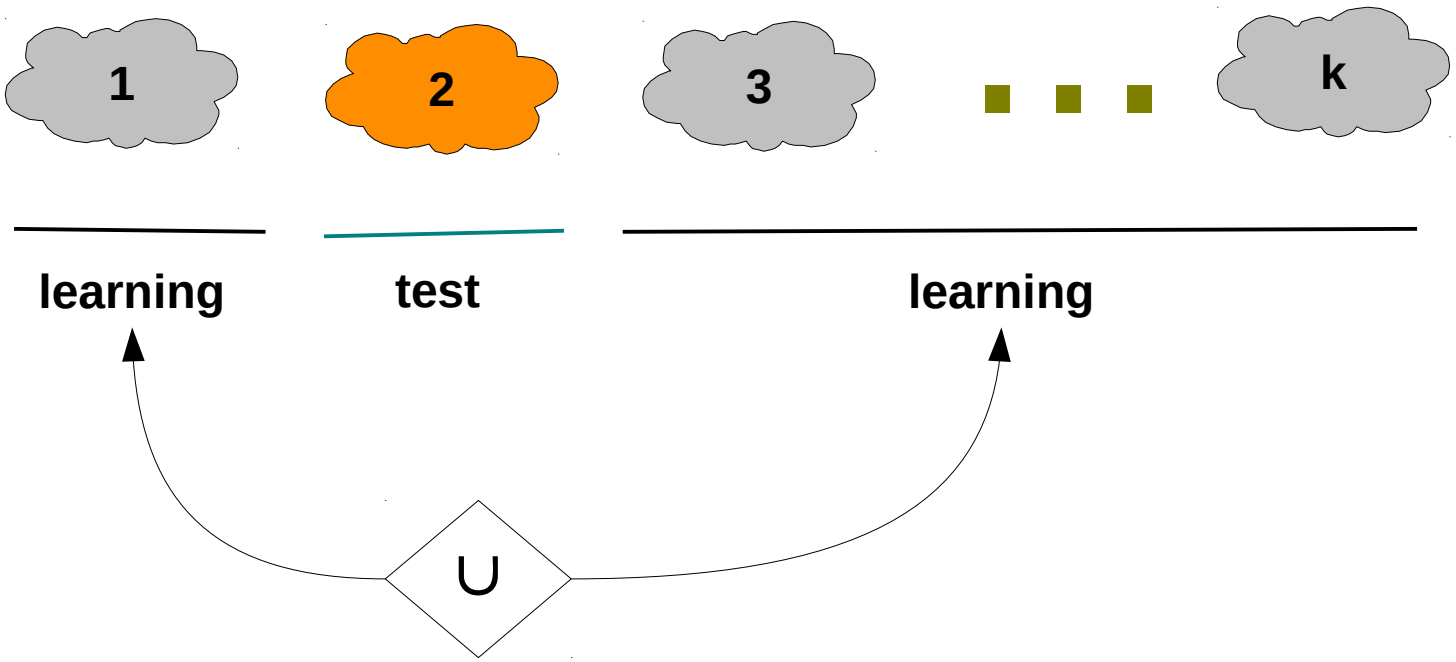
K-fold cross-validation



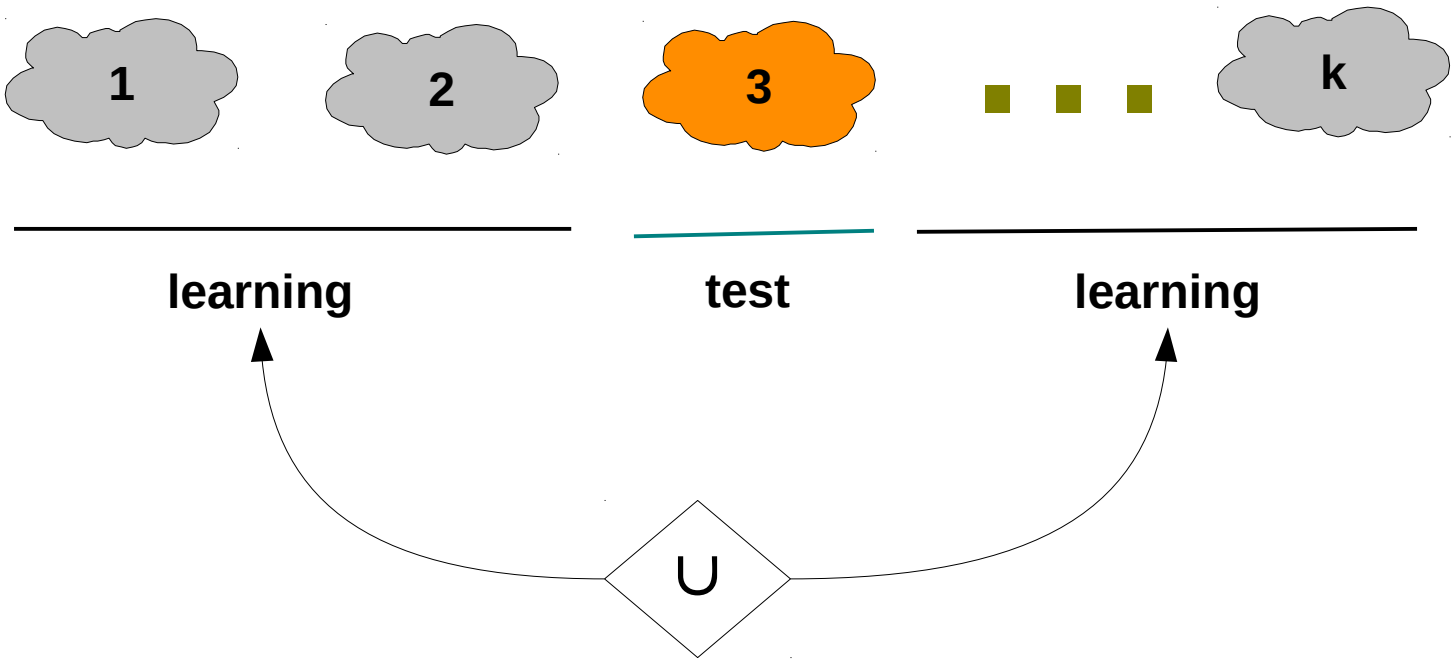
K-fold cross-validation



K-fold cross-validation



K-fold cross-validation



Ecc …

Metodi per la valutazione

- Sono tutte valutazioni fatte su test set:
  - **Holdout**: partiziono i dati disponibili in learning/test set
  - **Random subsampling**: le prestazioni potrebbero dipendere dalla partizione effettuata, eseguiamone diverse e facciamo la media
  - **Cross-validation**: come sopra ma cerchiamo di usare i dati in modo omogeneo
  - **Bootstrap**: se ci sono pochi dati partizionare può produrre insiemi troppo piccoli per essere significativi

Sampling con replacement: gli esempi su cui fare il training sono selezionati dall'insieme che sarà usato per il training ma non vengono rimossi da questo

Training set:

1) scelgo un'istanza e la aggiungo al training set
2) non rimuovo l'istanza dall'insieme originario !!
3) torno al punto 1)

NB: una stessa istanza può comparire più volte nel training set

Test set: insieme degli esempi originari non selezionati

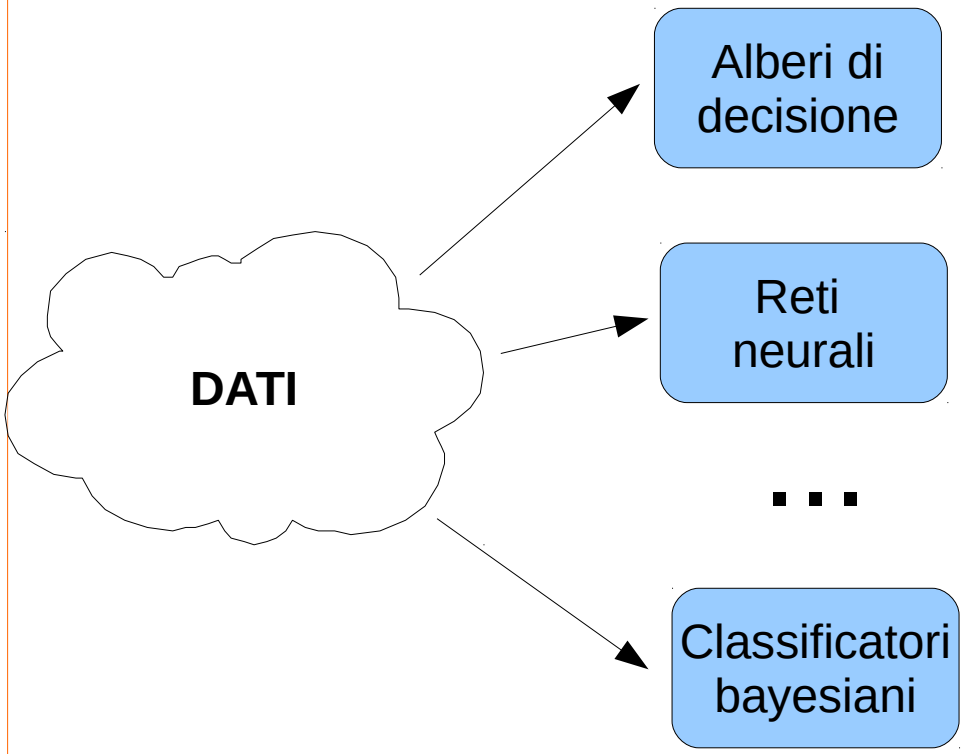
Fatti apprendimento e valutazione, si ripete il tutto per un numero di volte a piacere. Poi si calcola l'accuratezza media.

In molti casi produce una valutazione più accurata della cross-validation.

Ne esistono molte versioni ...

Confrontare modelli diversi

Problema



Ogni classificatore avrà un grado di accuratezza calcolato con una delle tecniche viste

In generale non si può contare sul fatto che i test siano stati fatti sugli stessi (sotto-)insiemi di dati!

Se uso la cross-validation o il random subsampling gli insiemi di learn/test cambiano ogni volta

Le accuratezze calcolate sono relative a basi diverse

Problema: l'accuratezza calcolata su un certo test set è una misura generale della bontà di un modello?

In altri termini …

Chiedo a un campione di 1000 persone quale marca di aranciata preferiscono e il 60% dice “la marca X”

Ora chiedo a un altro campione di 1000 persone della stessa città quale marca di aranciata preferiscono: *quanto è probabile che esattamente il 60% mi dica “la marca X”?*

In altri termini …

Chiedo a un campione di 1000 persone quale marca di aranciata preferiscono e il 60% dice “la marca X”:

Ora chiedo a un altro campione di 1000 persone della stessa città quale marca di aranciata preferiscono: *quanto è probabile che esattamente il 60% mi dica “la marca X”?*

Non basta dire “60%”, molto meglio prevedere di quanto si discosterà il risultato se cambio campione

Esempio: 60 ± 10 %

Cosa c'entrano le aranciate?

Dato un test set di 1000 istanze un classificatore mi dice che il 60% delle istanze è di correttamente classificato

Se io eseguo il test su un altro campione di 1000 istanze: *quanto è probabile che esattamente il 60% risulti correttamente classificato?*

Non basta dire "60%", molto meglio prevedere di quanto si discosterà il risultato se cambio campione

Esempio: 60 ± 10 %

Torniamo alle aranciate ...

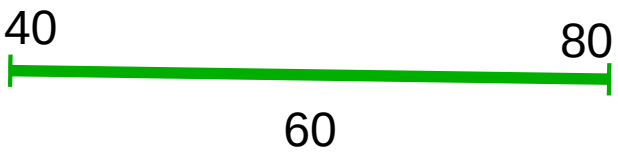
Intervallo di confidenza

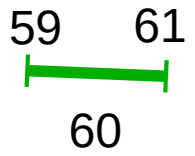
Se chiedo a un campione di 1000 persone quale marca di aranciata preferiscono e il 60% dice “la marca X”:

☺ si può essere **ragionevolmente certi** che *fra il 40 e l'80%* degli abitanti della città preferisce davvero la marca X (60 ± 20)

♠ Non si può essere altrettanto certi che fra il 59 e il 61% degli abitanti della città preferiscano la marca X (60 ± 1)

Intervallo di confidenza





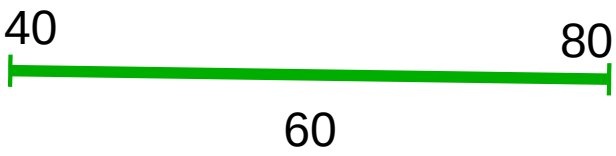
Livello di Confidenza

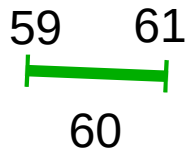
Se chiedo a un campione di 1000 persone quale marca di aranciata preferiscono e il 60% dice “la marca X”:

☺ si può essere ragionevolmente certi che fra il 40 e l'80% degli abitanti della città preferisce davvero la marca X (60 ± 20)

♠ Non si può essere altrettanto certi che fra il 59 e il 61% degli abitanti della città preferiscano la marca X (60 ± 1)

Intervallo di confidenza





Marca X ∈ [40, 80]: **95%**

Marca X ∈ [59, 61]: **75%**

Livello di confidenza

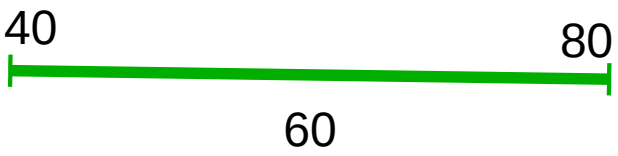
Tornando ai classificatori

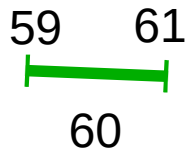
Se un modello costruito su un campione di 1000 istanze dice che il 60% è correttamente classificato

☺ si può essere ragionevolmente certi che fra il 40 e l'80% delle istanze di un altro campione sia davvero classificate correttamente (60 ± 20)

♠ Non si può essere altrettanto certi che fra il 59 e il 61% delle istanze di un altro campione qualsiasi siano classificate correttamente (60 ± 1)

Intervallo di confidenza



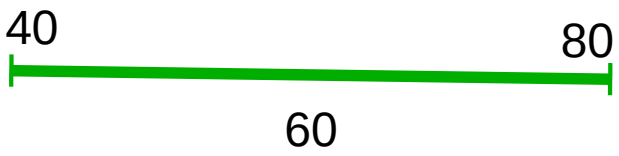


Corretti ∈ [40, 60]: **95%**

Corretti ∈ [59, 61]: **75%**

Livello di confidenza

Intervallo e Livello di Confidenza



Corretti ∈ [40, 80]: **95%**

Intervallo di confidenza: intervallo a cui si pensa un certo valore reale e ignoto appartenga

Livello di confidenza: probabilità che il valore reale e ignoto sia effettivamente compreso nell'intervallo dato

$$P(a \leqslant \theta \leqslant b) = 1 - \alpha$$

Estremi dell'intervallo di confidenza

Livello di confidenza

http://www.aiaccess.net/English/Glossaries/GlosMod/e_gm_confidence_interval.htm

Compiti possibili

- Dato un certo livello di confidenza (1 – α), calcolare l'ampiezza dell'intervallo di confidenza

  Esempio: data una stima dell'accuratezza pari a 92% arrivare a dire che, in generale, tale accuratezza varierà fra l'89% e il 95% con probabilità 95%

  Nota: Di solito si usano livelli di confidenza del 95% oppure del 99%

- Dato un intervallo di confidenza, calcolare il livello di confidenza
- Calcolare quanto deve essere ampio un campione per ottenere un certo livello di confidenza su un certo intervallo