Questa frase descrive il concetto di Online Bagging (OB), un adattamento dell'algoritmo di bagging tradizionale per l'apprendimento su stream di dati. Vediamo di scomporla e spiegarla passo per passo:

1. Nel bagging tradizionale, ogni classificatore base viene addestrato su un sottoinsieme di dati ottenuto tramite campionamento con sostituzione dall'insieme di training originale.

 In questo processo, ogni esempio può apparire K volte nel sottoinsieme di training di un classificatore, dove K segue una distribuzione binomiale.

 Nel contesto di Streaming Machine Learning (SML), il flusso di dati è potenzialmente infinito, quindi N (il numero totale di esempi) tende a essere molto grande.

 Quando N è grande, la distribuzione binomiale tende ad approssimare una distribuzione di Poisson con parametro λ=1.

 Sfruttando questa proprietà, l'Online Bagging, invece di effettuare un campionamento con sostituzione (difficile da realizzare in uno scenario di streaming), assegna a ogni esempio un peso secondo una distribuzione Poisson(1).

 La probabilità che un esempio abbia peso k è data da exp(-1)/k!, che è la funzione di massa di probabilità di una distribuzione Poisson con λ=1.

In pratica, questo approccio permette di simulare il processo di bagging in un contesto di streaming, dove non è possibile memorizzare tutti i dati e campionarli ripetutamente. Invece, ogni nuovo esempio viene pesato "al volo" secondo questa distribuzione, permettendo di mantenere la diversità tra i classificatori base tipica del bagging, ma in modo adatto all'apprendimento online.

Esempio: Bagging Tradizionale

Immaginiamo di avere un dataset di 1000 esempi e vogliamo creare 3 classificatori base.

 Per ogni classificatore, creiamo un nuovo dataset campionando 1000 volte con sostituzione dal dataset originale.

 Alcuni esempi potrebbero apparire più volte, altri potrebbero non apparire affatto.

3. Ogni classificatore viene addestrato sul proprio dataset.

Esempio: Online Bagging

Ora immaginiamo uno scenario di streaming dove gli esempi arrivano uno alla volta:

 Abbiamo 3 classificatori base inizializzati.

 Arriva un nuovo esempio.

 Per ogni classificatore, generiamo un peso secondo la distribuzione Poisson(1).

4. Il classificatore viene aggiornato usando l'esempio con il peso generato.

Vediamo come potrebbe funzionare per i primi 5 esempi:

Esempio 1:

- Classificatore A: peso = 1

- Classificatore B: peso = 0

- Classificatore C: peso = 2

Esempio 2:

- Classificatore A: peso = 1

- Classificatore B: peso = 1

- Classificatore C: peso = 0

Esempio 3:

- Classificatore A: peso = 0

- Classificatore B: peso = 2

- Classificatore C: peso = 1

Esempio 4:

- Classificatore A: peso = 1

- Classificatore B: peso = 1

- Classificatore C: peso = 1

Esempio 5:

- Classificatore A: peso = 3

- Classificatore B: peso = 0

- Classificatore C: peso = 1

Interpretazione:

 Un peso di 0 significa che il classificatore ignora quell'esempio.

 Un peso di 1 è come se l'esempio apparisse una volta nel dataset tradizionale.

 Un peso maggiore di 1 è come se l'esempio fosse ripetuto più volte.

Questo approccio simula il campionamento con sostituzione del bagging tradizionale, ma in modo adatto allo streaming:

1. Mantiene la diversità tra i classificatori.

 Non richiede di memorizzare l'intero dataset.

 Ogni esempio viene processato una sola volta al suo arrivo.

 La distribuzione Poisson(1) assicura che, in media, ogni esempio ha peso 1, ma con variazioni che simulano il campionamento casuale.

Questo metodo permette di applicare i benefici del bagging (come la riduzione della varianza e dell'overfitting) in un contesto di apprendimento online su flussi di dati potenzialmente infini