Fundamentos de la Programación Estadística y Data Mining en R
Unidad 4 Árboles de Decisión y Random Forests

Dr. Germán Rosati (Digital House/MTEySS/UNSAM)

Bagging de árboles de decisión

- Es posible aplicar el algoritmo de bagging para construir un ensamble de árboles de decisión (Breiman 1990)
- Procedimiento
  - Extraer B muestras con reposición del dataset
  - Para cada una de las muestras entrenar un árbol de clasificación
  - Con cada una de los B árboles construidos hacer una predicción sobre el test set
  - Agregar por «mayoría de votos» las predicciones para generar la predicción final

Bagging: árboles de decisión

- Más formalmente:
  - Un TrSet consiste en datos $\{(y_n, X_n), n = 1, \ldots, N\}$, donde y es una variable respuesta cuantitativa o cualitativa; Asumamos que tenemos un procedimiento para formar un predictor con ese TrSet $\varphi(X, TrS)$
  - Asumamos que tenemos, además, una secuencia de TrS_k cada uno de los cuales consiste en N observaciones independientes de la misma distribución subyacente a TrS
  - La idea es usar TrS_k para obtener un mejor predictor que un simple $\varphi(X, TrS)$

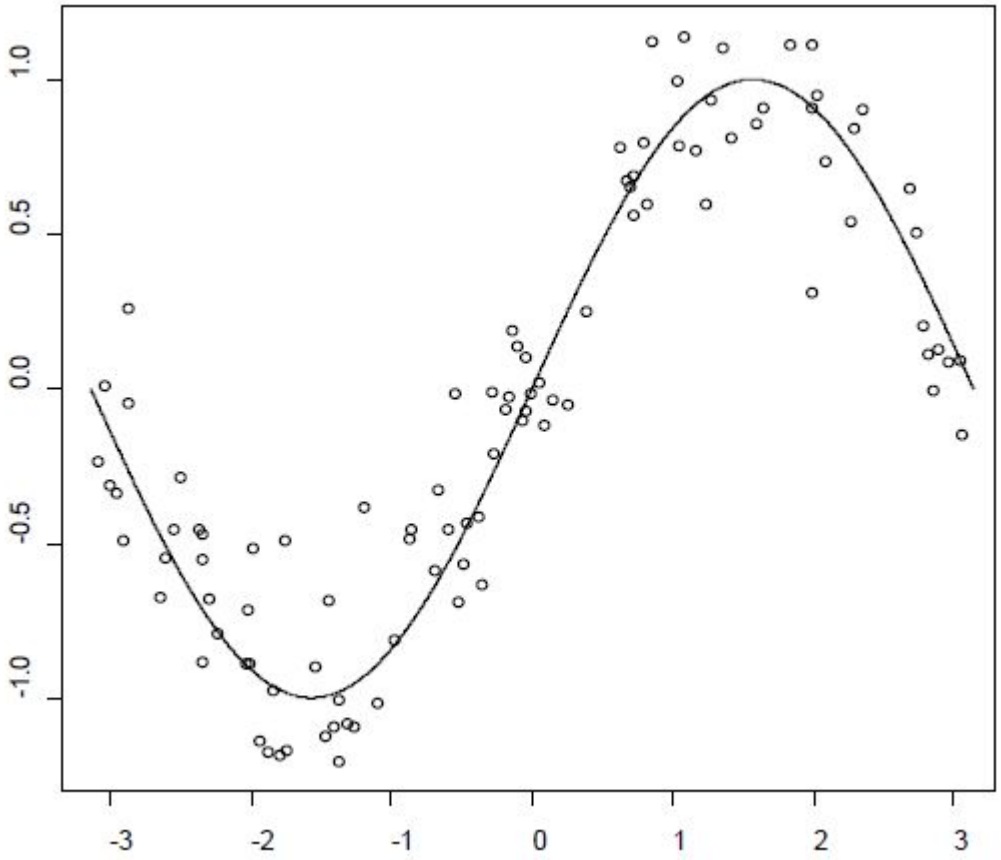
Bagging: árboles de decisión

- Si y es numérica, un procedimiento obvio es $\varphi_A(X) = E(\varphi(X, TrS_k))$
- Si y es cualitativa, la forma de agregar los predictores puede ser por alguna forma de "voto mayoritario"
- En bagging obtenemos réplicas con reposición del método dataset original

- Factor crítico: para que bagging funcione bien es necesario que los $\varphi(X, TrS)$ sean "inestables". Es decir que pequeños cambios en TrS generen cambios grandes en $\varphi(X, TrS)$. Ejemplos: árboles de decisión, redes neuronales, best subset regression.

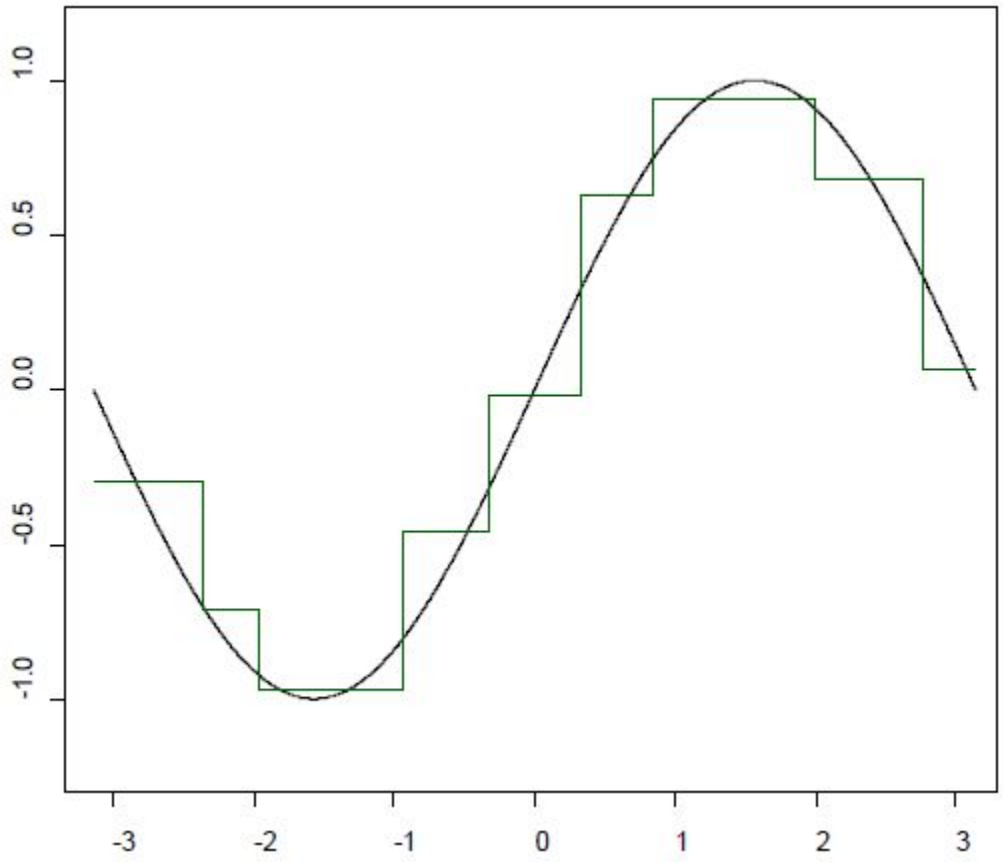
Bagging árboles: ejemplo

- Función generadora



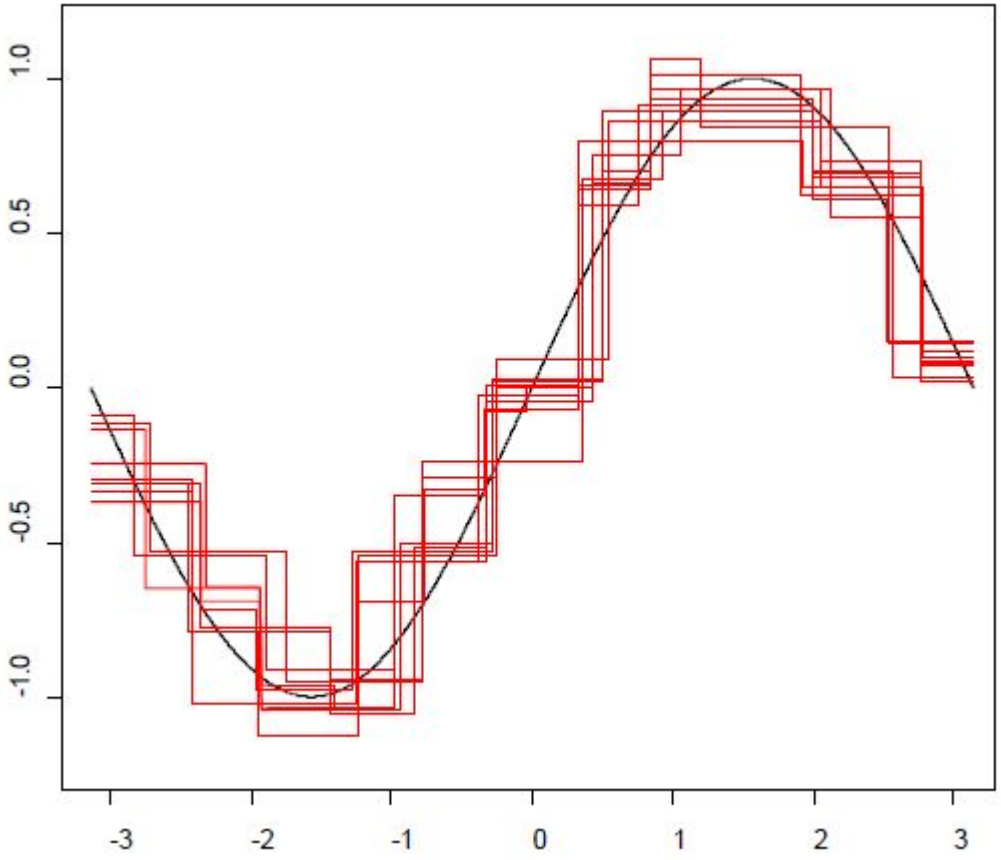
Bagging árboles: ejemplo

- Un solo árbol



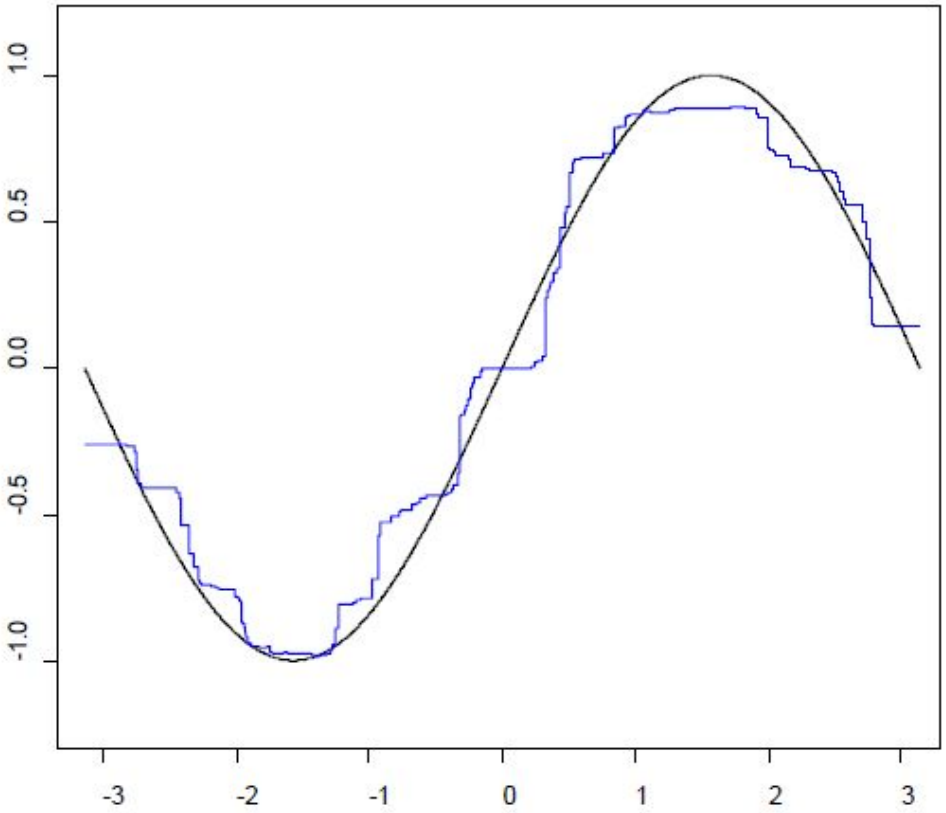
Bagging árboles: ejemplo

- 10 árboles



Bagging árboles: ejemplo

- Promedio de 100 árboles



Bagging árboles: importancia

- ¿Cómo medir la importancia de las variables a través de un modelo bagging de árboles de decisión?
- Problema de un bagging: dificultad en la interpretación.
- Un árbol «simple» era un modelo claramente interpretable y gráfico.
- Si construyo 2000 árboles de decisión… ¿cúal tiene sentido graficar para avanzar en la interpretación de la importancia de los predictores?

Bagging árboles: importancia

- Es necesario construir medidas agregadas que sirvan como resumen de la importancia de cada una de las variables a lo largo de todos los árboles construidos
  - RSS (suma de cuadrados de los residuos) en árboles de regresión: podemos calcular la cantidad en que el RSS se «achica» debido a los splits a lo largo de un predictor dado a lo largo de todos los B árboles. Valores grandes significan un predictor relevante
  - Gini en árboles de clasificación: parecido al anterior. Podemos agregar el valor total en que el índice de Gini es «achicado» por todos los splits en los que se usa un predictor en todos los árboles

Random forest

- Random forest es muy similar a un bagging de árboles de decisión
- La diferencia: además de generar variabilidad sobre los registros, se genera variabilidad sobre los predictores.
- En bagging, genero B predicciones a partir de
  - B remuestras bootstrap del TrS original y de
  - M predictores del TrS original
- De esta forma, en bagging entran el total de los M predictores.

Random forest

- Esto genera que los B árboles estén muy correlacionados. ¿Por qué?
- Random forest logra “descorrelacionar” los árboles a través de esta pequeño ajuste: muestrear (permutar) un subconjunto (m) de los M predictores.
- De esta forma, cada vez que el algoritmo evalúa un split de un árbol, solamente considera una muestra aleatoria simple de m E M. Solamente considera las m variables elegidas y selecciona la que mejor split genera.

Random forest

- Supongamos que hay un predictor muy “fuerte” y muchos otros “moderados”.
- Al hacer bagging, probablemente este predictor fuerte entre en todos los árboles (y seguramente, además, en el primer split).
- En consecuencia, todos los árboles van a ser “parecidos”, es decir, correlacionados. Lo mismo va a pasar con las predicciones.
- Al promediar muchos árboles muy correlacionadas, no se va a observar una ganancia en la reducción de la variancia del ensamble.

Random forest

- Random forests, lo soluciona obligando a considerar solamente un subset de predictores.
- Entonces, en promedio $(p - m)/p$ de todos los splits no van a considerar ese predictor fuerte y todos los otros predictores van a pesar más. Esto es "descorrelacionar" los árboles.
- En general, usar un valor pequeño de m debería ayudar cuando tenemos muchos predictores muy correlacionados.

Random forest

- Cada vez que el árbol genera un split nuevo, realiza una nueva selección de m variables.
- ¿Cuántas variables hay que seleccionar en cada split? Es decir, ¿cuál es valor de m?
  - Una opción $\mathrm{m} = \sqrt{M}$
  - Otra opción: seleccionarlo a través de cross-validation

Random forest - Interpretación

- ¿Cómo interpretar un clasificador random forest?
  - Importancia de las variables (decrecimiento en el MSE o en el error de clasificación de cada una de las variables a lo largo de todo el ensamble - ver más arriba)
- Otra opción (complementaria a la anterior) es comenzar a pensar en la “dependencia parcial”. Es decir, ¿cuál es la relación entre el predictor y la variable de respuesta?

Random forest – Dependencia parcial

- En un modelo de regresión lineal
  - $Y_i = \alpha_i + \sum_{k=1}^{K} \beta_k X_{i,k} + e_i$
- los efectos marginales de las variables independientes son los P β_p de la regresión. Cada uno es el "impacto" sobre la variable dependiente cuando la variable independiente varía en una unidad.
- El impacto es constante: independientemente del valor de X el impacto sobre Y es el mismo.

Random forest – Dependencia parcial

- En un modelo de regresión logística
  - $logit(Y_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$

- los efectos marginales de las variables independientes son los P β_p de la regresión. Pero su interpretación es más compleja porque la relación no es lineal. El impacto de cada X sobre Y es función de los valores de X.

Random forest – Dependencia parcial

- Idea básica: obtener una predicción para cada valor único de X_p manteniendo constante el efecto del resto de las X_p.
- Luego, graficar las predicciones para cada valor de X_p. Esto debería brindar una aproximación a la forma funcional de la relación entre X_p e Y
- Dado que random forest puede aproximar casi cualquier relación entre X e Y resulta útil para detectar relaciones no lineales sin necesidad de parametrizarlas previamente

Random forest – Dependencia parcial

- Para valor de la variable de interés se crea un nuevo dataset en el cual a todas las observaciones se asigna el mismo valor en la variable de interés
- Este dataset es “dropped-down” en el bosque y una predicción para cada dataset es obtenida.
- Se promedian las observaciones y se obtiene una predicción final

Random forest – Dependencia parcial

- Sean:
  - x_j: la variable predictora de interés
  - $\mathbf{X}_j$: la matriz del resto de las predictoras
  - y: la variable dependiente (cuantitativa)
  - $\hat{f}(\mathbf{X})$: el "bosque" construido
- Para x_j ordenar los valores únicos $V = \{x_j\}, i \in \{1, \dots n\}$, resultando en V^*, donde $|V^*| = K$
- Crear ***K*** nuevas matrices $\mathbf{X}_k = \left(\left[x_j = V^*\right], \mathbf{X}_j\right)$
- Para cada matriz ***K***, "correr" el bosque $\hat{f}(\mathbf{X})$ y obtener una predicción de $\hat{y}_k^*$
- Promediar los resultados de cada $\hat{y}_k^*$: $\hat{y}^* = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_k^*$
- Graficar V^* contra $\hat{y}^*$

Random forest – Dependencia parcial

- El resultado es “más” que los efectos marginales.
- Cada predicción se hace no “manteniendo constantes” los otros predictores sino que se usa toda la información disponible del resto de los predictores.
- Esto quiere decir que la relación graficada contiene todas las relaciones entre x_j e y, incluyendo los efectos “medios” de todas las interacciones de xj con el resto de los predictores $\mathbf{X}_j$. Por eso, se lo llama “dependencia parcial” y no “dependencia marginal”.

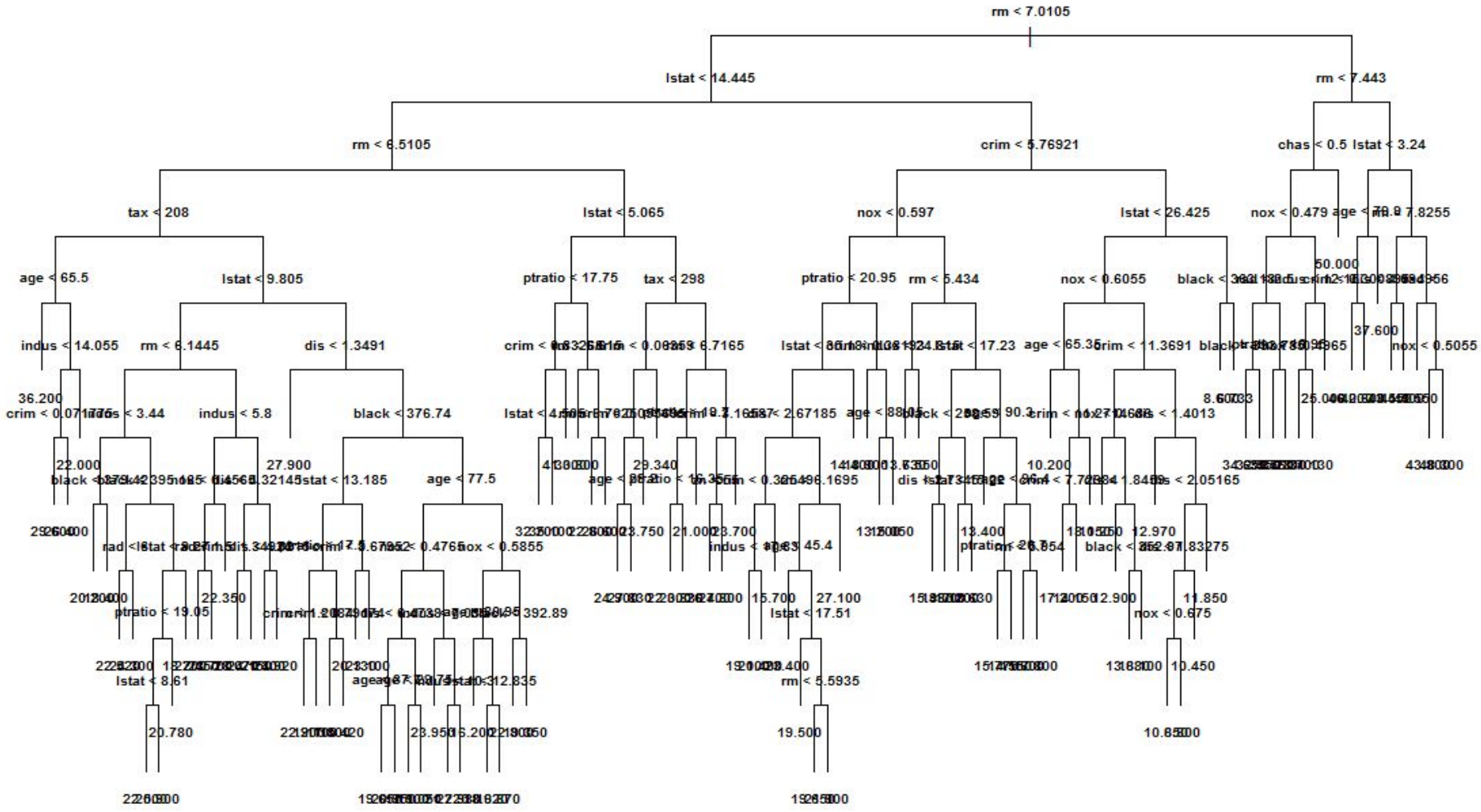
Random forest

- Ventajas según Breiman (2000):
  - Buena capacidad predictiva (comparable a boosting)
  - Robusto relativamente a ruido y a outliers
  - Más rápido que bagging y boosting
  - Da estimaciones de error e importancia de variables
- Desventajas
  - Funciona mejor en clasificación que en regresión

Random Forest - Ejemplo

- Datos: Boston dataset (ISLR package)
- Objetivo: predecir el precio mediano de las viviendas en diversos condados de Boston
  - 1. CRIM: per capita crime rate by town
    2. ZN: proportion of residential land zoned for lots over 25,000 sq.ft.
    3. INDUS: proportion of non-retail business acres per town
    4. CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
    5. NOX: nitric oxides concentration (parts per 10 million)
    6. RM: average number of rooms per dwelling
    7. AGE: proportion of owner-occupied units built prior to 1940
    8. DIS: weighted distances to five Boston employment centres
    9. RAD: index of accessibility to radial highways
    10. TAX: full-value property-tax rate per $10,000
    11. PTRATIO: pupil-teacher ratio by town
    12. B: 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
    13. LSTAT: % lower status of the population
    14. MEDV: Median value of owner-occupied homes in $1000's

Random Forest – Ej: árbol máximo



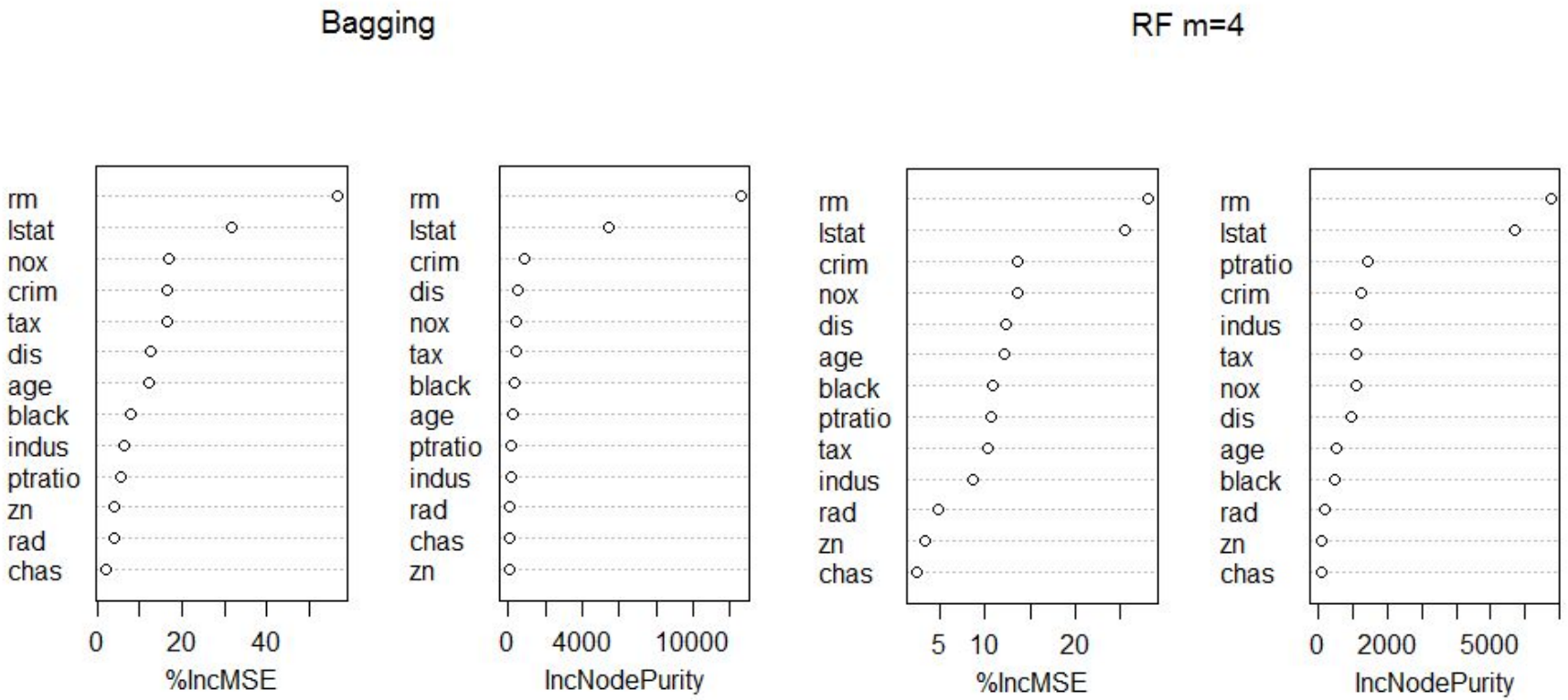
Comparación Random Forest y bagg

- Se estimaron un modelo bagging y 12 modelos random forest variando el parámetro mtry (es decir, cantidad de predictores que se muestrean en cada split).
- En cada modelo se entrenaron 500 árboles
- Se dividió el dataset en 50% como TrS y 50% como TeS

Comparación Random Forest y bagg

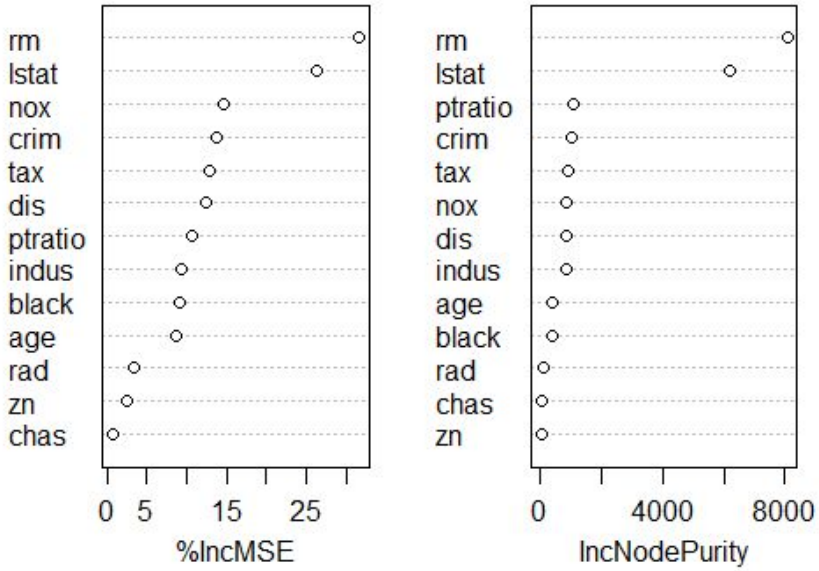
	OOB Error	Test Error
Bagging m=13	11.04	21.84
R. Forest m=1	19.08	27.3
R. Forest m=2	13.3	20.58
R. Forest m=3	11.63	18.89
R. Forest m=4	11.05	18.35
R. Forest m=5	10.64	18.32
R. Forest m=6	10.63	18.00
R. Forest m=7	10-42	18.63
R. Forest m=8	10.38	19.34
R. Forest m=9	10.55	19.56
R. Forest m=10	10.62	20.15
R. Forest m=11	10.84	21.08
R. Forest m=12	10.95	21.7

Random Forest - Ejemplo

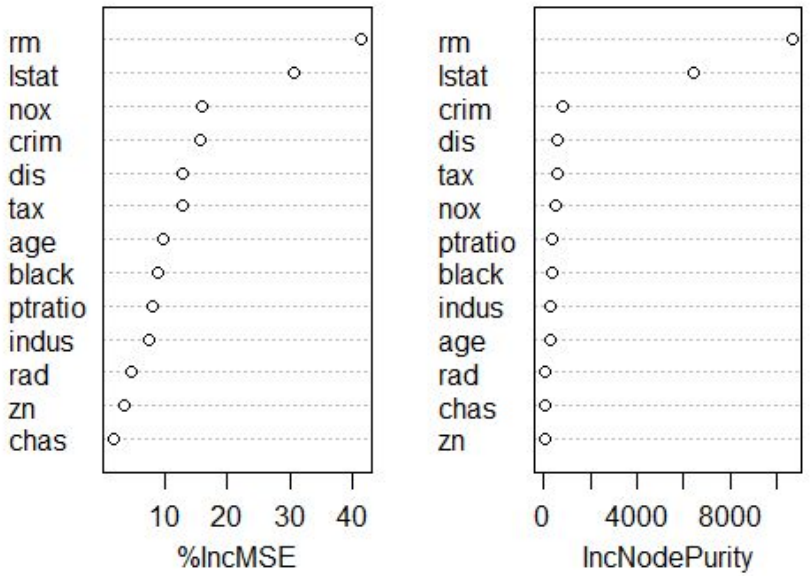


Random Forest - Ejemplo

RF m=6

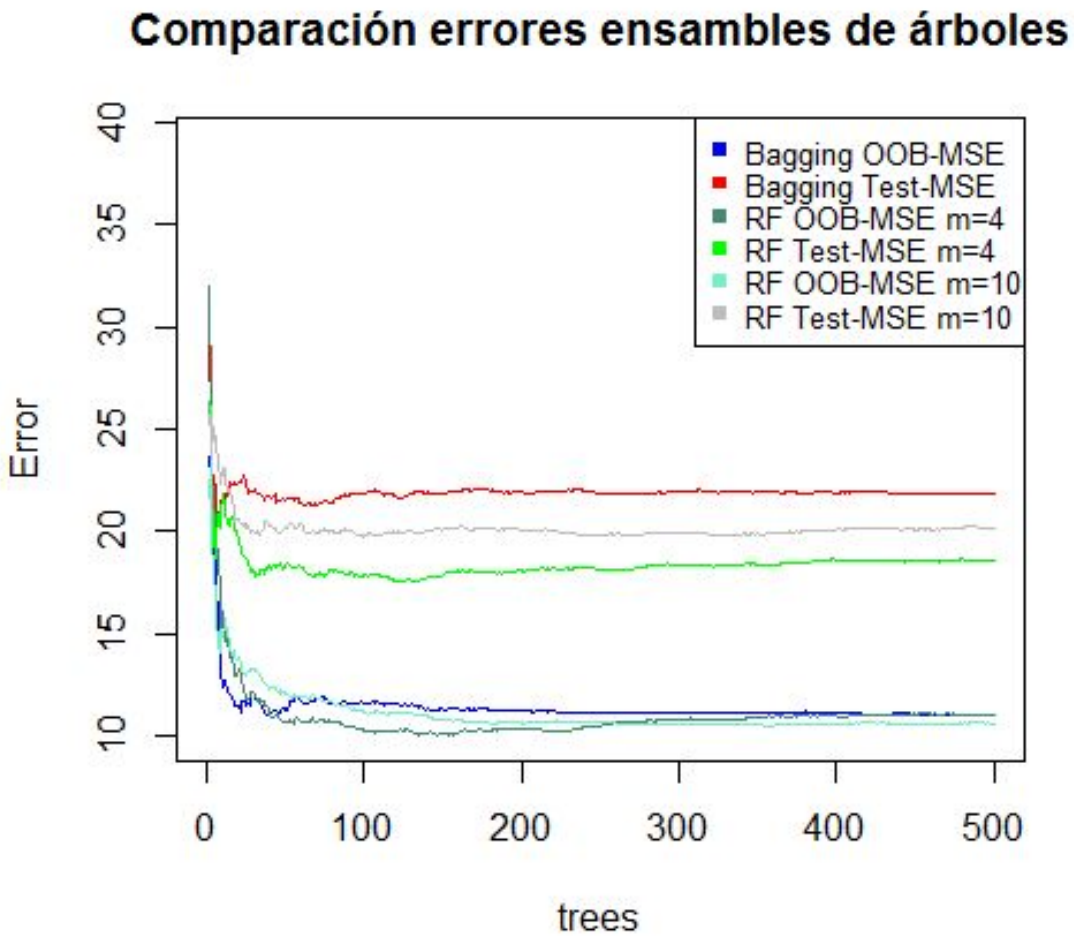


RF m=10



Random Forest - Ejemplo

- A partir de determinado tamaño de árboles el modelo no parece sobrefittear
- RF es el que menor error en el test set muestra



Random Forest - Ejemplo

Gráficos de dependencia parcial de algunas variables