



Machine Learning 101

Bagging y Random Forest



■ Motivación

- Combinar algoritmos, normalmente árboles, para mejorar sus prestaciones
- Proporcionan grandes prestaciones en problemas complejos



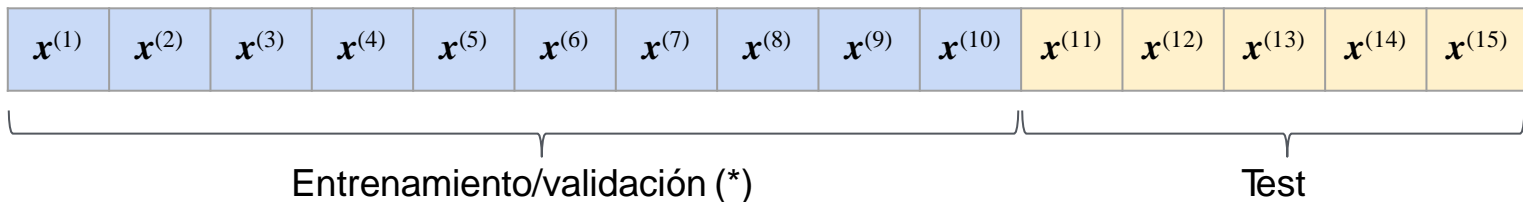
Índice

1. **Remuestreo Bootstrap**
2. Bagging
3. Random Forest
4. Importancia variables



Remuestreo Bootstrap

- Técnica estadística para cuantificar la incertidumbre de un estimador
 - En ML nos sirve para **medir las prestaciones de un algoritmo**
- Supongamos un problema de aprendizaje supervisado, donde disponemos de un conjunto de datos etiquetados $\{X, y\}$, con $N = 15$.

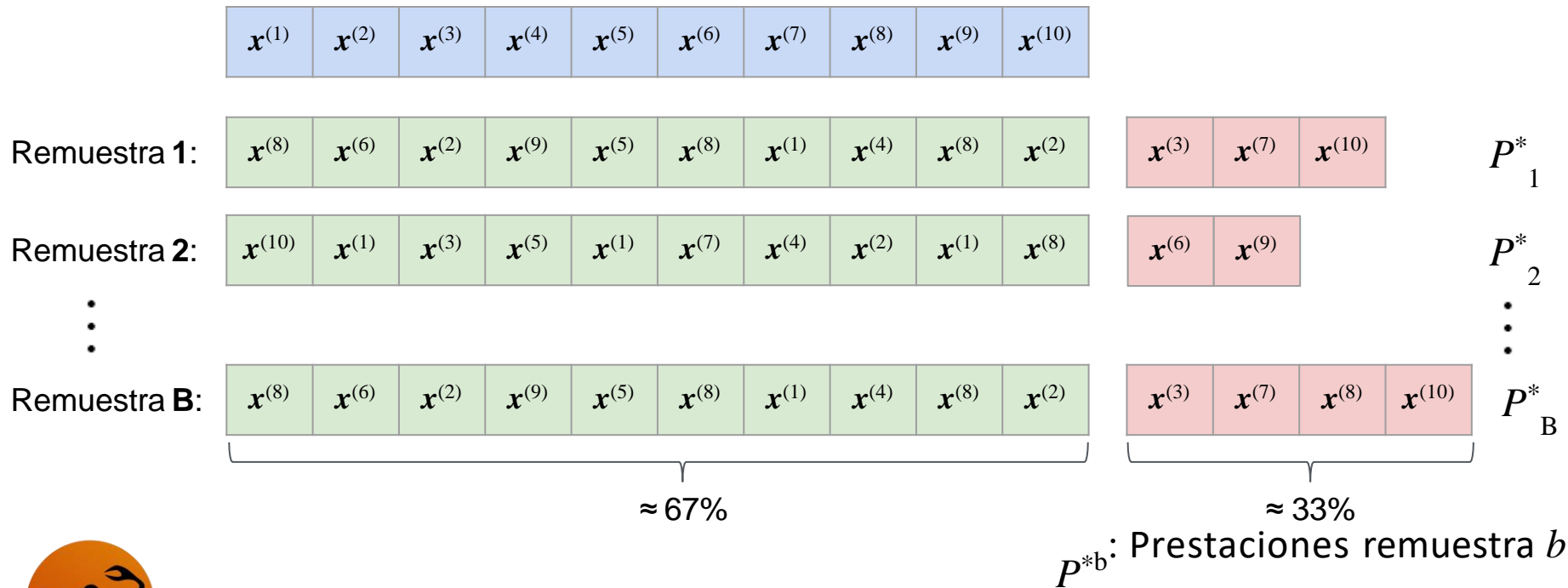


(*) numeración no es orden, los datos han sido ya aleatorizados



Remuestreo Bootstrap

- Bootstrap: remuestras con repetición



■ *Out-of-bag performance estimation*

- Out-of-bag, remuestra b: P_b^*
- Prestaciones totales

$$OOB = \frac{1}{B} \sum_{b=1}^B P_b^*$$

- Normalmente $B = 200-500$
- Al promediar reducimos la varianza del estimador (es similar cross-validation)



Índice

1. Remuestreo Bootstrap
2. **Bagging**
3. Random Forest
4. Importancia variables



■ Bagging: Bootstrap AGGregation

- Motivación: **reducir varianza** de los árboles de decisión (en función de la división los resultados pueden ser muy distintos)
- Utilizar *bootstrap* para **combinar árboles de decisión**:
 - Se construyen (entrenan) B árboles utilizando B remuestras

$\mathbf{x}^{(8)}$	$\mathbf{x}^{(6)}$	$\mathbf{x}^{(2)}$	$\mathbf{x}^{(9)}$	$\mathbf{x}^{(5)}$	$\mathbf{x}^{(8)}$	$\mathbf{x}^{(1)}$	$\mathbf{x}^{(4)}$	$\mathbf{x}^{(8)}$	$\mathbf{x}^{(2)}$
--------------------	--------------------	--------------------	--------------------	--------------------	--------------------	--------------------	--------------------	--------------------	--------------------

$\mathbf{x}^{(8)}$	$\mathbf{x}^{(6)}$	$\mathbf{x}^{(2)}$	$\mathbf{x}^{(9)}$	$\mathbf{x}^{(5)}$	$\mathbf{x}^{(8)}$	$\mathbf{x}^{(1)}$	$\mathbf{x}^{(4)}$	$\mathbf{x}^{(8)}$	$\mathbf{x}^{(2)}$
--------------------	--------------------	--------------------	--------------------	--------------------	--------------------	--------------------	--------------------	--------------------	--------------------

- Se combina la salida para predecir una nueva muestra: $\mathbf{x}^{(new)}$



■ Bagging

- Se combina la salida para predecir una nueva muestra: $\mathbf{x}^{(new)}$

- Regresión: $\hat{y} = \frac{1}{B} \sum_{b=1}^B f_{b,TREE}^*(\mathbf{x}^{(new)})$

- Clasificación: *majority vote* $\hat{y} = \max_{k=1,\dots,K} \left\{ \sum_{b=1}^B f_{b,TREE}^*(\mathbf{x}^{(new)}) \right\}$

- Se estiman las prestaciones mediante Out-Of-Bag



■ *Bagging: pros and cons*

- OK
 - mejoran las prestaciones sustancialmente
- KO
 - Si hay uno o varios predictores fuertes, puede que los B árboles generados sean bastante similares, por lo que no estamos reduciendo la varianza dado que los árboles están altamente correlacionados



Índice

1. Remuestreo Bootstrap
2. Bagging
3. **Random Forest**
4. Importancia variables



■ *Random forest*

- Motivación: **decorrelacionar** árboles remuestrados
- Utilizar *bootstrap* para **combinar árboles de decisión**:
 - Se construyen (entrenan) B árboles utilizando B remuestras
 - En la construcción de cada árbol, para cada *split* se fuerza a utilizar un subconjunto aleatorio de $m < d$ predictores
- Normalmente $m = \sqrt{d}$
- Si $m = d$, entonces es Bagging
- Si el número de predictores relevantes es pequeño, y alta dimensionalidad, peligro de *overfitting*



Índice

1. Remuestreo Bootstrap
2. Bagging
3. Random Forest
4. **Importancia variables**

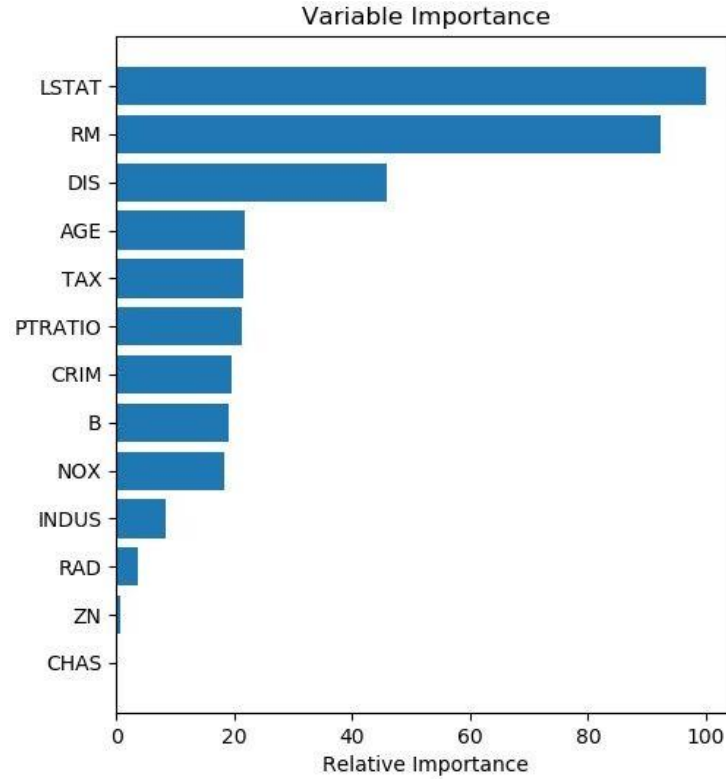


■ Importancia de las variables

- Con la agregación de árboles se pierde interpretabilidad
- No obstante se puede extraer una **medida de la importancia** de cada variable
 - Cuánto mejoran las prestaciones en los *splits* asociados a dicha variable (*ESL, página 368*)
 - En otras palabras: para cada split de cada árbol construido, se mide la mejora en prestaciones debido a la variable por la que se particiona el árbol.
- Medida relativa: se escala entre 0-100
- Puede aplicarse a un árbol individual, pero no es concluyente
- Se puede utilizar como ranking en selección de características, ¡pero hay que hacerlo bien! (wrapper)



■ Importancia de las variables



■ Referencias

- Introduction to Statistical Learning
 - Capítulo 5, sección 2
 - Capítulo 8, sección 2
- The Elements of Statistical Learning
 - Capítulo 10, sección 13
 - Capítulo 15
- Hands On Machine Learning.
 - Capítulo 7



Let's code!

