ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Introducción a Bosques Aleatorios

Andrés G. Abad, Ph.D.

Agenda

Introducción al problema de clasificación

Bosques aleatorios (random forests)
- Introducción a los bosques aleatorios
- Error Out-of-bag
- Importancia de las variables (características)

Referencias Bibliográficas

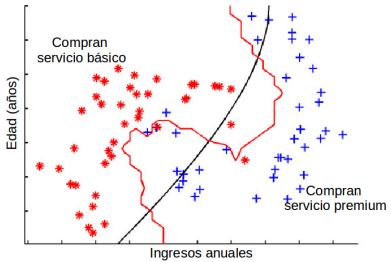
Definición del problema de clasificación I

- Un objeto $\mathbf{x} = [x_1, \ldots, x_p]$, con características x_i, pertenece exactamente a una clases $c \in \{1, 2, \ldots, C\}$.
- Asumimos que tenemos un conjunto de datos

$$\mathcal{D} = \{(\mathbf{x}^{(1)}, c^{(1)}), \ldots, (\mathbf{x}^{(n)}, c^{(n)})\}$$

- Buscamos una función $\hat{f}$ que asigne $\mathbf{x}^{(i)}$ a $c^{(i)}$ lo mejor posible:

$$\hat{f} = \arg\min_{f} \mathbb{P}_{(\mathbf{x},c)}[\mathbb{1}(f(\mathbf{x}) \neq c)]$$

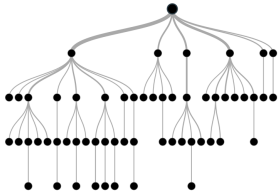


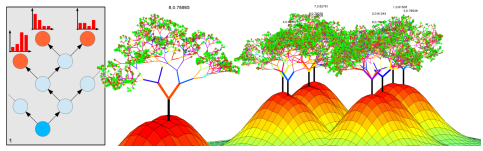
- Objeto $\mathbf{x}$ pertenece a una de dos clases: $\{Basico, Premium\}$
- Objeto $\mathbf{x}$ medidos en dos características: x_1 ingresos anuales, y x_2 edad en años
- Dos clasificadores $\hat{f}$'s: convexo-cuadrático (linea negra) y no-convexo (linea roja)

Agenda

Bosques aleatorios (random forests) I

Los árboles de clasificación sufrir de sobre-ajuste (bajo sesgo - alta variabilidad)





- Los bosques aleatorios (introducidos en Breiman [2001]) combinan varios árboles de decisión cada uno entrenado en diferentes partes del conjunto de entrenamiento para reducer la variabilidad
  - reducen el error de predicción
- El conjunto de árboles de decisión operan a manera de conjunto de *expertos* votando por la predicción de la clase del objeto
  - El objeto es asignado a la clase con más votos

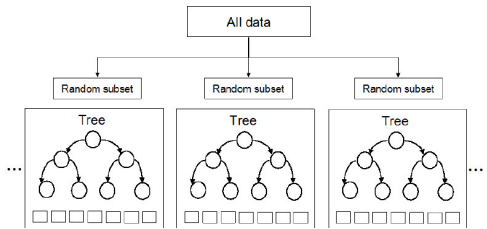
Tree bagging I

Tree bagging (bootstrap aggregating)

Para $b = 1, \ldots, B$:

1. Tome una muestra con reemplazo de tamaño n de datos X_b y Y_b
2. Entrene un árbol de clasificación $\hat{f}_b$ en la muestra X_b y Y_b.

Con los B árboles entrenados se conforma un clasificador $\hat{f}$ que asigna el objeto $\mathbf{x}$ a la clase con más *votos* entre los B árboles



Bosque aleatorio I

Bosque aleatorio

Se sigue casi el mismo procedimiento que en Tree Bagging con la siguiente adición:

- En cada nodo se escoge solo un subconjunto de las características como candidatas
- Si se tienen p características, en Hastie et al. [2003] se recomienda escoger $\sqrt{p}$ (rendondeado hacia abajo)

Out-of-bag error I

Es un método para predecir el error de predicción para bosques aleatorios

Def. Out-of-bag error (OOB)

Es el promedio de los errores de predicción sobre cada objeto de entrenamiento $\mathbf{x}_i$, utilizando solamente los árboles que no tenian a $\mathbf{x}_i$ en su muestra de entrenamiento.

Importancia de las variables (características) I

Es un método para estimar la importancia de cada característica (variables) en la predicción.

Def. Importancia de las variables (características)

1. Se obtiene el error OOB para cada objeto de la muestra y se promedio sobre todos los árboles
2. Para medir la importancia de la característica $\mathbf{x}_j$ permutamos el valor de la característica en la muestra y calculamos de nuevo el error OOB en la nueva muestra permutada
3. La medida de importancia de la característica $\mathbf{x}_j$ se mide como el promedio sobre todos los árboles de las diferencias del error OOB entre la muestras antes y después de la permutación

Referencias Bibliográficas I

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.

Hastie, T., Tibshirani, R., and Friedman, J. (2003). *The Elements of Statistical Learning*. Springer.