Stats Under the Stars[3]
Gruppo Call of Data

Artico Igor, D'Angelo Laura, Diquigiovanni Jacopo, Rossini Jacopo

28 maggio 2017

Light GBM e Bayesian Optimization

Light GBM è una particolare variante di *gradient boosting*, con alcune modifiche che lo rendono particolarmente vantaggioso. Si basa su alberi di classificazione, ma la scelta della foglia da "splittare" a ogni passo viene fatta in modo più efficace. Mentre il boosting opera una crescita dell'albero in profondità, Light GBM opera la scelta combinando due criteri: da un lato un'ottimizzazione basata su Gradient Descent, dall'altro, per evitare problemi di overfitting, viene posto un limite alla profondità massima. Questo tipo di crescita viene detta *leaf wise* ed è schematizzata in Figura 2, mentre in Figura 1 è rappresentato il metodo di split del boosting tradizionale.
Light GBM presenta numerosi vantaggi:

- **velocità di elaborazione**: poiché non fa crescere completamente gli alberi, e grazie al *binning* delle variabili (procedura che opera una divisione di queste in sottogruppi sia per velocizzare i calcoli sia come metodo di regolarizzazione), la procedura presentata è mediamente un'ordine di grandezza più veloce rispetto ad algoritmi simili.
- **utilizzo più parsimonioso della memoria**: la procedura di *binning* comporta un utilizzo meno intensivo della memoria.
- **migliore accuratezza** rispetto agli usuali algoritmi di boosting: poichè utilizza una procedura leaf-wise, gli alberi ottenuti risultano più complessi. Allo stesso tempo, per evitare l'overfitting, viene posto un limite sulla profondità massima.
- l'algoritmo è facilmente **parallelizzabile**.

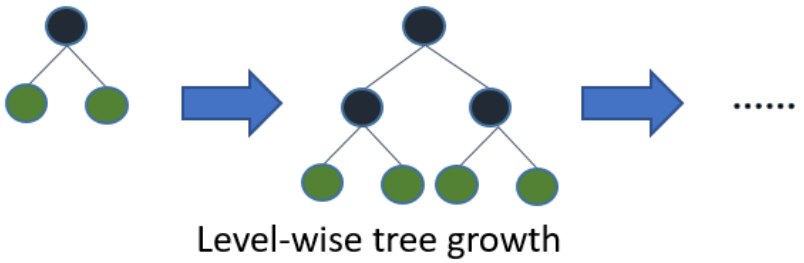

Figura 1: Rappresentazione della crescita degli alberi usata nel boosting tradizionale

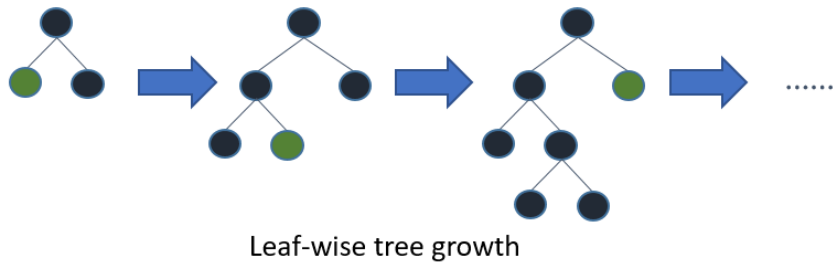

Figura 2: Rappresentazione della crescita degli alberi usata in Light GBM

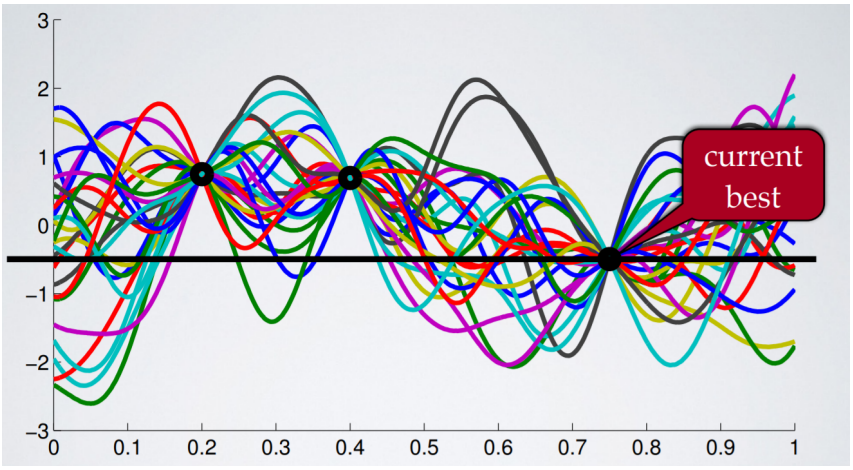

Figura 3: Idea dell'ottimizzatore bayesiano: possibili funzioni di perdita

Grazie alla velocità di calcolo offerta dal nostro modello, abbiamo deciso di raffinare il nostro metodo di ricerca dei parametri implementando un algoritmo di ottimizzazione bayesiana per i parametri di Light GBM.
La metodologia utilizzata valuta iterativamente possibili *set* di parametri in modo da trovarne valori competitivi. Per ottimizzare quindi la funzione di perdita (nel nostro caso l'accuratezza di classificazione delle prime 10000 osservazioni, ordinate per probabilità prevista in ordine decrescente) procederemo a calcolare su una serie di punti i valori di interesse, e approssimeremo le zone prive di punti ipotizzando che la funzione da ottimizzare sia relizzazione di un processo gaussiano (Figura 3). Tale ipotesi permette di quantificare l'incertezza nelle zone in questione tramite la creazione di bande di confidenza, e grazie al processo di integrazione, anche individuare aree in cui i parametri sembrano più promettenti. I criteri di scelta dell'insieme di parametri che verranno utilizzati da Light GBM sono molteplici, e variano dal miglioramento atteso al limite di confidenza superiore. Questo procedimento ci garantisce un'imparzialità assoluta nelle decisioni, oltre a permetterci di esplorare in maniera molto approfondita lo spazio parametrico: tale approccio si rivela particolarmente utile, in quanto LGBM esiste da poco meno di due mesi e il *know-how* riguardo alla gestione dei parametri è certamente qualcosa di poco noto alla maggior parte degli utenti. Un approccio robusto come questo è quindi considerato opportuno.

Nella fase finale, grazie alle numerose iterazioni dell'ottimizzatore e ai molteplici modelli così ottenuti, è stato possibile combinare i risultati predittivi dei migliori 10 modelli in un *ensemble*, creata con una media pesata, al fine di migliorare ulteriormente le capacità predittive complessive.