T4 Árboles, bosques, bagging y boosting

Índice

1 CART

Classification and regression trees (CART): árboles que particionan el espacio de entrada recursivamente hasta alcanzar las hojas, cada una de ellas caracterizada por la región en la que se aplica y su (modelo de) predicción correspondiente

- Dada una entrada $\boldsymbol{x} \in \mathbb{R}^D$, un árbol puede verse como un conjunto de reglas de decisión anidadas hasta alcanzar las hojas
- Cada regla de decisión o **nodo interno** i define un **split paralelo a un eje:** compara una característica d_i de la entrada con un umbral t_i y, si $\boldsymbol{x}_{d_i} \leq t_i$, $\boldsymbol{x}$ se sigue procesando por la rama izquierda; si no, se procesa por la derecha
- El procesamiento de $\boldsymbol{x}$ termina al llegar a un nodo hoja, donde se especifica la salida predicha para toda entrada dentro de su región asociada, esto es, la región acorde con los splits definidos en sus nodos antecesores

1.1 Árbol de regresión

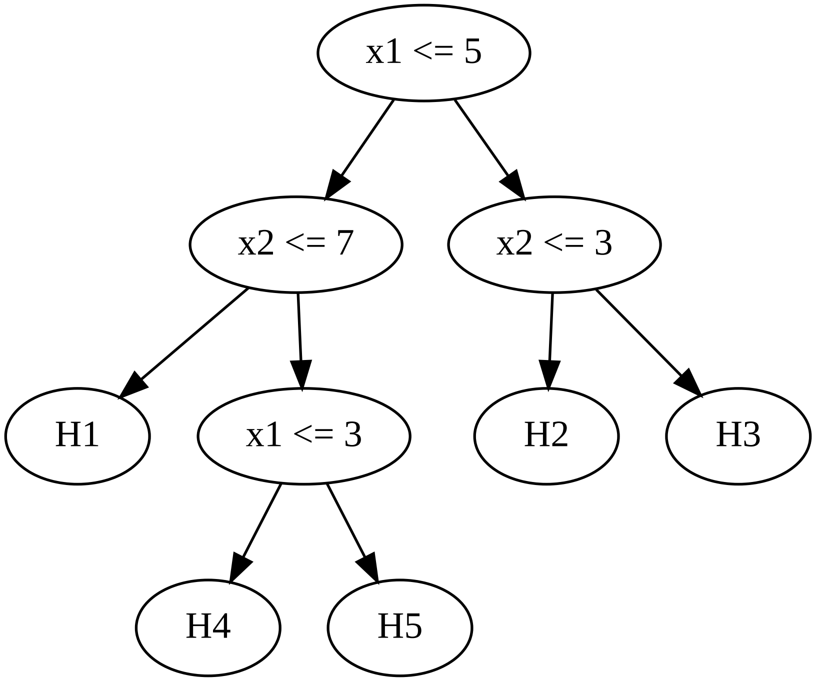
Árbol de regresión: $\boldsymbol{\theta} = \{(R_j, w_j) : j = 1 : J\}$, de J hojas, donde R_j y w_j denotan la región y salida asociadas a la hoja j

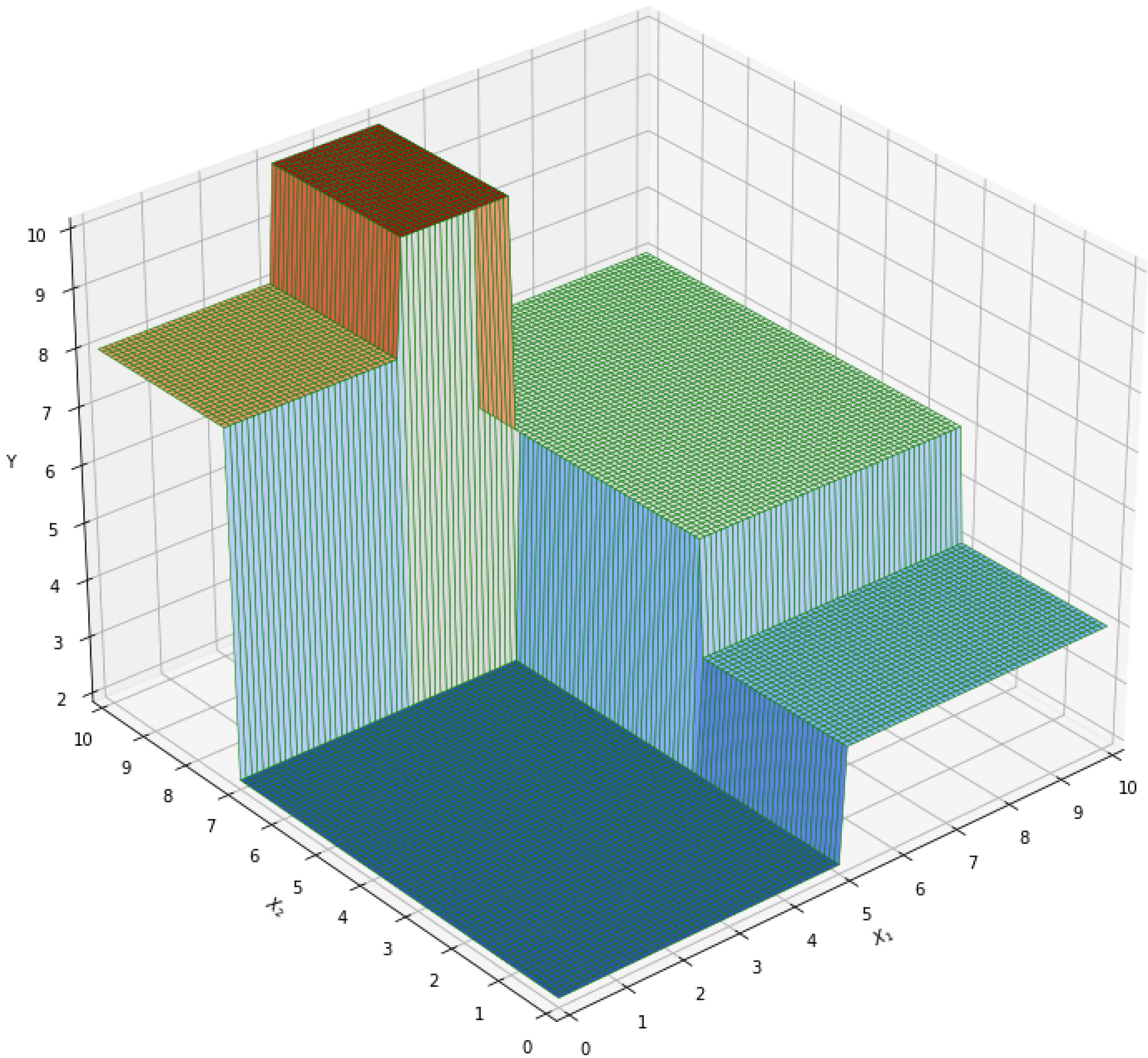
$$f(\boldsymbol{x}; \boldsymbol{\theta}) = \sum_{j=1}^{J} w_j \, \mathbb{I}(\boldsymbol{x} \in R_j)$$

Aprendizaje de las salidas: se suele usar la media de las salidas de los datos de entrenamiento en cada hoja

Ejemplo: árbol de regresión para entradas 2d con 4 nodos internos y 5 hojas

Out[1]:




Y
10
9
8
7
6
5
4
3
2
X_2
0
1
2
3
4
5
6
7
8
9
10
X_1
0
1
2
3
4
5
6
7
8
9
10

1.2 Árbol de clasificación

Árbol de clasificación: las hojas contienen una distribución sobre las etiquetas de clase en lugar de una respuesta promedio

1.2 Ajuste

Ajuste de un árbol de J hojas: minimización en $\boldsymbol{\theta} = \{(R_j, w_j) : j = 1 : J\}$ de la pérdida empírica

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_n \ell(y_n, f(\boldsymbol{x}; \boldsymbol{\theta})) = \sum_j \sum_{\boldsymbol{x}_n \in R_j} \ell(y_n, w_j)$$

Dificultad: la pérdida no es diferenciable pues debemos aprender la estructura del árbol y encontrar una óptima es un problema NP-duro

Soluciones aproximadas: métodos voraces como **CART, C4.5 o ID3,** que hacen crecer el árbol añadiendo un nodo en cada iteración

1.2.1 Splits de un nodo

Splits de un nodo i con base en una característica j: a partir de los datos que alcanzan i, $\mathcal{D}_i = \{(\boldsymbol{x}_n, y_n)\}$

- Si j es real, podemos definir un split por cada umbral $t \in \mathcal{T}_j$, con $\mathcal{T}_j = \{x_{nj}\}$ por ejemplo, tal que:

$$\begin{aligned}\mathcal{D}_i^L(j,t) &= \{(\boldsymbol{x}_n, y_n) \in \mathcal{D}_i : x_{nj} \leq t\}\\ \mathcal{D}_i^R(j,t) &= \{(\boldsymbol{x}_n, y_n) \in \mathcal{D}_i : x_{nj} > t\}\end{aligned}$$

- Si j es categórica, podemos definir un split por cada t igual a K_j valores posibles

$$\begin{aligned}\mathcal{D}_i^L(j,t) &= \{(\boldsymbol{x}_n, y_n) \in \mathcal{D}_i : x_{nj} = t\}\\ \mathcal{D}_i^R(j,t) &= \{(\boldsymbol{x}_n, y_n) \in \mathcal{D}_i : x_{nj} \neq t\}\end{aligned}$$

1.2.2 Elección del mejor split

Coste o impureza de un nodo: definimos una función $c(\cdot)$ para evaluarla, independiente del tamaño del nodo

Mejor split de un nodo i: uno que reduzca al máximo su coste, esto es, que minimice la suma de los costes normalizados de los hijos

$$(j_i, t_i) = \arg \min_{j\in\{1,\dots,D\}} \min_{t\in\mathcal{T}_j} \frac{|\mathcal{D}_i^L(j,t)|}{|\mathcal{D}_i|}\, c(\mathcal{D}_i^L(j,t)) + \frac{|\mathcal{D}_i^R(j,t)|}{|\mathcal{D}_i|}\, c(\mathcal{D}_i^R(j,t))$$

Coste de un nodo en regresión: se suele usar el **error cuadrático medio** (con respecto a la respuesta media del nodo)

$$c(\mathcal{D}_i) = \frac{1}{|\mathcal{D}_i|} \sum_{(\boldsymbol{x}_n, y_n)\in\mathcal{D}_i} (y_n - \bar{y})^2 \qquad \text{con} \qquad \bar{y} = \frac{1}{|\mathcal{D}_i|} \sum_{(\boldsymbol{x}_n, y_n)\in\mathcal{D}_i} y_n$$

Coste de un nodo en clasificación: las funciones usuales se basan en la distribución empírica de las clases en el nodo

$$\hat{\pi}_{ic} = \frac{1}{|\mathcal{D}_i|} \sum_{(\boldsymbol{x}_n, y_n)\in\mathcal{D}_i} \mathbb{I}(y_n = c)$$

Índice Gini: mide el error de clasificación esperado en el nodo i; esto es, la probabilidad de que un dato al azar se clasifique mal si su clase se determina aleatoriamente según las probabilidades de las clases

$$G_i = \sum_c \hat{\pi}_{ic}(1-\hat{\pi}_{ic}) = \sum_c \hat{\pi}_{ic} - \sum_c \hat{\pi}_{ic}^2 = 1 - \sum_c \hat{\pi}_{ic}^2$$

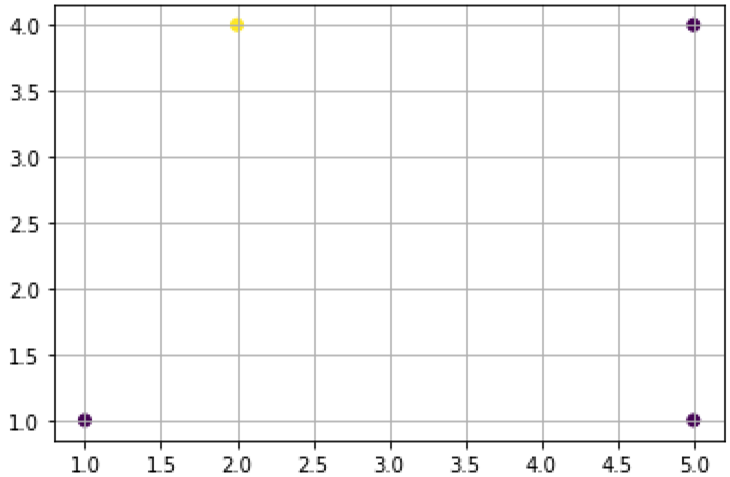
Entropía: de la distribución empírica de las clases en el nodo

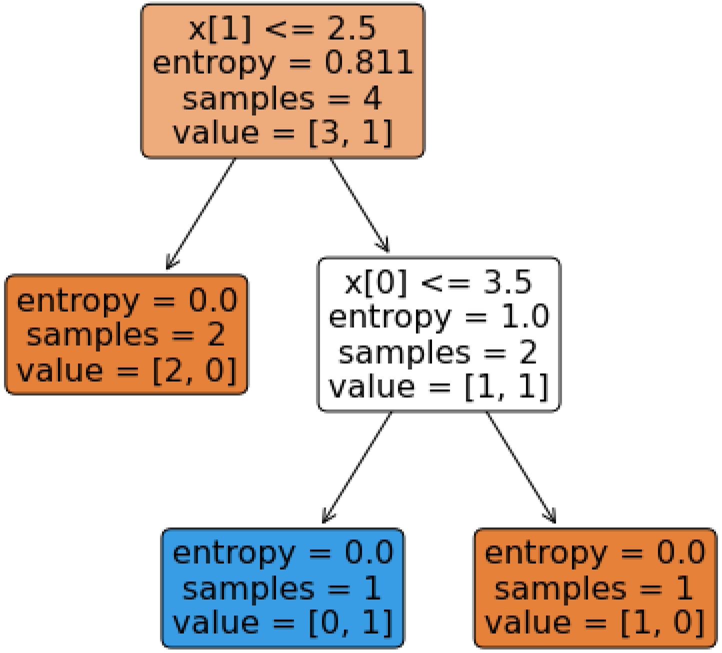
$$H_i = \mathbb{H}(\hat{\boldsymbol{\pi}}_i) = -\sum_c \hat{\pi}_{ic} \log \hat{\pi}_{ic}$$

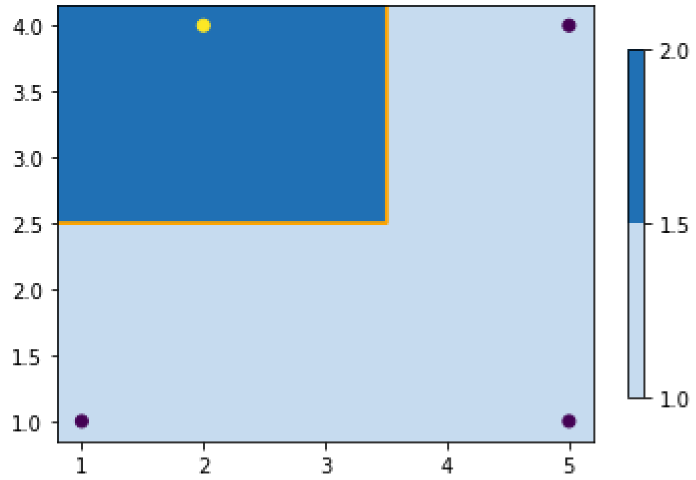
- **Mínima entropía:** $\hat{\pi}_{ic} = \delta(c = c^*) \;\Rightarrow\; H_i = -1\log 1 - \sum_{c\neq c^*} 0\,\log 0 = 0$

- **Máxima entropía:** $\hat{\pi}_{ic} = 1/C \;\Rightarrow\; H_i = -\sum_{c=1}^{C} \frac{1}{C} \log \frac{1}{C} = -\log \frac{1}{C} = \log C$

Ejemplo: $\mathcal{D}_i = \{((1,1)^t, 1), ((2,4)^t, 2), ((5,1)^t, 1), ((5,4)^t, 1)\}$




x[1] <= 2.5
entropy = 0.811
samples = 4
value = [3, 1]
entropy = 0.0
samples = 2
value = [2, 0]
x[0] <= 3.5
entropy = 1.0
samples = 2
value = [1, 1]
entropy = 0.0
samples = 1
value = [0, 1]
entropy = 0.0
samples = 1
value = [1, 0]


4.0
3.5
3.0
2.5
2.0
1.5
1.0
1
2
3
4
5
2.0
1.5
1.0

1.3 Regularización

Sobre-entrenamiento: si dejamos que un árbol crezca incontroladamente, podemos ajustarlo de manera que no cometa ningún error en entrenamiento (salvo ruido de etiquetas), pero funcione mal con datos futuros

Regularización: se suelen aplicar técnicas que limitan el tamaño del árbol

Aproximación directa: parar el crecimiento al tener pocos ejemplos en un nodo o alcanzar una profundidad máxima

Aproximación alternativa: dejar crecer el árbol al máximo y luego podarlo de hijos a padres mediante fusión de hijos

1.4 Características perdidas

Ventaja: a diferencia de otros modelos discriminativos, el manejo de datos con características perdidas es sencillo con árboles

Splits subrogados: heurístico estándar que, en caso de pérdida de una variable en inferencia, emplea variables de reserva que inducen particiones similares a las que induce la variable perdida

Variables categóricas perdidas: se añade un nuevo valor "perdido" y los datos se tratan como complemente observados

1.5 Ventajas e incovenientes

1.5.1 Ventajas

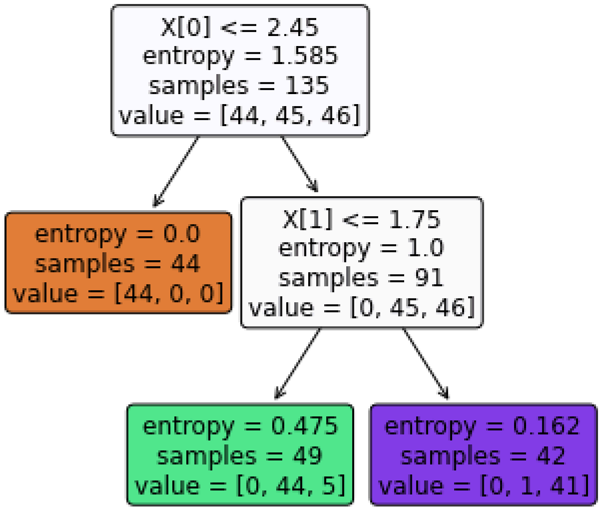
- Son fáciles de interpretar
- Manejan fácilmente entradas mixtas, discretas y continuas
- Son insensibles a transformaciones monótonas de las entradas ya que los puntos de split se basan en la ordenación de los datos, por lo que no es necesario estandarizarlos

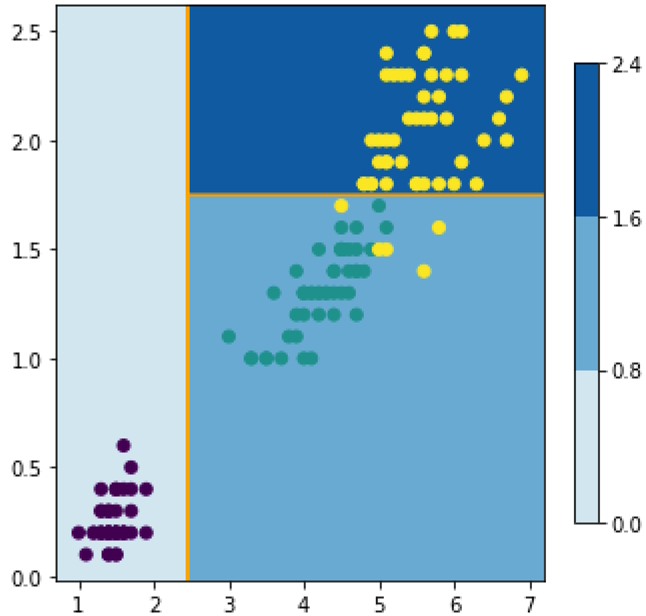
- Realizan selección de variables automáticamente
- De ajuste rápido y fácil escalado a grandes conjuntos de datos
- Pueden manejar características de entrada perdidas

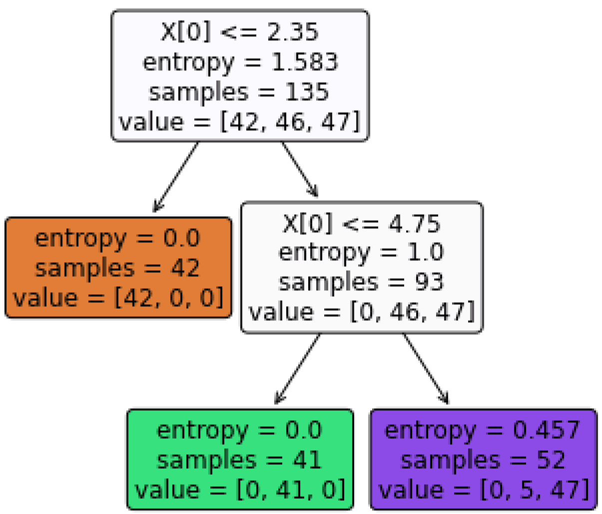
1.5.2 Incovenientes

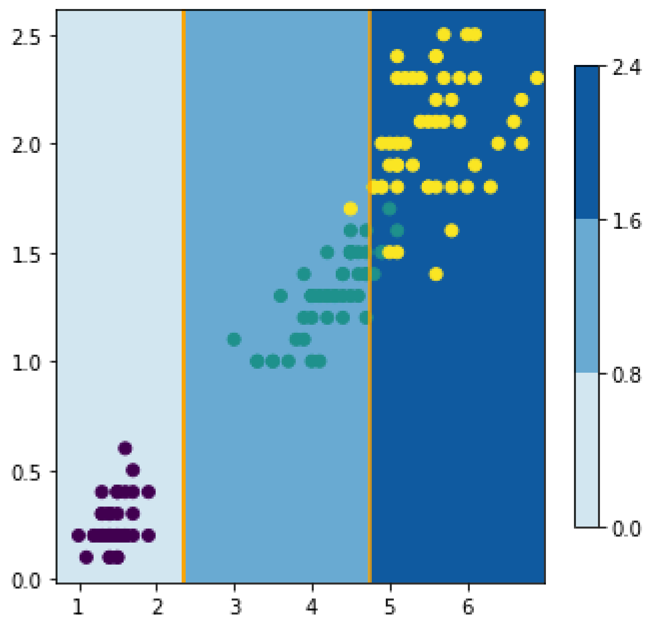
- No son muy precisos, debido en parte a su construcción voraz
- **Inestabilidad:** pequeños cambios en los datos de entrada pueden tener grandes consecuencias en la estructura del árbol

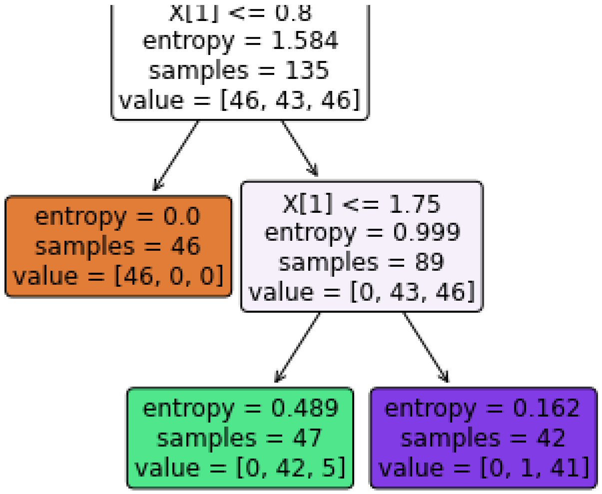
Ejemplo: inestabilidad con iris

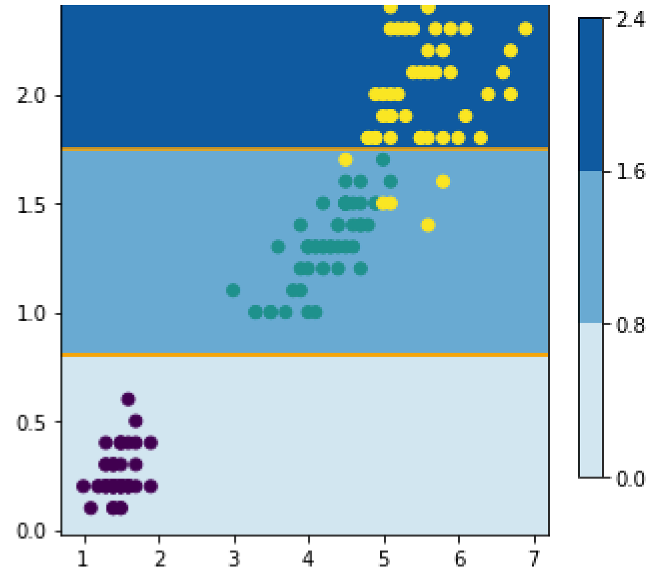

X[0] <= 2.45
entropy = 1.585
samples = 135
value = [44, 45, 46]
entropy = 0.0
samples = 44
value = [44, 0, 0]
X[1] <= 1.75
entropy = 1.0
samples = 91
value = [0, 45, 46]
entropy = 0.475
samples = 49
value = [0, 44, 5]
entropy = 0.162
samples = 42
value = [0, 1, 41]


2.4
1.6
0.8
0.0


X[0] <= 2.35
entropy = 1.583
samples = 135
value = [42, 46, 47]
entropy = 0.0
samples = 42
value = [42, 0, 0]
X[0] <= 4.75
entropy = 1.0
samples = 93
value = [0, 46, 47]
entropy = 0.0
samples = 41
value = [0, 41, 0]
entropy = 0.457
samples = 52
value = [0, 5, 47]


2.4
1.6
0.8
0.0





2 Aprendizaje de ensambles

Los árboles constituyen un estimador de alta varianza: pequeñas perturbaciones de los datos resultan en predicciones muy distintas

2.1 Aprendizaje de ensambles

Aprendizaje de ensambles: reduce la varianza de los árboles promediando M modelos base $\{f_m\}$

$$f(y \mid \boldsymbol{x}) = \frac{1}{M} \sum_{m=1}^{M} f_m(y \mid \boldsymbol{x})$$

Ensamble en regresión: estimador de sesgo similar al de los modelos base pero, en general, de mejor precisión por la menor varianza

Ensamble en clasificación: la salida se decide por el **método comité,** esto es, por voto mayoritario

Probabilidad de acierto de un comité: M modelos base independientes para clasificación binaria; todos con probabilidad de acierto θ

- Dada una muestra la clase 1, la clase escogida por el modelo base m puede verse com una Bernoulli $Y_m \in \{0,1\}$, para todo m
- Así, la suma de los votos a la clase 1, $S=\sum_m Y_m$, es una binomial $\operatorname{Bin}(M,\theta)$
- En definitiva, la probabilidad de acierto del comité puede hallarse a partir de la función de distribución binomial:

$$p = P(S > M/2) = 1 - B(M/2, M, \theta) \qquad (B \text{ es la función de distribución binomial})$$

```
1000 predictores independientes con theta = 0.5100 acertarán con p = 0.7261
```

2.2 Bagging

Bagging (bootstrap aggregating): ensambla M modelos ajustados con diferentes versiones de los datos, obtenidas por boostraping (muestreo con reemplazamiento)

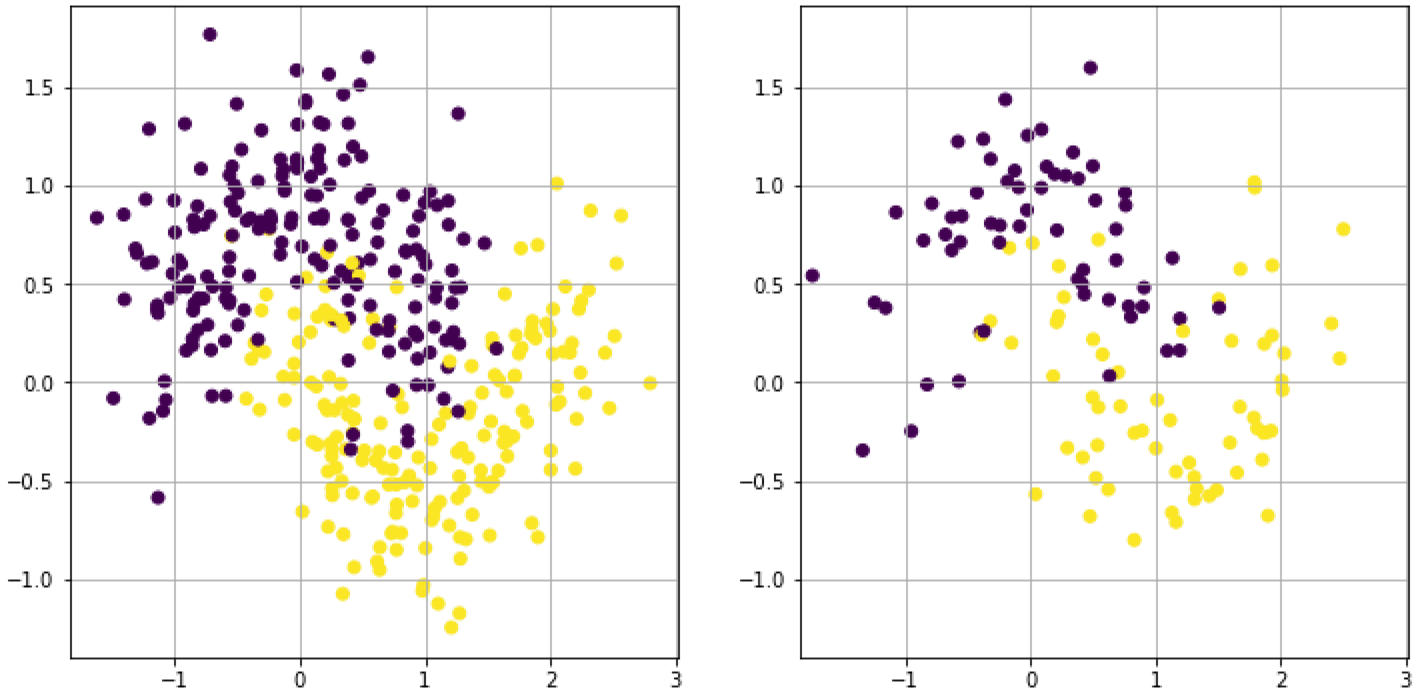
Desventaja: cada modelo base ve un 63% de datos aprox.; en el límite, la probabilidad de que un dato no se seleccione es

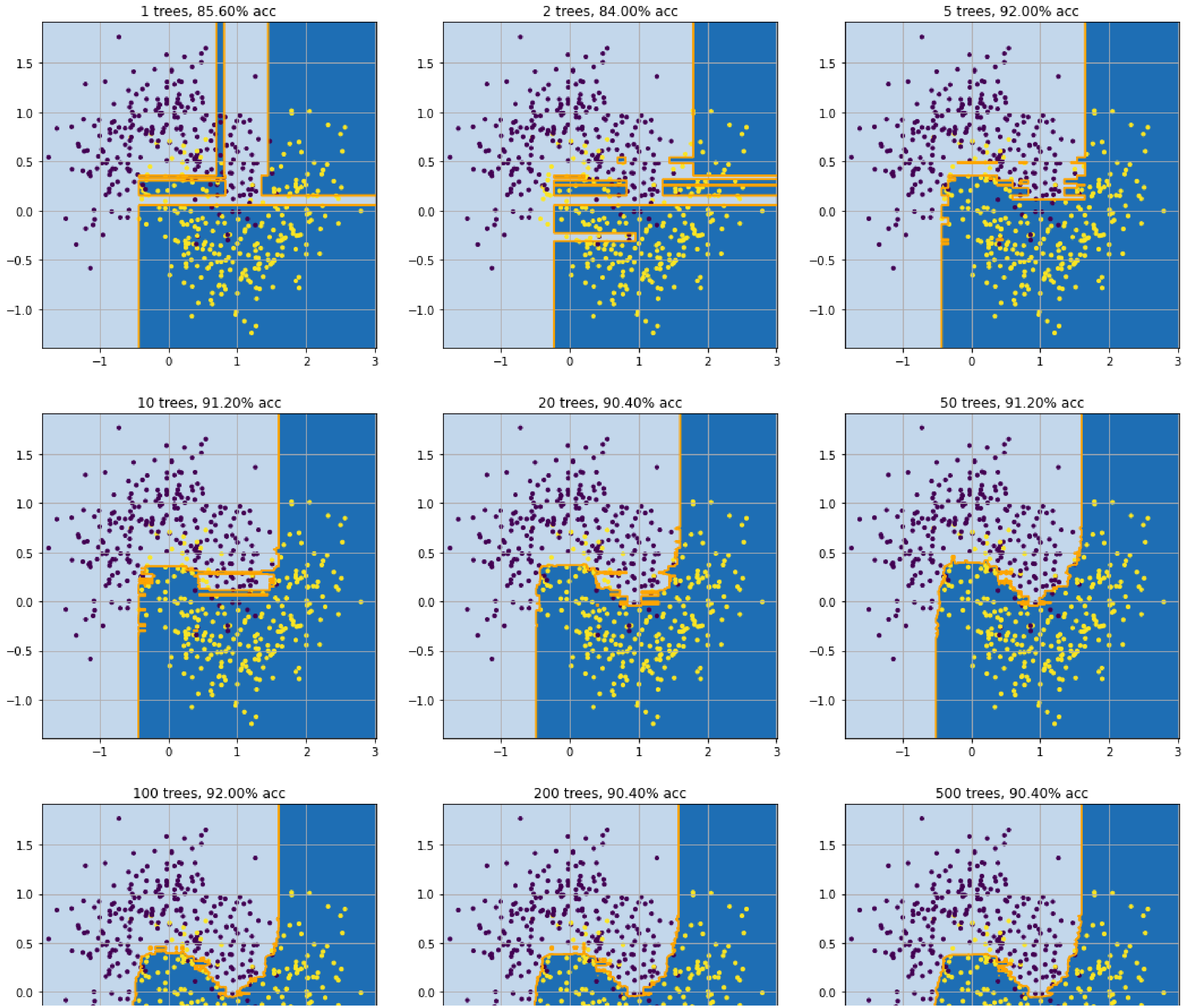
$$p = \lim_{N\to\infty}(1-1/N)^N = e^{-1} \approx 0.37$$

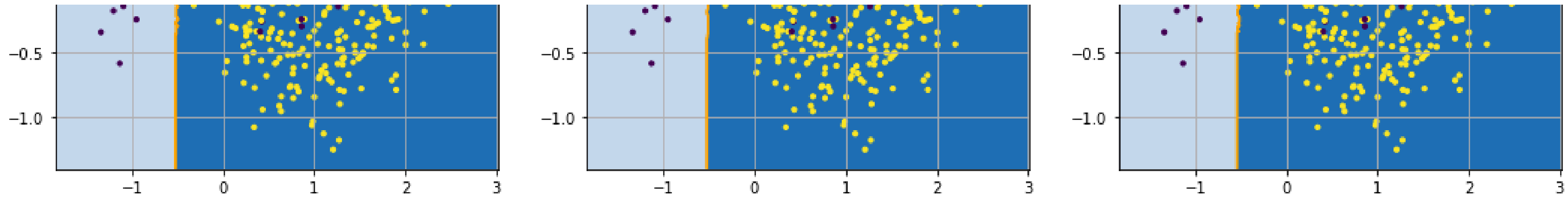
Ventaja: el 37% de muestras **out-of-bag** puede usarse en test

Ventaja principal: el ensamble no depende demasiado de ningún dato individual, lo que favorece mayor robustez y generalización

Ejemplo: bagging de árboles


1.5
1.0
0.5
0.0
−0.5
−1.0
−1
0
1
2
3
1.5
1.0
0.5
0.0
−0.5
−1.0
−1
0
1
2
3


1 trees, 85.60% acc
2 trees, 84.00% acc
5 trees, 92.00% acc
10 trees, 91.20% acc
20 trees, 90.40% acc
50 trees, 91.20% acc
100 trees, 92.00% acc
200 trees, 90.40% acc
500 trees, 90.40% acc



Ejemplo: clasificación de correos en spam y no-spam

```
Bagged 10 trees, test err 5.9%
Bagged 50 trees, test err 5.5%
Bagged 100 trees, test err 5.4%
Bagged 200 trees, test err 5.5%
Bagged 300 trees, test err 5.5%
Bagged 400 trees, test err 5.4%
Bagged 500 trees, test err 5.6%
```

2.3 Random forests

Random forests: variante de bagging de árboles que mejora la decorrelación de modelos base mediante aleatorización, no solo de datos, sino también de variables de entrada; así, la característica de split j_i se optimiza sobre un conjunto aleatorio $S_i \subseteq \{1,\dots,D\}$,

$$(j_i, t_i) = \operatorname*{arg\,min}_{j\in S_i} \min_{t\in\mathcal{T}_j} \frac{|\mathcal{D}_i^L(j,t)|}{|\mathcal{D}_i|}\, c(\mathcal{D}_i^L(j,t)) + \frac{|\mathcal{D}_i^R(j,t)|}{|\mathcal{D}_i|}\, c(\mathcal{D}_i^R(j,t))$$

Ventaja frente a bagging: los bosques suelen ser más precisos que bagging pues muchas características son irrelevantes

Ventaja frente a boosting: los aprendices pueden entrenarse en paralelo, cosa que no puede hacerse en boosting

Ejemplo: clasificación de correos en spam y no-spam

```
RF 10 trees, test err 6.3%
RF 50 trees, test err 5.0%
RF 100 trees, test err 4.9%
RF 200 trees, test err 4.8%
RF 300 trees, test err 4.9%
RF 400 trees, test err 4.8%
RF 500 trees, test err 4.8%
```

3 Boosting

3.1 Boosting

Modelo aditivo de funciones base adaptativas: ensamble visto como suma de modelos base, no necesariamente árboles

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{m=1}^{M} \beta_m \, F_m(\boldsymbol{x};\boldsymbol{\theta}_m)$$

Objetivo: minimizar la pérdida empírica (con regularizador)

$$\mathcal{L}(f) = \sum_{i=1}^{N} \ell(y_i, f(\boldsymbol{x}_i))$$

Boosting (potenciación): ajusta secuencialmente modelos aditivos de clasificadores binarios, $F_m \in \{-1, +1\}$

- Primero ajusta F_1 a los datos y se ponderan con más peso los errores
- Luego ajusta F_2 a los datos ponderados en el paso anterior
- El proceso sigue hasta llegar a M componentes
- Si la precisión de cada **weak learner** F_m es mejor que el azar (50%), la del **strong learner** f será aún mejor

Ventaja frente a bagging y bosques: ofrece mejores resultados pues reduce el sesgo del aprendiz fuerte ajustando árboles que dependen unos de otros; bagging y bosques solo reducen la varianza ajustando árboles independientes

Evolución: propuesto en aprendizaje PAC para clasificación binaria con pérdida específica; actualmente se plantea bajo un marco estadístico general, con pérdidas diversas para extender su aplicación a regresión, clasificación multi-clase, ranking, etc.

Ejemplo: clasificación de correos en spam y no-spam

```
Boosting 10 trees, test err 6.2%
Boosting 50 trees, test err 5.4%
Boosting 100 trees, test err 4.7%
Boosting 200 trees, test err 4.6%
Boosting 300 trees, test err 4.8%
Boosting 400 trees, test err 4.6%
Boosting 500 trees, test err 4.4%
```

3.2 Modelado aditivo por etapas hacia adelante

Forward stagewise additive modeling (FSAM): optimiza la empírica con pérdida genérica y f modelo aditivo

Objetivo FSAM (para el modelo base m): empírica con pérdida genérica, $\ell(y, \hat{y})$

$$L_m(\beta, \boldsymbol{\theta}) = \sum_{i=1}^{N} \ell(y_i, f_{m-1}(\boldsymbol{x}_i) + \beta F(\boldsymbol{x}_i; \boldsymbol{\theta}))$$

Minimización del objetivo y reajuste del modelo: $(\beta_m, \boldsymbol{\theta}_m) = \operatorname{argmin}_{\beta, \boldsymbol{\theta}} \; L_m(\beta, \boldsymbol{\theta})$

$$f_m(\boldsymbol{x}) = f_{m-1}(\boldsymbol{x}) + \beta_m F_m(\boldsymbol{x}) \qquad \text{con} \qquad F_m(\boldsymbol{x}) = F(\boldsymbol{x}; \boldsymbol{\theta}_m)$$

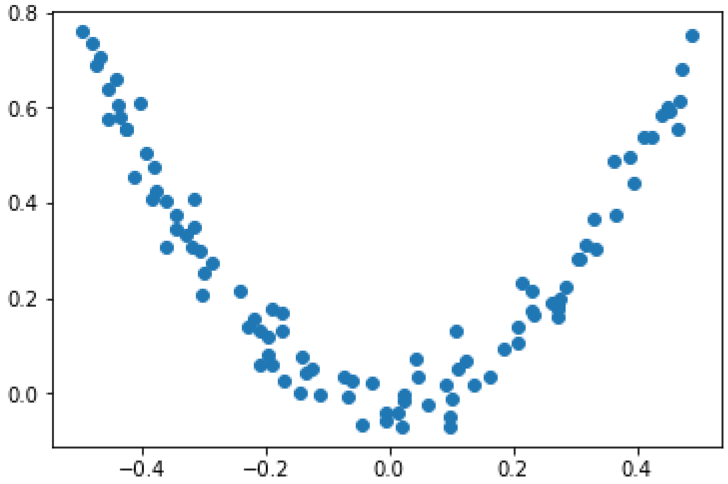
3.3 Boosting mínimos cuadrados

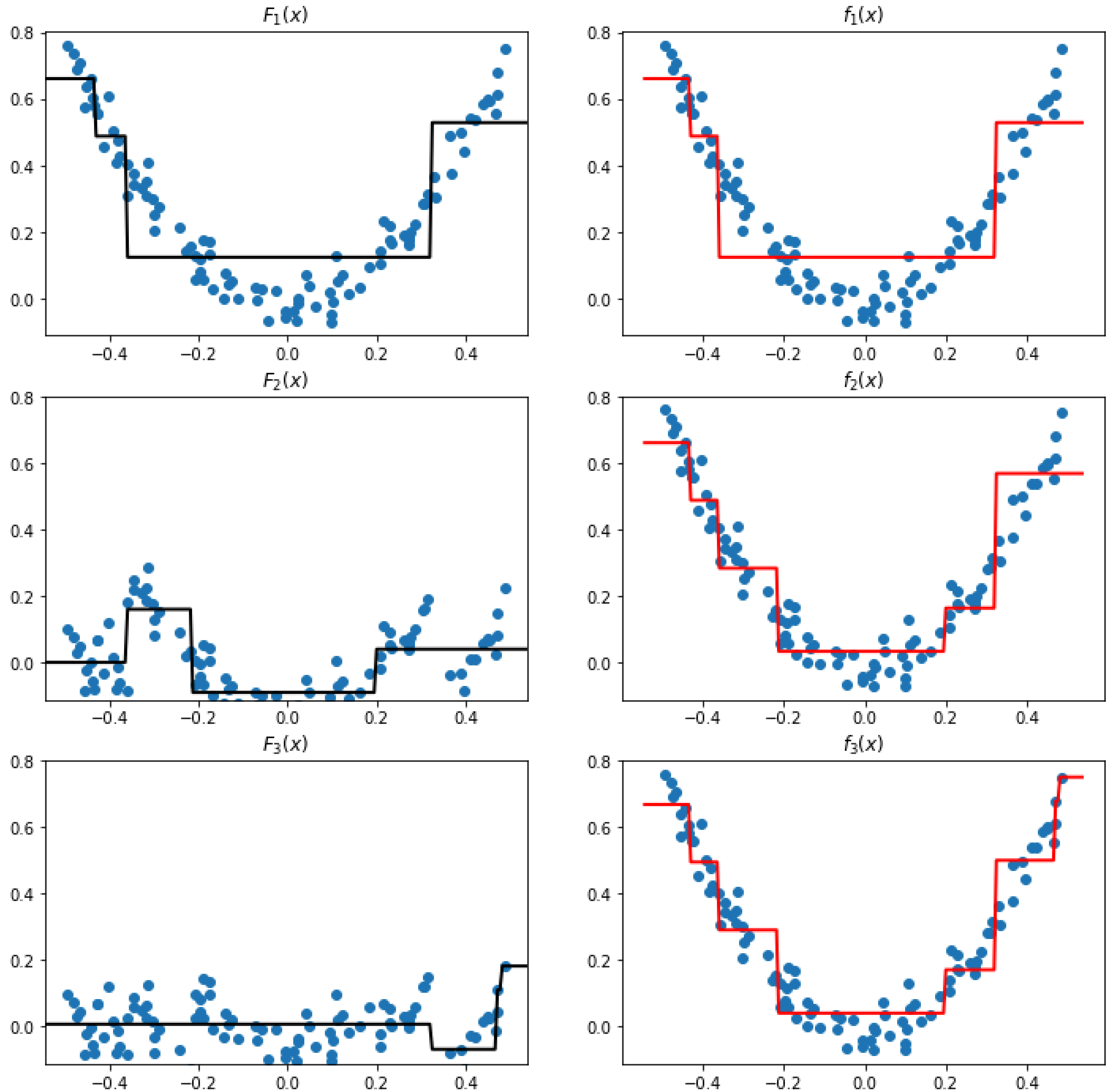
Objetivo FSAM con pérdida cuadrática: $\ell(y, \hat{y}) = (y - \hat{y})^2$

$$L_m(\beta, \boldsymbol{\theta}) = \sum_{i=1}^{N} (r_{im} - \beta F(\boldsymbol{x}_i; \boldsymbol{\theta}))^2 \qquad \text{con residuos} \qquad r_{im} = y_i - f_{m-1}(\boldsymbol{x}_i)$$

Boosting mínimos cuadrados: minimiza el objetivo fijando $\beta=1$ y ajustando F a los residuos

Ejemplo: regresión simple con boosting mínimos cuadrados




$F_1(x)$
$f_1(x)$
$F_2(x)$
$f_2(x)$
$F_3(x)$
$f_3(x)$

3.4 AdaBoost

3.4.1 Objetivo FSAM con pérdida exponencial

Si $\ell(\tilde{y}, \hat{y}) = \exp(-\tilde{y}\hat{y})$ con $\tilde{y} \in \{-1, +1\}$

$$\begin{aligned}
L_m(\beta, \boldsymbol{\theta}) &= \sum_{i=1}^{N} \exp(-\tilde{y}_i(f_{m-1}(\boldsymbol{x}_i) + \beta F(\boldsymbol{x}_i; \boldsymbol{\theta}))) \\
&= \sum_{i=1}^{N} w_{im} \exp(-\beta \tilde{y}_i F(\boldsymbol{x}_i; \boldsymbol{\theta})) \qquad \text{con} \qquad w_{im} = \exp(-\tilde{y}_i f_{m-1}(\boldsymbol{x}_i)) \\
&= e^{\beta} \sum_{\tilde{y}_i \neq F(\boldsymbol{x}_i;\boldsymbol{\theta})} w_{im} + e^{-\beta} \sum_{\tilde{y}_i = F(\boldsymbol{x}_i;\boldsymbol{\theta})} w_{im} \\
&= e^{\beta} \sum_{\tilde{y}_i \neq F(\boldsymbol{x}_i;\boldsymbol{\theta})} w_{im} + e^{-\beta} \left(\sum_{i=1}^{N} w_{im} - \sum_{\tilde{y}_i \neq F(\boldsymbol{x}_i;\boldsymbol{\theta})} w_{im} \right) \\
&= (e^{\beta} - e^{-\beta}) \sum_{i=1}^{N} w_{im} \mathbb{I}(\tilde{y}_i \neq F(\boldsymbol{x}_i; \boldsymbol{\theta})) + e^{-\beta} \sum_{i=1}^{N} w_{im}
\end{aligned}$$

3.4.2 Minimización del objetivo en dos pasos

Primero hallamos $\boldsymbol{\theta}_m$ a partir de los datos ponderados:

$$\boldsymbol{\theta}_m = \operatorname*{argmin}_{\boldsymbol{\theta}} \sum_{i=1}^{N} w_{im} \mathbb{I}(\tilde{y}_i \neq F(\boldsymbol{x}_i; \boldsymbol{\theta}))$$

Luego obtenemos β_m mediante minimización en β de $L_m(\beta, \boldsymbol{\theta}_m)$:

$$\beta_m = \operatorname*{argmin}_{\beta}\ L_m(\beta, \boldsymbol{\theta}_m) = \frac{1}{2} \log \frac{1 - \mathrm{err}_m}{\mathrm{err}_m} \qquad \text{con} \qquad \mathrm{err}_m = \frac{1}{\sum_{i=1}^{N} w_{im}} \sum_{i=1}^{N} w_{im} \mathbb{I}(\tilde{y}_i \neq F_m(\boldsymbol{x}_i))$$

3.4.3 Adaboost

Adaboost: halla $F_m(\cdot)$ y β_m en la iteración m y reajusta el modelo

Pesos de los datos para la primera iteración: $w_{i1} = 1/N$

Pesos de los datos para la iteración $m+1$: se calculan tras hallar $F_m(\cdot)$ y β_m en la iteración m

$$\begin{aligned}
w_{i,m+1} &= \exp(-\tilde{y}_i f_m(\boldsymbol{x}_i)) \\
&= \exp(-\tilde{y}_i f_{m-1}(\boldsymbol{x}_i) - \tilde{y}_i \beta_m F_m(\boldsymbol{x}_i)) \\
&= w_{im} \exp(-\tilde{y}_i \beta_m F_m(\boldsymbol{x}_i)) \\
&= w_{im} \exp(\beta_m (2\mathbb{I}(\tilde{y}_i \neq F_m(\boldsymbol{x}_i)) - 1)) \\
&= w_{im} \exp(2\beta_m \mathbb{I}(\tilde{y}_i \neq F_m(\boldsymbol{x}_i))) \exp(-\beta_m)
\end{aligned}$$

El factor $\exp(-\beta_m)$ se puede ignorar ya que es constante para todos los datos en el objetivo FSAM de la iteración $m+1$. Así pues, los pesos de los datos para la iteración $m+1$ son:

$$w_{i,m+1} = \begin{cases} w_{im} \exp(2\beta_m) & \text{si } \tilde{y}_i \neq F_m(\boldsymbol{x}_i) \\ w_{im} & \text{en otro caso} \end{cases}$$

Modelo ajustado para clasificación binaria: $f(\boldsymbol{x}) = \operatorname{sgn}(\sum_m \beta_m F_m(\boldsymbol{x}))$

Modelos para regresión y clasificación multi-clase: se usan variantes de Adaboost convenientemente adaptadas

3.4.4 Propiedades de Adaboost

Sensibilidad a outliers: ya que los pesos de los datos mal clasificados crecen exponencialmente

Dificultad para estimar probabilidades: en teoría, el riesgo de un modelo $f(\boldsymbol{x})$ con pérdida exponencial es

$$\mathbb{E}[\exp(-\tilde{y} f(\boldsymbol{x})) \mid \boldsymbol{x}] = p(\tilde{y} = 1 \mid \boldsymbol{x}) \exp(-f(\boldsymbol{x})) + p(\tilde{y} = -1 \mid \boldsymbol{x}) \exp(f(\boldsymbol{x}))$$

Derivando con respecto a $f(\boldsymbol{x})$ e igualando a cero, tenemos que el modelo de mínimo riesgo teórico halla la mitad de la log-odds:

$$f(\boldsymbol{x}) = \frac{1}{2}\log\frac{p(\tilde{y}=1 \mid \boldsymbol{x})}{p(\tilde{y}=-1 \mid \boldsymbol{x})}$$

Aunque no obtenemos probabilidades directamente, este resultado justifica la aplicación del operador signo al modelo

Ejemplo: clasificación de correos en spam y no-spam

```
AdaBoosting 10 trees, test err 10.9%
AdaBoosting 50 trees, test err 7.2%
AdaBoosting 100 trees, test err 5.9%
AdaBoosting 200 trees, test err 5.7%
AdaBoosting 300 trees, test err 5.7%
AdaBoosting 400 trees, test err 5.5%
AdaBoosting 500 trees, test err 5.5%
```

3.5 LogitBoost

Predicción probabilística: usamos el modelo aditivo para predecir la mitad de la log-odds

$$p(\tilde{y} \mid \boldsymbol{x};\boldsymbol{\theta}) = \sigma(\tilde{y}a) \qquad \text{con} \qquad a = 2f(\boldsymbol{x};\boldsymbol{\theta})$$

Objetivo FSAM con log-pérdida y $\beta=1$: $\ell(\tilde{y},\boldsymbol{\theta};\boldsymbol{x}) = -\log p(\tilde{y} \mid \boldsymbol{x};\boldsymbol{\theta})$

$$\begin{aligned}
L_m(\boldsymbol{\theta}) &= -\sum_{i=1}^N \log(\sigma(\tilde{y}_i 2[f_{m-1}(\boldsymbol{x}_i) + F(\boldsymbol{x}_i;\boldsymbol{\theta})])) \\
&= \sum_{i=1}^N \log(1+\exp(-2\tilde{y}_i[f_{m-1}(\boldsymbol{x}_i) + F(\boldsymbol{x}_i;\boldsymbol{\theta})]))
\end{aligned}$$

LogitBoost: algoritmo de Newton para minimizar este objetivo directamente

3.6 Gradient boosting

Gradient boosting: FSAM visto como descenso por gradiente para un problema de minimización en un espacio funcional

$$\hat{\boldsymbol{f}} = \underset{\boldsymbol{f}}{\operatorname{argmin}}\, \mathcal{L}(\boldsymbol{f}) \quad \text{con} \quad \mathcal{L}(\boldsymbol{f}) = \sum_{i=1}^{N} \ell(y_i, f(\boldsymbol{x}_i))$$

Funciones base simplificadas: valores en el conjunto de entrenamiento, $\boldsymbol{f} = (f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_N))^t$

Descenso por gradiente: escoge la "dirección" de máximo descenso, esto es, la del neg-gradiente de $\mathcal{L}(\boldsymbol{f})$ en $\boldsymbol{f}_{m-1}$, $\boldsymbol{g}_m$

$$\boldsymbol{f}_m = \boldsymbol{f}_{m-1} - \beta_m \boldsymbol{g}_m \qquad \text{con} \qquad g_{im} = \left[\frac{\partial \ell(y_i, f(\boldsymbol{x}_i))}{\partial f(\boldsymbol{x}_i)}\right]_{f_{m-1}(\boldsymbol{x}_i)}$$

Factor de aprendizaje: β_m puede escogerse por búsqueda lineal

Funciones base generalizadas: para poder generalizar, se ajusta un aprendiz débil al neg-gradiente con pérdida cuadrática

$$F_m = \underset{F}{\operatorname{argmin}} \sum_{i=1}^{N} (-g_{im} - F(\boldsymbol{x}_i))^2$$

3.6.1 Algoritmo básico

El algoritmo básico prescinde de β_m pero incluye un **shrinkage factor** $0 < \nu \leq 1$ para facilitar la regularización:

1. Inicializar $f_0(\boldsymbol{x}) = \operatorname{argmin}_F \sum_{i=1}^{N} \ell(y_i, F(\boldsymbol{x}_i))$
2. **for** $m = 1 : M$ **do**
3. Calcular el neg-gradiente o (pseudo-)**residuo** $r_{im} = -\left[\frac{\partial \ell(y_i, f(\boldsymbol{x}_i))}{\partial f(\boldsymbol{x}_i)}\right]_{f_{m-1}(\boldsymbol{x}_i)}$

4. Usar el aprendiz débil para hallar $F_m = \operatorname{argmin}_F \sum_{i=1}^N (r_{im} - F(\boldsymbol{x}_i))^2$
5. Actualizar $f_m(\boldsymbol{x}) = f_{m-1}(\boldsymbol{x}) + \nu F_m(\boldsymbol{x})$
6. Devolver $f(\boldsymbol{x}) = f_M(\boldsymbol{x})$

3.6.2 Regresión

Salidas: $y_i \in \mathbb{R}$

Pérdida cuadrática o su mitad: $\ell(y_i, f(\boldsymbol{x}_i)) = \frac{1}{2}(y_i - f(\boldsymbol{x}_i))^2$ (como boosting mínimo cuadrados)

Residuo de la pérdida cuadrática: $r_i = y_i - f(\boldsymbol{x}_i)$

Pérdida valor absoluto: $\ell(y_i, f(\boldsymbol{x}_i)) = |y_i - f(\boldsymbol{x}_i)|$

Residuo de la pérdida valor absoluto: $r_i = \operatorname{sgn}(y_i - f(\boldsymbol{x}_i))$

3.6.3 Clasificación binaria

Salidas: $\tilde{y}_i \in \{-1, +1\}$

Pérdida exponencial: $\ell(\tilde{y}_i, f(\boldsymbol{x}_i)) = \exp(-\tilde{y}_i f(\boldsymbol{x}_i))$ (como Adaboost)

Residuo de la pérdida exponencial: $r_i = \tilde{y}_i \exp(-\tilde{y}_i f(\boldsymbol{x}_i))$

Log-pérdida binaria: $\ell(\tilde{y}_i, f(\boldsymbol{x}_i)) = \log(1 + \exp(-\tilde{y}_i f(\boldsymbol{x}_i)))$ (como LogitBoost)

Residuo de la log-pérdida binaria:

$$r_i = -\frac{1}{1 + \exp(-\tilde{y}_i f(\boldsymbol{x}_i))} \exp(-\tilde{y}_i f(\boldsymbol{x}_i))(-\tilde{y}_i) = \tilde{y}_i \frac{1}{1 + \exp(\tilde{y}_i f(\boldsymbol{x}_i))} = \tilde{y}_i \sigma(-\tilde{y}_i f(\boldsymbol{x}_i))$$

3.6.4 Clasificación multiclase

Salidas: $y_i \in \{1, \dots, C\}$

Log-pérdida: se ajustan C modelos aditivos, uno por cada clase, cuyas predicciones se normalizan mediante una softmax

$$\ell(y_i, f_1(\boldsymbol{x}_i), \dots, f_C(\boldsymbol{x}_i)) = -\sum_c \mathbb{I}(y_i = c) \log \pi_{ic} \quad \text{con} \quad \pi_{ic} = S(f_1(\boldsymbol{x}_i), \dots, f_C(\boldsymbol{x}_i))_c = \frac{\exp(f_c(\boldsymbol{x}_i))}{\sum_{c'=1}^C \exp(f_{c'}(\boldsymbol{x}_i))}$$

Residuo de la log-pérdida: para cada clase c

$$\begin{aligned} r_{ic} &= -\frac{\partial \ell(y_i, f_1(\boldsymbol{x}_i), \dots, f_C(\boldsymbol{x}_i))}{\partial f_c(\boldsymbol{x}_i)} \\ &= \frac{\partial}{\partial f_c(\boldsymbol{x}_i)} \sum_{\tilde{c}} \mathbb{I}(y_i = \tilde{c}) \log \pi_{i\tilde{c}} \\ &= \mathbb{I}(y_i = c) \frac{1}{\pi_{ic}} \pi_{ic}(1 - \pi_{ic}) \\ &= \mathbb{I}(y_i = c)(1 - \pi_{ic}) \end{aligned}$$

3.6.5 Gradient tree boosting

Gradient tree boosting: gradient boosting con árbol de regresión como aprendiz débil

$$F_m = \sum_{j=1}^{J_m} w_{jm} \mathbb{I}(\boldsymbol{x} \in R_{jm})$$

- R_{jm} y w_{jm} son la región y salida asociadas a la hoja j del árbol añadido en la iteración m
- La salida puede ser un escalar o, más generalmente, un vector (de probabilidades, por ejemplo)

Aprendizaje de las regiones: CART sobre residuos

Aprendizaje de las salidas: minimización del riesgo empírico con los datos de la hoja

$$\hat{w}_{jm} = \underset{w}{\operatorname{argmin}} \sum_{\boldsymbol{x}_i \in R_{jm}} \ell(y_i, f_{m-1}(\boldsymbol{x}_i) + w)$$

Aprendizaje de las salidas con pérdida cuadrática: $\hat{w}_{jm}$ es la media empírica de los residuos de la hoja

3.6.6 XGBoost

Extreme gradient boosting (XGBoost): implementación muy popular de gradient tree boosting con algunos refinamientos

- Objetivo regularizado
- Aproximación de segundo orden de la pérdida
- Muestreo de características en nodos internos
- Técnicas algorítmicas varias para mejorar la escabilidad

4 Interpretación de ensambles de árboles

4.1 Importancia de características

Importancia de una característica k en un árbol T: suma de ganancias (reducciones de coste) en los nodos v_j que la usan

$$R_k(T) = \sum_j G_j \mathbb{I}(v_j = k)$$

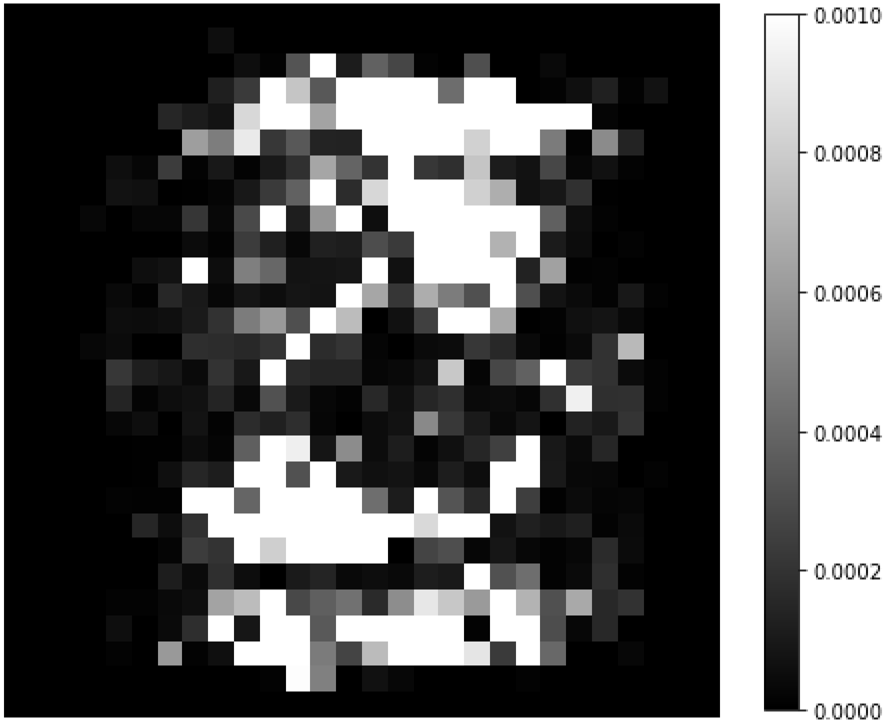
Importancia de una característica k en un ensamble de M árboles: extensión mediante promediado

$$R_k = \frac{1}{M} \sum_{m=1}^{M} R_k(T_m)$$

Normalización de importancias: suelen normalizarse con respecto a la máxima (100%)

Ejemplo: importancias para clasificador (de dígitos escogidos en) MNIST

```
(13834, 784) (13834,)
```



4.2 Gráficos de dependencia parcial

Gráfico de dependencia parcial: muestra la predicción del modelo en función de una (x_k) o dos (x_j, x_k) características

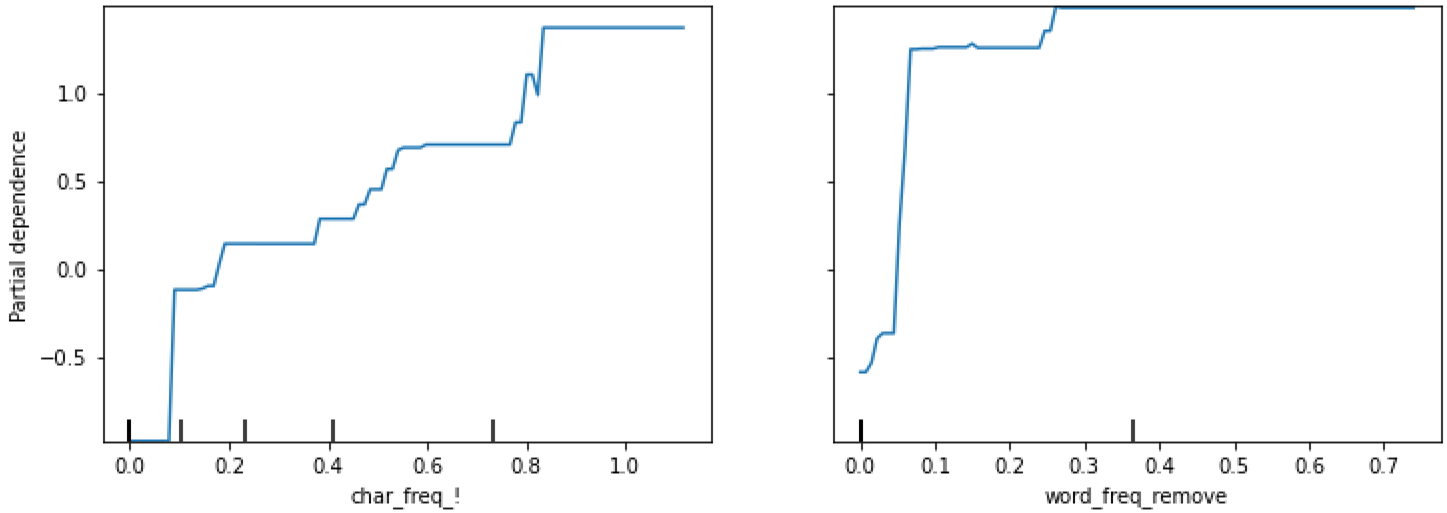
$$\bar{f}_k(x_k) = \frac{1}{N}\sum_{n=1}^{N} f(\boldsymbol{x}_{n,-k}, x_k)$$
$$\bar{f}_{jk}(x_j, x_k) = \frac{1}{N}\sum_{n=1}^{N} f(\boldsymbol{x}_{n,-jk}, x_j, x_k)$$

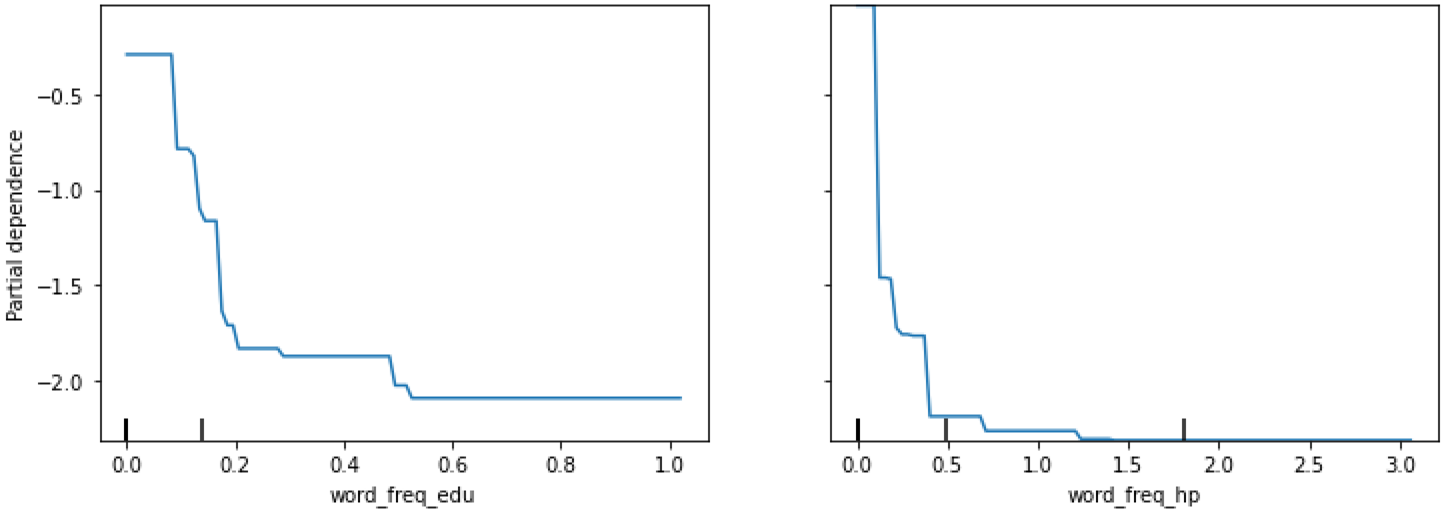
Gráfico de dependencia parcial en clasificación binaria: muestra la log-odds en función de x_k (y x_k)

Ejemplo: dependencia parcial de log-odds de la clase spam; aumenta con la frecuencia de ch! y remove; edu y hp la disminuyen

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4601 entries, 0 to 4600
Data columns (total 57 columns):
 #   Column                      Non-Null Count  Dtype  
---  ------                      --------------  -----  
 0   word_freq_make              4601 non-null   float64
 1   word_freq_address           4601 non-null   float64
 2   word_freq_all               4601 non-null   float64
 3   word_freq_3d                4601 non-null   float64
 4   word_freq_our               4601 non-null   float64
 5   word_freq_over              4601 non-null   float64
 6   word_freq_remove            4601 non-null   float64
 7   word_freq_internet          4601 non-null   float64
 8   word_freq_order             4601 non-null   float64
 9   word_freq_mail              4601 non-null   float64
 10  word_freq_receive           4601 non-null   float64
 11  word_freq_will              4601 non-null   float64
 12  word_freq_people            4601 non-null   float64
 13  word_freq_report            4601 non-null   float64
 14  word_freq_addresses         4601 non-null   float64
 15  word_freq_free              4601 non-null   float64
 16  word_freq_business          4601 non-null   float64
 17  word_freq_email             4601 non-null   float64
 18  word_freq_you               4601 non-null   float64
 19  word_freq_credit            4601 non-null   float64
 20  word_freq_your              4601 non-null   float64
 21  word_freq_font              4601 non-null   float64
 22  word_freq_000               4601 non-null   float64
 23  word_freq_money             4601 non-null   float64
 24  word_freq_hp                4601 non-null   float64
 25  word_freq_hpl               4601 non-null   float64
 26  word_freq_george            4601 non-null   float64
 27  word_freq_650               4601 non-null   float64
 28  word_freq_lab               4601 non-null   float64
 29  word_freq_labs              4601 non-null   float64
 30  word_freq_telnet            4601 non-null   float64
 31  word_freq_857               4601 non-null   float64
 32  word_freq_data              4601 non-null   float64
 33  word_freq_415               4601 non-null   float64
 34  word_freq_85                4601 non-null   float64
 35  word_freq_technology        4601 non-null   float64
```

```
 36  word_freq_1999              4601 non-null   float64
 37  word_freq_parts             4601 non-null   float64
 38  word_freq_pm                4601 non-null   float64
 39  word_freq_direct            4601 non-null   float64
 40  word_freq_cs                4601 non-null   float64
 41  word_freq_meeting           4601 non-null   float64
 42  word_freq_original          4601 non-null   float64
 43  word_freq_project           4601 non-null   float64
 44  word_freq_re                4601 non-null   float64
 45  word_freq_edu               4601 non-null   float64
 46  word_freq_table             4601 non-null   float64
 47  word_freq_conference        4601 non-null   float64
 48  char_freq_;                 4601 non-null   float64
 49  char_freq_(                 4601 non-null   float64
 50  char_freq_[                 4601 non-null   float64
 51  char_freq_!                 4601 non-null   float64
 52  char_freq_$                 4601 non-null   float64
 53  char_freq_#                 4601 non-null   float64
 54  capital_run_length_average  4601 non-null   float64
 55  capital_run_length_longest  4601 non-null   int64
 56  capital_run_length_total    4601 non-null   int64
dtypes: float64(55), int64(2)
memory usage: 2.0 MB
```




Partial dependence
−0.5
−1.0
−1.5
−2.0
0.0
0.2
0.4
0.6
0.8
1.0
word_freq_edu
0.0
0.5
1.0
1.5
2.0
2.5
3.0
word_freq_hp