Arboles de clasificación (Parte II).

César Gómez

21 de octubre de 2020

Métodos de ensamble

Agregando muchos árboles de regresión, utilizando métodos como *bagging*, *random forest* (bosques aleatorios) y *boosting*. El desempeño a nivel de predicción de los árboles puede ser dramáticamente mejorado

- Hasta el momento, los árboles que se están construyendo, padecen el inconveniente de presentar una varianza alta.
- **Bagging** o *Bootstrap aggregation* es un procedimiento de propósito general para reducir la varianza de un método de aprendizaje estadístico.

Métodos de ensamble

Agregando muchos árboles de regresión, utilizando métodos como *bagging*, *random forest* (bosques aleatorios) y *boosting*. El desempeño a nivel de predicción de los árboles puede ser dramáticamente mejorado

- Hasta el momento, los árboles que se están construyendo, padecen el inconveniente de presentar una varianza alta.
- **Bagging** o *Bootstrap aggregation* es un procedimiento de propósito general para reducir la varianza de un método de aprendizaje estadístico.
- Consiste en realizar B muestras **Bootstrap** de los datos de entrenamiento, se genera un árbol no podado por cada muestra bootstrap que se ajuste a la misma y se obtienen predicciones

$$\hat{f}^{*1}(x), \hat{f}^{*2}(x), \ldots, \hat{f}^{*B}(x).$$

Métodos de ensamble

Agregando muchos árboles de regresión, utilizando métodos como *bagging*, *random forest* (bosques aleatorios) y *boosting*. El desempeño a nivel de predicción de los árboles puede ser dramáticamente mejorado

- Hasta el momento, los árboles que se están construyendo, padecen el inconveniente de presentar una varianza alta.
- **Bagging** o *Bootstrap aggregation* es un procedimiento de propósito general para reducir la varianza de un método de aprendizaje estadístico.
- Consiste en realizar B muestras **Bootstrap** de los datos de entrenamiento, se genera un árbol no podado por cada muestra bootstrap que se ajuste a la misma y se obtienen predicciones

$$\hat{f}^{*1}(x), \hat{f}^{*2}(x), \ldots, \hat{f}^{*B}(x).$$

- Estas predicciones se promedian para obtener un modelo de aprendizaje estadístico con baja varianza

$$\frac{1}{B}$$

- El número de árboles ajustados B no es un parámetro crítico en el contexto de bagging. en la práctica se promedian cientos o hasta miles de árboles con el fin de reducir la varianza. Simplemente se utiliza un valor de B suficientemente grande para que el error se estabilice.

- El número de árboles ajustados B no es un parámetro crítico en el contexto de bagging. en la práctica se promedian cientos o hasta miles de árboles con el fin de reducir la varianza. Simplemente se utiliza un valor de B suficientemente grande para que el error se estabilice.
- En el caso de árboles de clasificación con todos los árboles ajustados a la muestra bootstrap, la predicción bagging se obtiene al final con las predicciones mediante un *voto de mayoría.*

Estimación del error “out of bag”

- Sucede que hay una forma fácil de estimar el error de un modelo “bagged”, sin la necesidad de recurrir a una validación cruzada.
- Recuérdese que la clave en *bagging* consiste en que se generan árboles que se ajustan a muestras bootstrap de los datos de entrenamiento.

Estimación del error "out of bag"

- Sucede que hay una forma fácil de estimar el error de un modelo "bagged", sin la necesidad de recurrir a una validación cruzada.
- Recuérdese que la clave en *bagging* consiste en que se generan árboles que se ajustan a muestras bootstrap de los datos de entrenamiento.
- También sucede que en promedio cada árbol *bagged* utiliza aproximadamente solo 2/3 de las observaciones. A la tercera parte remanente de las observaciones se les denomina, observaciones "**out-of-bag**" (OBB).

- Esto producirá aproximadamente $B/3$ predicciones para la i-ésima observación. Con estas predicciones se puede calcular un MSE en el caso de árboles de regresión o el *error de clasificación* en el caso de los árboles de clasificación.

- Esto producirá aproximadamente $B/3$ predicciones para la i-ésima observación. Con estas predicciones se puede calcular un MSE en el caso de árboles de regresión o el *error de clasificación* en el caso de los árboles de clasificación.
- El abordaje OOB para estimar el error en el conjunto de prueba, es particularmente conveniente cuando se emplea *bagging* en grandes conjuntos de datos para los cuales una validación cruzada deviene computacionalmente intensa.

Ejercicio2 Chp5

Derive la probabilidad de que una observación dada sea parte de una muesta bootstrap. Suponga que se tiene una muestra bootstrap de un conjunto con n observaciones.

1. Cuál es la probabilidad de que la primera observación bootstrap no sea la j-ésima observación de la muestra original?.

Ejercicio2 Chp5

Derive la probabilidad de que una observación dada sea parte de una muesta bootstrap. Suponga que se tiene una muestra bootstrap de un conjunto con n observaciones.

1. Cuál es la probabilidad de que la primera observación bootstrap no sea la j-ésima observación de la muestra original?.
2. Cuál es la probabilidad de que la segunda observación bootstrap no sea la j-ésima observación de la muestra original?.

Ejercicio2 Chp5

Derive la probabilidad de que una observación dada sea parte de una muesta bootstrap. Suponga que se tiene una muestra bootstrap de un conjunto con n observaciones.

1. Cuál es la probabilidad de que la primera observación bootstrap no sea la j-ésima observación de la muestra original?.
2. Cuál es la probabilidad de que la segunda observación bootstrap no sea la j-ésima observación de la muestra original?.
3. Argumente que la probabilidad de que la j-ésima observación *no esté* en la muestra bootstrap es $(1 - 1/n)^n$.

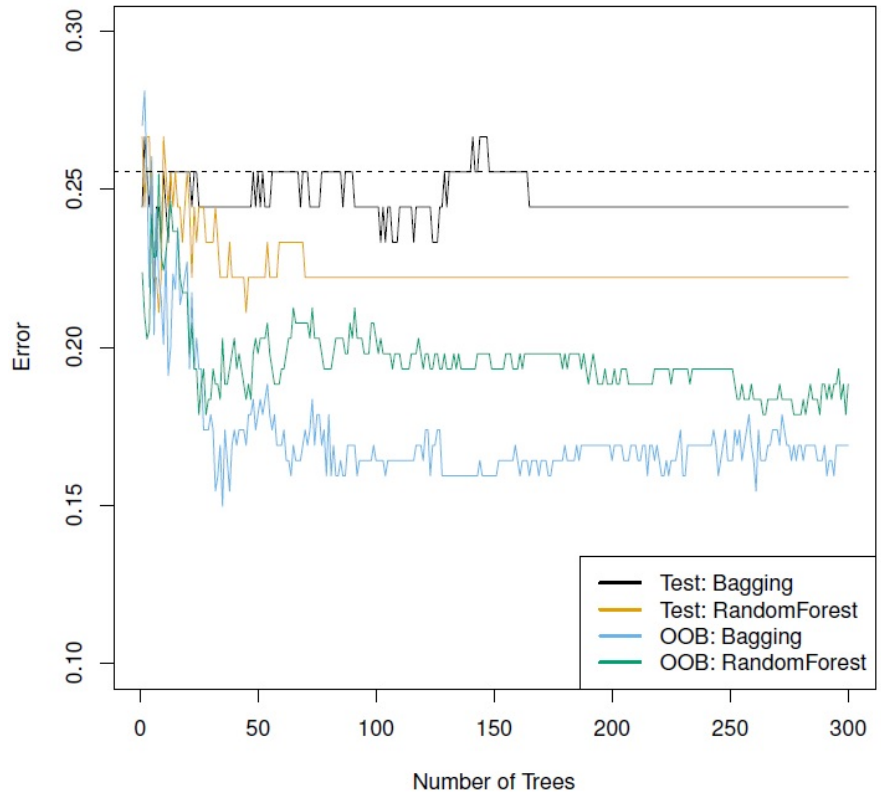

Figura 1: OOB (out-of-bag).

Midiendo la importancia de las variables

- Es difícil interpretar un modelo bagged. En el caso de árboles, por ejemplo, a pesar de que un árbol es interpretable, el modelo obtenido promediando la predicción de muchos árboles es más difícil de ser interpretada.

Midiendo la importancia de las variables

- Es difícil interpretar un modelo bagged. En el caso de árboles, por ejemplo, a pesar de que un árbol es interpretable, el modelo obtenido promediando la predicción de muchos árboles es más difícil de ser interpretada.
- Con el *bagging* se consigue un aumento drástico en la precisión de las predicciones a costas de una perdida de interpretabilidad.

Midiendo la importancia de las variables

- Es difícil interpretar un modelo bagged. En el caso de árboles, por ejemplo, a pesar de que un árbol es interpretable, el modelo obtenido promediando la predicción de muchos árboles es más difícil de ser interpretada.
- Con el *bagging* se consigue un aumento drástico en la precisión de las predicciones a costas de una perdida de interpretabilidad.
- Sin embargo en un modelo *bagged* se puede obtener un resumen de la importancia de cada variable. Por ejemplo registrando la cantidad en que decrece RSS debido a cada uno de los predictores en el caso de árboles de regresión y promediando sobre todos los árboles.

Midiendo la importancia de las variables

- Es difícil interpretar un modelo bagged. En el caso de árboles, por ejemplo, a pesar de que un árbol es interpretable, el modelo obtenido promediando la predicción de muchos árboles es más difícil de ser interpretada.
- Con el *bagging* se consigue un aumento drástico en la precisión de las predicciones a costas de una perdida de interpretabilidad.
- Sin embargo en un modelo *bagged* se puede obtener un resumen de la importancia de cada variable. Por ejemplo registrando la cantidad en que decrece RSS debido a cada uno de los predictores en el caso de árboles de regresión y promediando sobre todos los árboles.
- También en el caso de árboles de clasificación se puede registrar la cantidad aculada en que decrece el índice de *Gini* o la *entropía*.

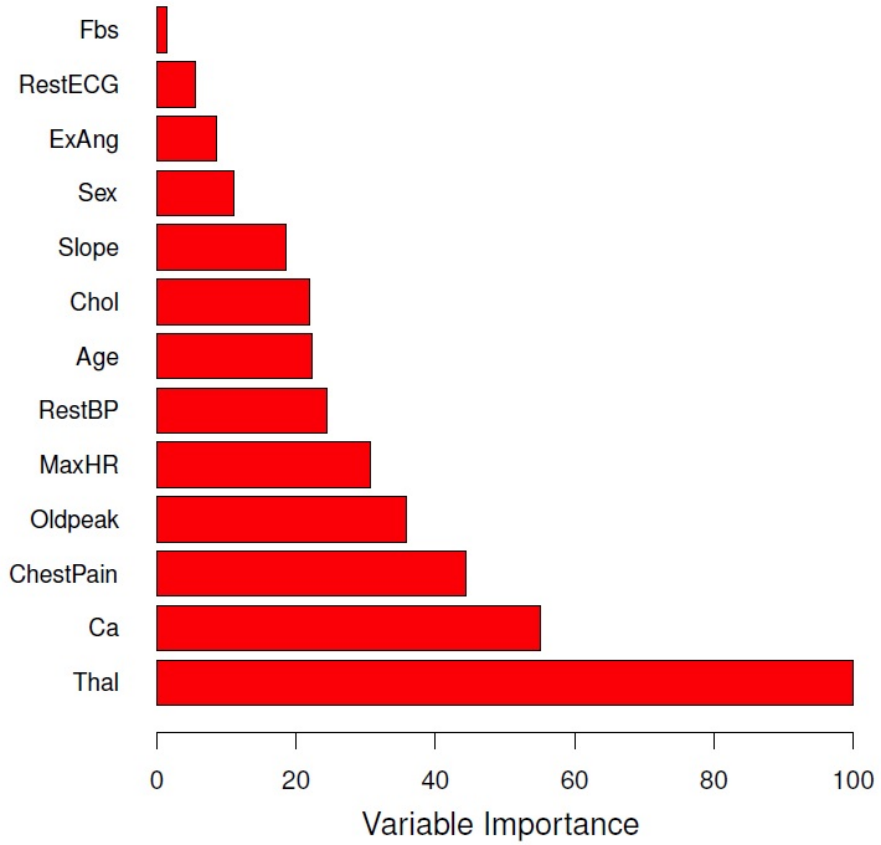
Fbs
RestECG
ExAng
Sex
Slope
Chol
Age
RestBP
MaxHR
Oldpeak
ChestPain
Ca
Thal
0
20
40
60
80
100
Variable Importance

Figura 2: .

Bosques Aleatórios (Random Forest)

- La metodología de bosques aleatorios proporciona una mejora sobre los árboles *bagged* que consiste una pequeña modificación que des correlaciona los árboles.

Bosques Aleatórios (Random Forest)

- La metodología de bosques aleatorios proporciona una mejora sobre los árboles *bagged* que consiste una pequeña modificación que des correlaciona los árboles.
- Como en *bagging* se construye un número de árboles de decisión en muestras bootstrap.

Bosques Aleatórios (Random Forest)

- La metodología de bosques aleatorios proporciona una mejora sobre los árboles *bagged* que consiste una pequeña modificación que des correlaciona los árboles.
- Como en *bagging* se construye un número de árboles de decisión en muestras bootstrap.
- Pero en el proceso de construcción de los árboles, cada vez que se considera el paso de seccionamiento solo se considera una muestra aleatoria de m variables del conjunto completo de p predictores.

$$\text{Típicamente} \quad m \approx \sqrt{p}.$$

- Es decir cuando se construye un árbol aleatorio, en cada bifurcación (split) del árbol, no se permite al algoritmo siquiera considerar una mayoría de los predictores disponibles.

- Es decir cuando se construye un árbol aleatorio, en cada bifurcación (split) del árbol, no se permite al algoritmo siquiera considerar una mayoría de los predictores disponibles.
- Esto puede sonar contra intuitivo, pero cuando hay uno o varios predictores dominantes, la colección de árboles *bagged* que se van construyendo tendrán a involucrar estos predictores y por lo tanto los arboles terminarán siendo muy correlacionados.

- Es decir cuando se construye un árbol aleatorio, en cada bifurcación (split) del árbol, no se permite al algoritmo siquiera considerar una mayoría de los predictores disponibles.
- Esto puede sonar contra intuitivo, pero cuando hay uno o varios predictores dominantes, la colección de árboles *bagged* que se van construyendo tendrán a involucrar estos predictores y por lo tanto los arboles terminarán siendo muy correlacionados.
- Así en promedio $(p - m)/p$ de las bifurcaciones en los árboles no considerarán uno de los predictores dominantes.
  Considerar solo una porción aleatoria de los predictores, ayuda a mitigar este efecto de árboles correlacionados y por lo tanto se puede disminuir más aún la varianza.

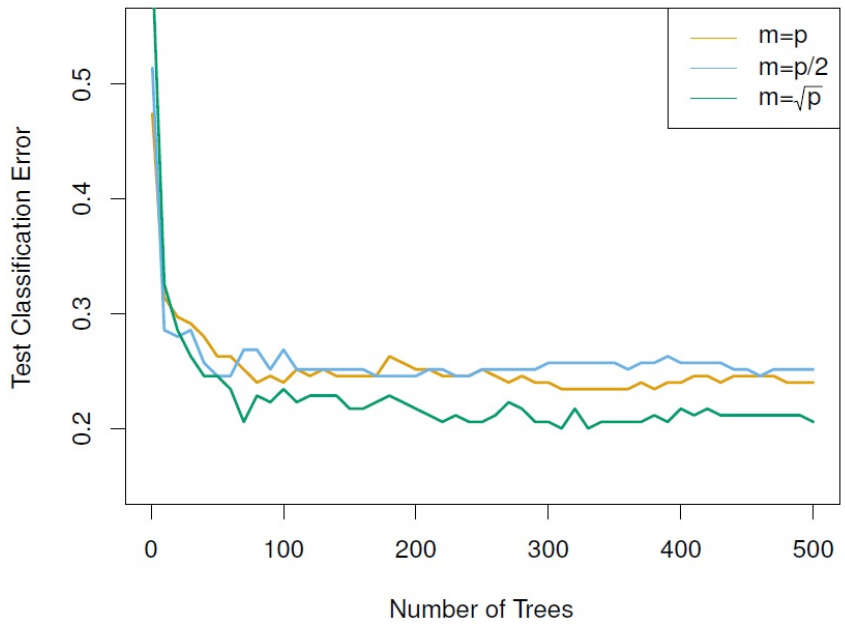

Figura 3: Resultados de la realización de bosques aleatorios en un conjunto de datos de expresión génica de 15 clases y $p = 500$ predictores para predecir cáncer versus normal. Los errores en el conjunto de prueba se muestran en función del número de árboles. Cada línea coloreada corresponde a diferentes valores de m. Bosque aleatorio ($m < p$) presenta una pequeña mejora sobre bagging ($m = p$). Un solo árbol de clasificación posee una tasa de error del 45.7 % .

Boosting

- **Boosting** es otra metodología para mejorar la predicción proporcionada por un árbol de decisión.
- Como Bagging, *Boosting* es un procedimiento general que puede ser aplicado a muchos métodos de aprendizaje estadístico para regresión y clasificación.

Boosting

- **Boosting** es otra metodología para mejorar la predicción proporcionada por un árbol de decisión.
- Como Bagging, *Boosting* es un procedimiento general que puede ser aplicado a muchos métodos de aprendizaje estadístico para regresión y clasificación.
- En *Boosting* en vez de ajustar un árbol de decisión grande, lo cual tiene el potencial de de sobre ajustar los datos, en forma secuencial se adicionan árboles donde cada uno es ajustado a los residuos del anterior. De forma más precisa véase el algoritmo.

Algoritmo Boosting para árboles de regresión.

1. Tómese $\hat{f}(x) = 0$ y $r_i = y_i$ para todo i en el conjunto de entrenamiento.

Algoritmo Boosting para árboles de regresión.

1. Tómese $\hat{f}(x) = 0$ y $r_i = y_i$ para todo i en el conjunto de entrenamiento.
2. Para $b = 1, 2, \ldots, B$ repetir:
   (a) Ajuste un árbol con d bifurcaciones ($d + 1$ nodos terminales) a los datos de entrenamiento (X, r).
   (b) Actualice $\hat{f}$ adicionando una versión "contraida" del árbol nuevo

$$\hat{f} \leftarrow \hat{f} + \lambda \hat{f}^b(x). \tag{2}$$

Algoritmo Boosting para árboles de regresión.

1. Tómese $\hat{f}(x) = 0$ y $r_i = y_i$ para todo i en el conjunto de entrenamiento.
2. Para $b = 1, 2, \ldots, B$ repetir:
   (a) Ajuste un árbol con d bifurcaciones ($d + 1$ nodos terminales) a los datos de entrenamiento (X, r).
   (b) Actualice $\hat{f}$ adicionando una versión "contraida" del árbol nuevo

$$\hat{f} \leftarrow \hat{f} + \lambda \hat{f}^b(x). \tag{2}$$

   (c) Actualise los residuos,

$$r_i \leftarrow r_i - \lambda \hat{f}^b(x_i). \tag{3}$$

Algoritmo Boosting para árboles de regresión.

1. Tómese $\hat{f}(x) = 0$ y $r_i = y_i$ para todo i en el conjunto de entrenamiento.
2. Para $b = 1, 2, \ldots, B$ repetir:
   (a) Ajuste un árbol con d bifurcaciones ($d + 1$ nodos terminales) a los datos de entrenamiento (X, r).
   (b) Actualice $\hat{f}$ adicionando una versión "contraida" del árbol nuevo

$$\hat{f} \leftarrow \hat{f} + \lambda \hat{f}^b(x). \tag{2}$$

   (c) Actualise los residuos,

$$r_i \leftarrow r_i - \lambda \hat{f}^b(x_i). \tag{3}$$

3. Obtenga finalmente el modelo *boosted*

$$\hat{f} = \sum_{b=1}^{B} \lambda \hat{f}^b(x). \tag{4}$$

Boosting posee tres parámetros:

1. El número de árboles B. A diferencia de *bagging* y *bosque aleatorio*, boosting puede puede causar sobre ajuste si el valor de B es muy grande, de todas maneras, el sobre ajuste si ocurre lo hace de manera lenta. Es posible utilizar validación cruzada para seleccionar el valor de B.

Boosting posee tres parámetros:

1. El número de árboles B. A diferencia de *bagging* y *bosque aleatorio*, boosting puede puede causar sobre ajuste si el valor de B es muy grande, de todas maneras, el sobre ajuste si ocurre lo hace de manera lenta. Es posible utilizar validación cruzada para seleccionar el valor de B.
2. El parámetro de *encogimiento* λ, que en general es un número positivo pequeño, controla la tasa a la cuál *boosting* aprende (tasa de aprendizaje). Valores típicos son 0.01 o 0.001 y la elección puede depender del problema. Un valor muy pequeño de λ puede requerir un valor muy grande de B para alcanzar un buen desempeño.

3. El número de bifurcaciones (splits) d, en cada árbol, que controla la complejidad del ensamble boosted. “A menudo $d = 1$ funciona bien, en cuyo caso cada árbol es una horqueta. En este caso el ensamble boosted está ajustando un modelo aditivo, devido a que cada termino solo envuelve una sola variable. de manera más general d representa la *profundidad de interacción* y controla el orden de la interacción en el modelo boosted, debido a que d splits envuelven a lo sumo d variables.

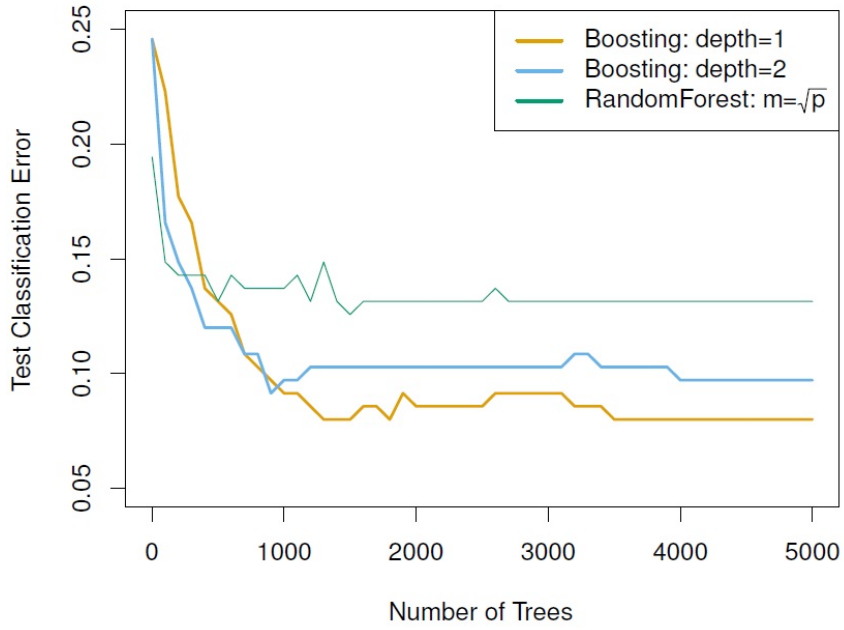

Figura 4: Resultados de la realización de bosques aleatorios en un conjunto de datos de expresión génica de 15 clases para predecir cáncer versus normal. El error de prueba es mostrado en función del número de árboles. Para 2 modelos boosted, $\lambda = 0.01$ con profundidad de 1, se desempeña ligeramente mejor que árboles de profundidad 2, y ambos a su vez poseen mejor desempeño que un bosque aleatorio, sin embargo los errores estándar son alrededor de 0.02, haciendo que ninguna de estas diferencias sea significativa. La tasa del error de prueba es del 24 % para un solo árbol .