



Machine Learning 101

Boosted Trees

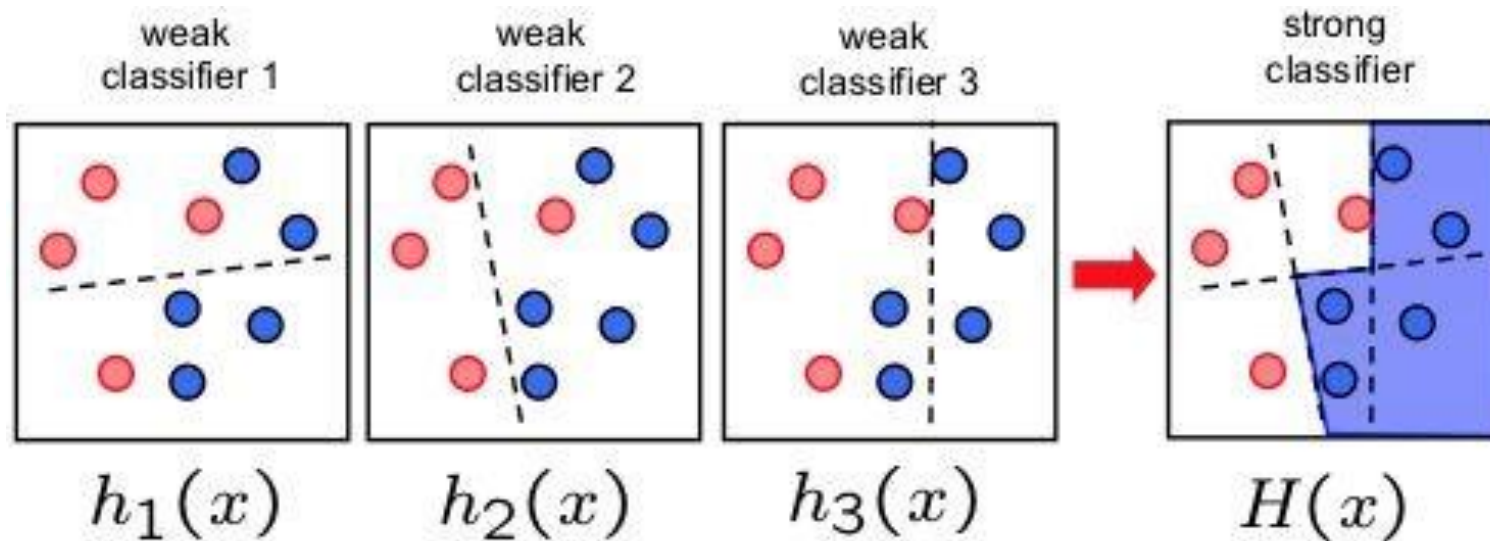


■ Introducción

- Una de las ideas más brillantes en machine learning de los últimos tiempos
- Proceso iterativo basado en la combinación lineal de algoritmos sencillos (*weak classifiers/regressors*)
- Aplicable a cualquier algoritmo de ML. Normalmente árboles (Boosted Trees)
- Una de los algoritmos más utilizados para competir y ganar en Kaggle



Intuición

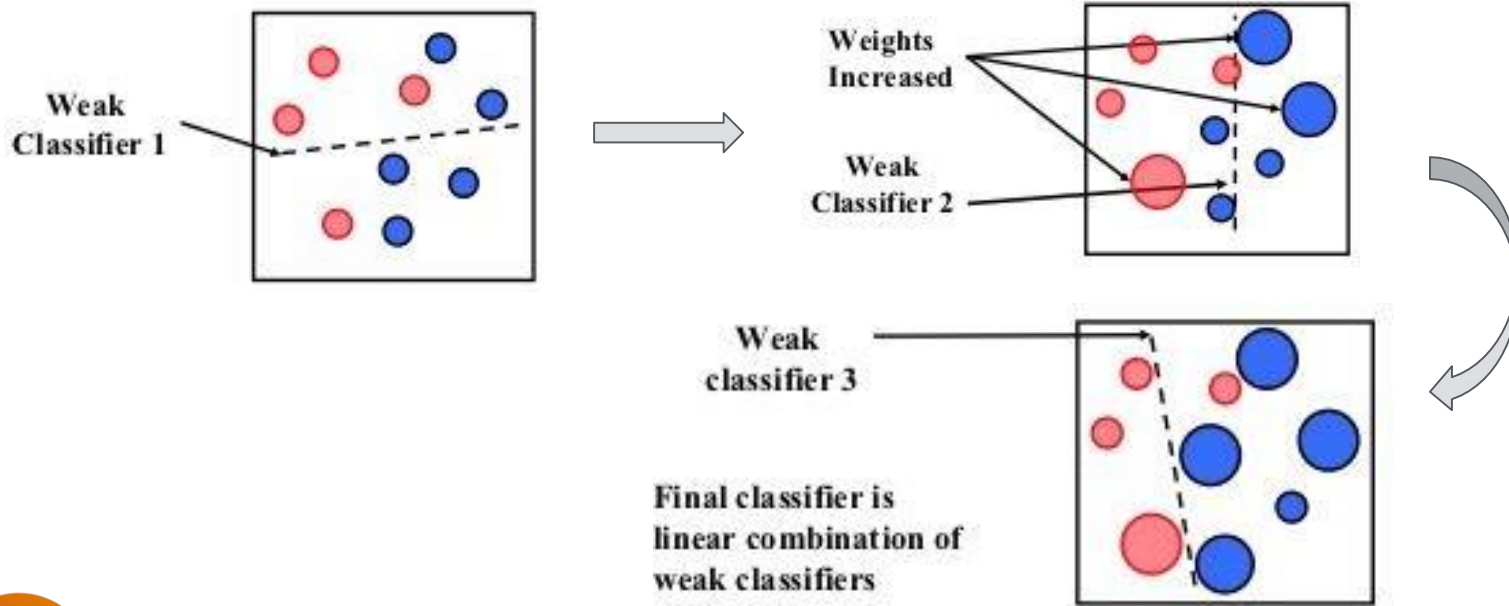


$$H(x) = \text{sign}(\alpha_1 h_1(x) + \alpha_2 h_2(x) + \alpha_3 h_3(x))$$



Proceso iterativo

- En cada iteración le damos más peso a los errores



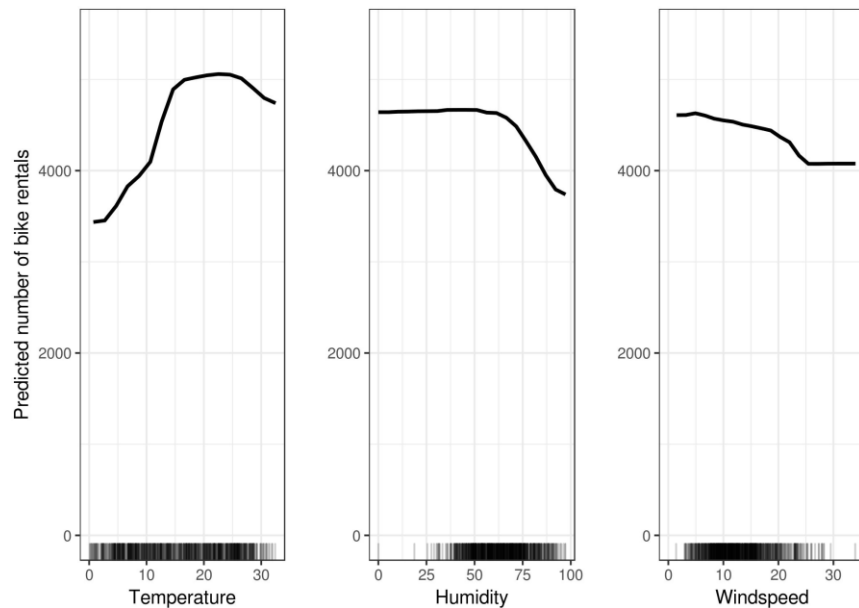
■ Hiperparámetros

- **Número de árboles** (iteraciones). Si es muy alto peligro de overfitting. Seleccionamos con validación cruzada
- **Tasa de aprendizaje** (alpha). Número positivo de valor pequeño, normalmente 0.01, 0.001. Está relacionado con el número de iteraciones. Si alpha es pequeño, se necesitarán más iteraciones para que las prestaciones converjan
- **Profundidad del árbol**. Controlamos la complejidad del árbol. Idealmente interesa que sea pequeña, así la capacidad de generalización es mayor. Se recomienda comenzar con valores pequeños.



■ Interpretabilidad: *partial dependence plots*

- Efecto marginal de cada característica con variable target predicha (regresión)

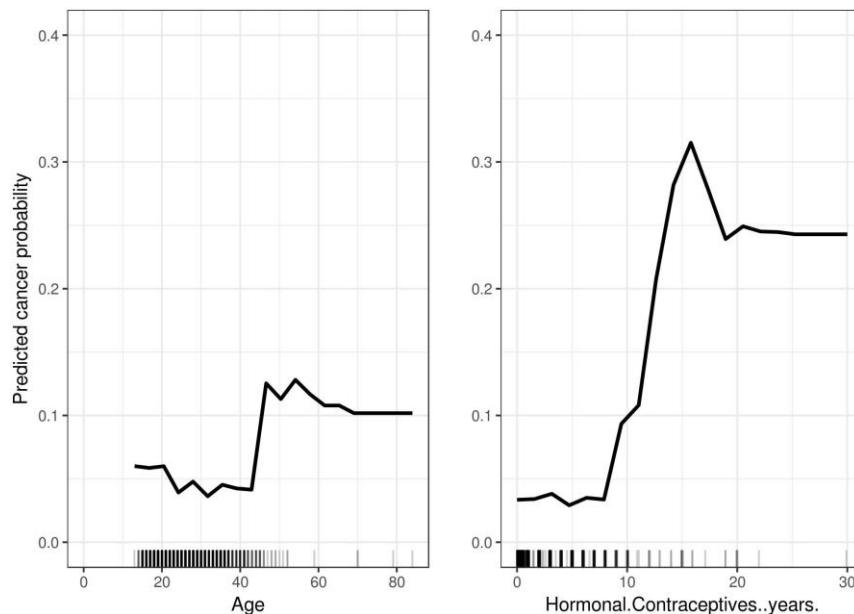


[interpretable machine learning](https://www.keencoding.io)



■ Interpretabilidad: *partial dependence plots*

- Efecto marginal de cada característica con variable target predicha (clasificación)



[interpretable machine learning](https://www.keenecoding.io)



■ Implementaciones

- sklearn: [Gradient Boosting](#)
- [XGBoost](#)
 - 5x más rápido que sklearn GradientBoosting
- [LightGBM](#)
 - ¿más eficiente todavía?



■ Árboles de decisión (*tree-based models*)

- Relaciones no lineales
- Sin necesidad de escalar variables
- Árbol aislado, muy interpretable (si no es muy profundo)
- Random Forest, algoritmo muy robusto, es un buen benchmark
- Mejores prestaciones, si se eligen con cuidado los parámetros libres
 - La potencia sin control no sirve de nada



■ Referencias

- The Elements of Statistical Learning.
 - Capítulo 10
- Introduction to Statistical Learning.
 - Capítulo 8, sección 3



Let's code!

