Gradient Boosted Tree

Rappel

Boosting tree correspond à une agrégation de modèles adaptatifs les uns des autres. Les modèles ici sont donc des arbres CART (d'où le "boosted tree"). Pourquoi parle t-on de *gradient* boosted tree ?

Théorie

On veut donc construire un modèle h_M tel que $h_M(x) = \sum_{t=1}^{M} \alpha \,.\, \delta_t(x)$ dans l'espoir de minimiser $E(L(h_M(X),\ Y))$ avec L une fonction de coût. Soit $\{(x_i,\ y_i)\}_{i=1,\ ...,\ n}$ des réalisations du couple $(X,\ Y)$ et soit $h_{m-1}(x)$ posé.

Sachant que $h_m(x)\ =\ h_{m-1}(x)\ +\ \alpha \,.\, \delta_m(x)$ avec α une constante, on cherche $\delta_m(x)$ un modèle d'arbre tel que :

$$\sum_{i=1}^{n} L(y_i,\ h_m(x_i))\ <\ \sum_{i=1}^{n} L(y_i,\ h_{m-1}(x_i)) \qquad (1)$$

$$\sum_{i=1}^{n} L(y_i,\ h_{m-1}(x_i)\ +\ \alpha \,.\, \delta_m(x)\ )\ <\ \sum_{i=1}^{n} L(y_i,\ h_{m-1}(x_i)) \qquad (2)$$

$x\ \to L(y,\ x)$ étant strictement convexe, on sait que :

$$L(y,\ x\ -\ h \,.\, \nabla_x L(y,\ x)\ )\ <\ L(y,\ x) \quad \forall\ x \neq x_{min} \qquad \text{pour } h \text{ assez petit} \qquad (3)$$

Cela se démontre avec un développement de Taylor à l'ordre 1. On a donc en remplaçant x et h dans (3) respectivement par $h_{m-1}(x_i)$ et α :

$$\sum_{i=1}^{n} L(y_i,\ h_{m-1}(x_i)\ +\ \alpha \,.\, g_i\ )\ <\ \sum_{i=1}^{n} L(y_i,\ h_{m-1}(x_i)\ ) \qquad (4)$$

avec $g_i\ =\ -\nabla_{h_{m-1}(x_i)} L(y_i,\ h_{m-1}(x_i))$ (appelé negative gradient ou résidus) et α assez petit. g_i étant dépendant de y_i, il nous faut les approcher avec un arbre de régression au sens de la norme $L2$ pour conserver l'inégalité (4). On va donc fitter un arbre de régression δ_m sur les negative gradient g_i.

On peut donc mettre au points un algorithme où on initialise $h_0(x)$ par exemple la moyenne des réalisation y_i dans un problème de régression, puis on fit les arbres δ_k sur les negative gradient $g_i\ =\ -\nabla_{h_{k-1}(x_i)} L(y_i,\ h_{k-1}(x_i))$ à chaque itération. On choisira un pas α assez petit. Le caractère strictement convexe de L assure une convergence de l'algorithme vers un minimum global.

Aller plus loin

Pour éviter l'over fitting, XGBoost propose une pénalisation des arbres, et en partant d'un développement de Taylor à l'ordre 2 de (2) on aboutit à une solution approchée d'arbres originaux.