Tree-Based Methods

Ciencia de Datos y Econometría Aplicada

Ignacio Sarmiento-Barbieri

Universidad de los Andes

Agenda

1. Bagging and Forests
   - Bagging
   - Random Forests

Agenda

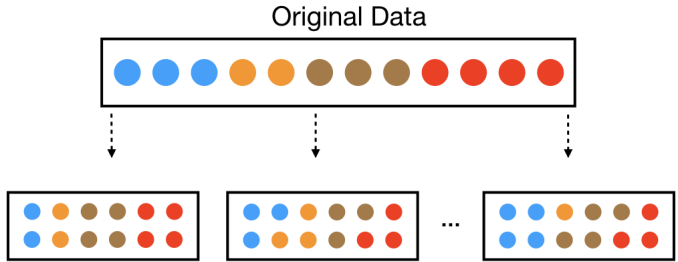
Bagging

- Problema con CART: pocos robustos.
- Idea: podemos mejorar mucho el rendimiento mediante la agregación

Bagging

- Bagging:
  - Obtenga muestras aleatorias $(X_i^b, Y_i^b)_{i=1}^N$ de la muestra observada (bootstrap).

Bagging



Bagging

- Bagging:
  - Obtenga muestras aleatorias $(X_i^b, Y_i^b)_{i=1}^N$ de la muestra observada (bootstrap).

Bagging

- Bagging:
  - Obtenga muestras aleatorias $(X_i^b, Y_i^b)_{i=1}^N$ de la muestra observada (bootstrap).
  - Para cada muestra, ajuste un árbol de regresión $\hat{f}^b(x)$
  - Promedie las muestras de bootstrap

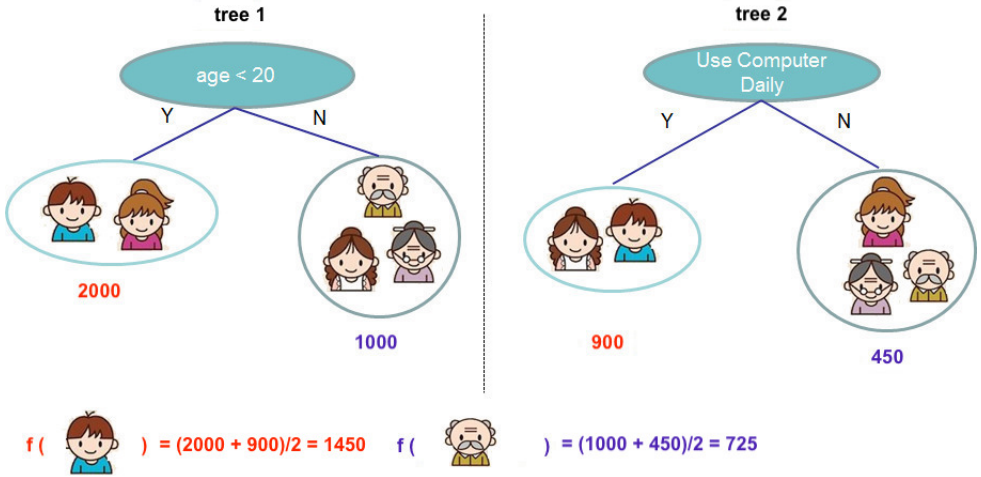
$$\hat{f}_{bag} = \frac{1}{B} \sum_{b=1}^{B} \hat{f}^b(x) \tag{1}$$

Bagging

- Bagging:
  - Obtenga muestras aleatorias $(X_i^b, Y_i^b)_{i=1}^N$ de la muestra observada (bootstrap).
  - Para cada muestra, ajuste un árbol de regresión $\hat{f}^b(x)$
  - Promedie las muestras de bootstrap

$$\hat{f}_{bag} = \frac{1}{B} \sum_{b=1}^{B} \hat{f}^b(x) \tag{1}$$

Bagging


tree 1
age < 20
Y
N
2000
1000
tree 2
Use Computer Daily
Y
N
900
450
f ( ) = (2000 + 900)/2 = 1450
f ( ) = (1000 + 450)/2 = 725

Ejemplo

photo from https://www.dailydot.com/parsec/batman-1966-labels-tumblr-twitter-vine/

Bagging

Out-of-Bag Error: Intuición

- En Bagging, entrenamos múltiples modelos sobre muestras bootstrap.
- Cada muestra contiene $\sim$63% de los datos originales.
- Estas muestras no usadas se llaman Out-of-Bag (OOB).

Bagging

Out-of-Bag Error: Matemáticamente

Sean:

- $\{(x_i, y_i)\}_{i=1}^n$: conjunto de entrenamiento.
- $T_1, \ldots, T_B$: arboles entrenados en cada muestra bootstrap.
- $\mathcal{B}_i \subset \{1, \ldots, B\}$: índices de modelos donde x_i no fue usado (OOB).

Entonces, el error OOB es:

$$E_{OOB} = \frac{1}{n} \sum_{i=1}^{n} L \left(y_i, \frac{1}{|\mathcal{B}_i|} \sum_{b \in \mathcal{B}_i} T_b(x_i) \right)$$

- L función de pérdida (error cuadrático, $0-1$, etc.).
- Estima el **error de generalización, fuera de muestra** sin necesidad de un set de validación.

Bagging

Importancia de Variables por Permutación: Intuición

- En modelos tipo Random Forests, queremos saber:

  ¿Qué tan importante es cada variable para las predicciones?

- La idea:
  - Medimos cuánto empeora el desempeño del modelo si rompemos la relación entre una variable y el objetivo.
- ¿Cómo rompemos esa relación?
  - Permutando los valores de la variable, dejando las otras igual.
- Si la variable es importante, el error debería aumentar.

Bagging

Importancia por Permutación: Definición

- Sea E_{OOB}: error del modelo usando los datos OOB.
- Para cada variable j:
  - Permutamos la columna x_j en los datos OOB.
  - Calculamos el nuevo error: $E_{\text{perm}(j)}$.
- La importancia de la variable j es:

$$\text{VI}(j) = E_{\text{perm}(j)} - E_{\text{OOB}}$$

- Si x_j es irrelevante, el error no debería cambiar.
- Si x_j es clave, el error aumentará.

Agenda

1. Bagging and Forests
   - Bagging
   - Random Forests

Random Forests

- Problema con el bagging: si hay un predictor fuerte, diferentes árboles son muy similares entre sí.
- Bosques (forests): reduce la correlación entre los árboles en el boostrap.
- Si hay p predictores, en cada partición use solo $m < p$ predictores, elegidos al azar.
- Bagging es forests con $m = p$ (usando todo los predictores en cada partición).
- m es un hiper-parámetro, $m = \sqrt{(p)}$ es un benchmark

Ejemplo

photo from https://www.dailydot.com/parsec/batman-1966-labels-tumblr-twitter-vine/