Ensembles

Álvaro Barbero Jiménez
alvaro.barbero.jimenez@gmail.com

1. Introducción

Definición de ensemble

Un “**ensemble**” o conjunto de clasificadores es una agrupación de varios modelos de aprendizaje, que se construyen para resolver el mismo problema de aprendizaje automático y toman **decisiones en conjunto** para realizar predicciones.

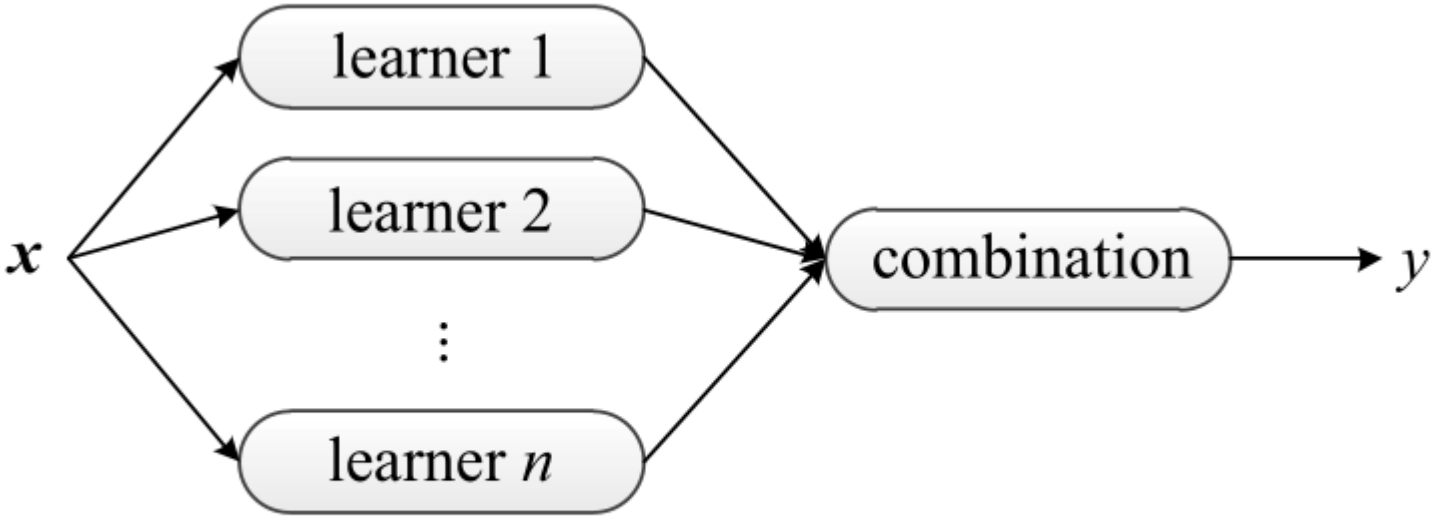


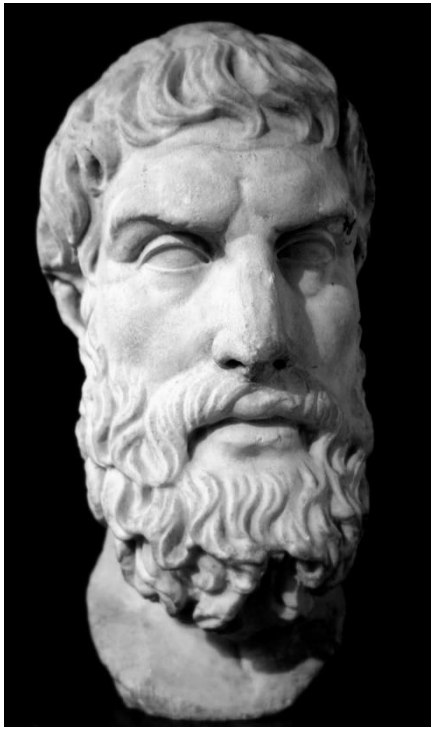
FIGURE 1.9: A common ensemble architecture.

Filosofía de los ensembles

Navaja de Occam
(Guillermo de Occam, 1287–1347)

Cuando varias hipótesis explican un mismo fenómeno, debe preferirse la más simple.

Principio de las explicaciones múltiples
(Epicuro, 341-270 a.c.)

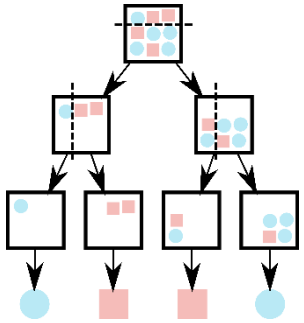


Si varias teorías son consistentes con las observaciones, deben mantenerse todas.

Tipos de modelos base

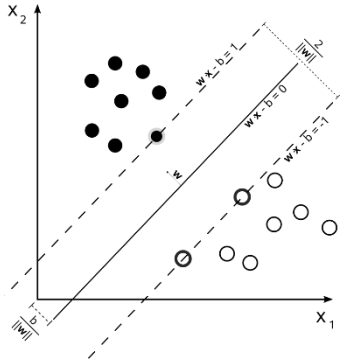
Construir un ensemble implica entrenar una serie de **modelos base**, y luego combinar sus predicciones de alguna manera. Existen dos grandes categorías de modelos base.

Modelos débiles



Modelos poco potentes que funcionan ligeramente mejor que generar predicciones basadas en los prioris de clases o la media de regresión. Rápidos de entrenar. Árboles sencillos, regresión lineal, etc…

Modelos fuertes



Modelos potentes que requieren mucho esfuerzo de entrenamiento pero suelen conseguir buenos resultados. SVMs, redes neuronales, etc…

Métricas deseables en ensembles

Los modelos base de un ensemble se construyen de manera que se maximice la combinación de dos métricas, a menudo contrapuestas:

Precisión

La precisión de la predicción combinada de un ensemble depende de la precisión de cada uno de sus modelos base.

Diversidad

Si todos los modelos base son idénticos o muy parecidos, combinar sus predicciones no aporta nada.

Todos es mejor que el más preciso

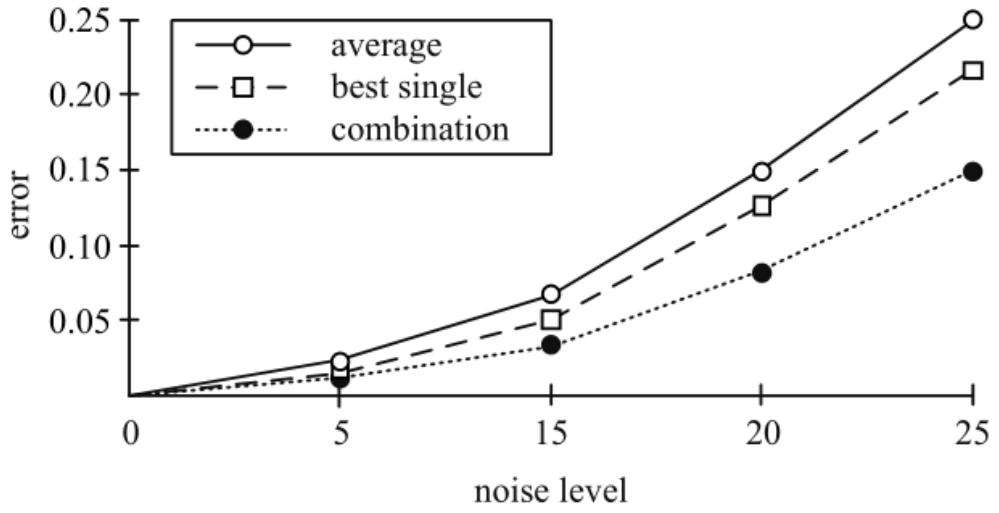


FIGURE 1.10: A simplified illustration of Hansen and Salamon [1990]'s observation: Ensemble is often better than the best single.

Caso histórico: Netflix Prize

En 2006, Netflix inició un concurso con el objetivo de mejorar en un 10% de error relativo su sistema de recomendación de películas en DVD. El premio: 1.000.000$

La solución ganadora resultó ser la combinación en un ensemble de todos los algoritmos creados por tres equipos participantes (BellKor, Pragmatic Chaos y BigChaos). El ensemble incluía:

- Factorización de matrices
- Vecinos próximos
- Modelos de regresión
- Máquinas de Boltzmann


CASEY JOHNSTON, ARS TECHNICA BUSINESS 04.16.12 8:20 AM

NETFLIX NEVER USED ITS $1 MILLION ALGORITHM DUE TO ENGINEERING COSTS

Para combinar los modelos base se utilizaron tanto redes neuronales como Gradient Boosting.

https://www.wired.com/2012/04/netflix-prize-costs/

Caso histórico: Xbox Kinect

Microsoft

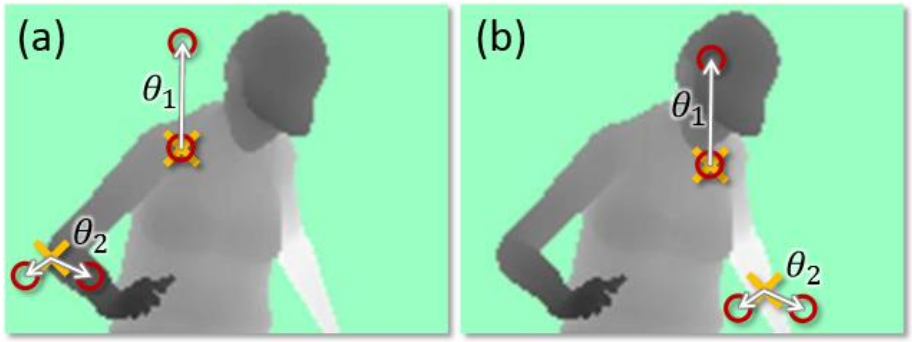


Figure 3. **Depth image features.** The yellow crosses indicates the pixel $\mathbf{x}$ being classified. The red circles indicate the offset pixels as defined in Eq. 1. In (a), the two example features give a large depth difference response. In (b), the same two features at new image locations give a much smaller response.

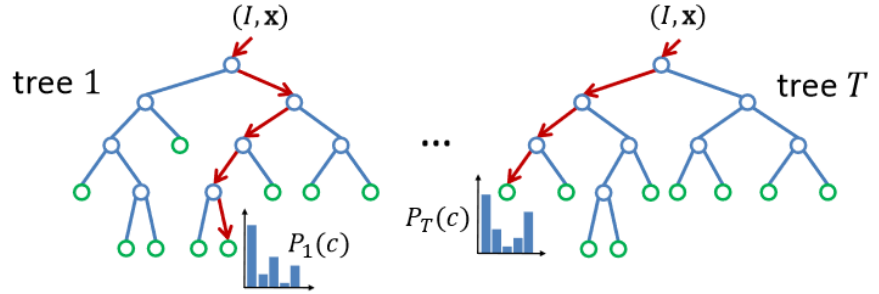
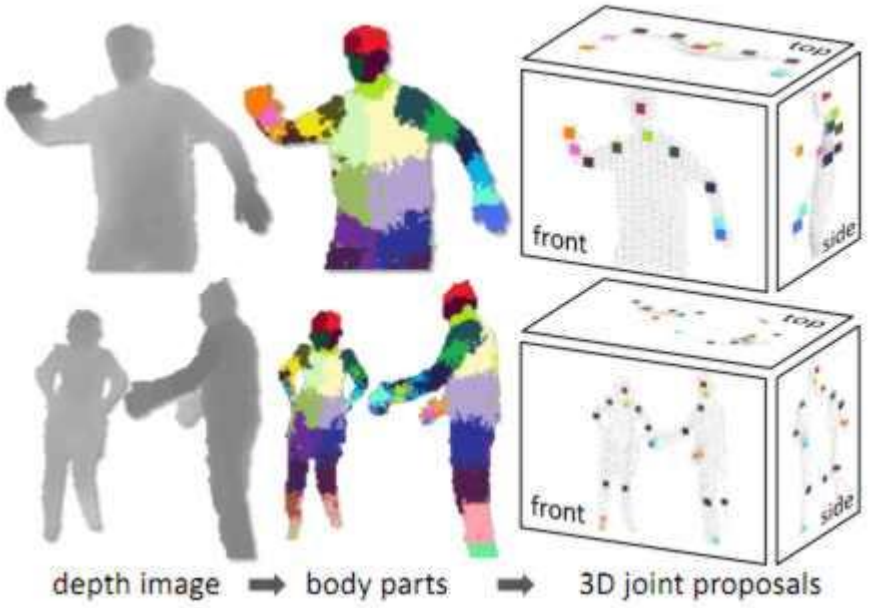
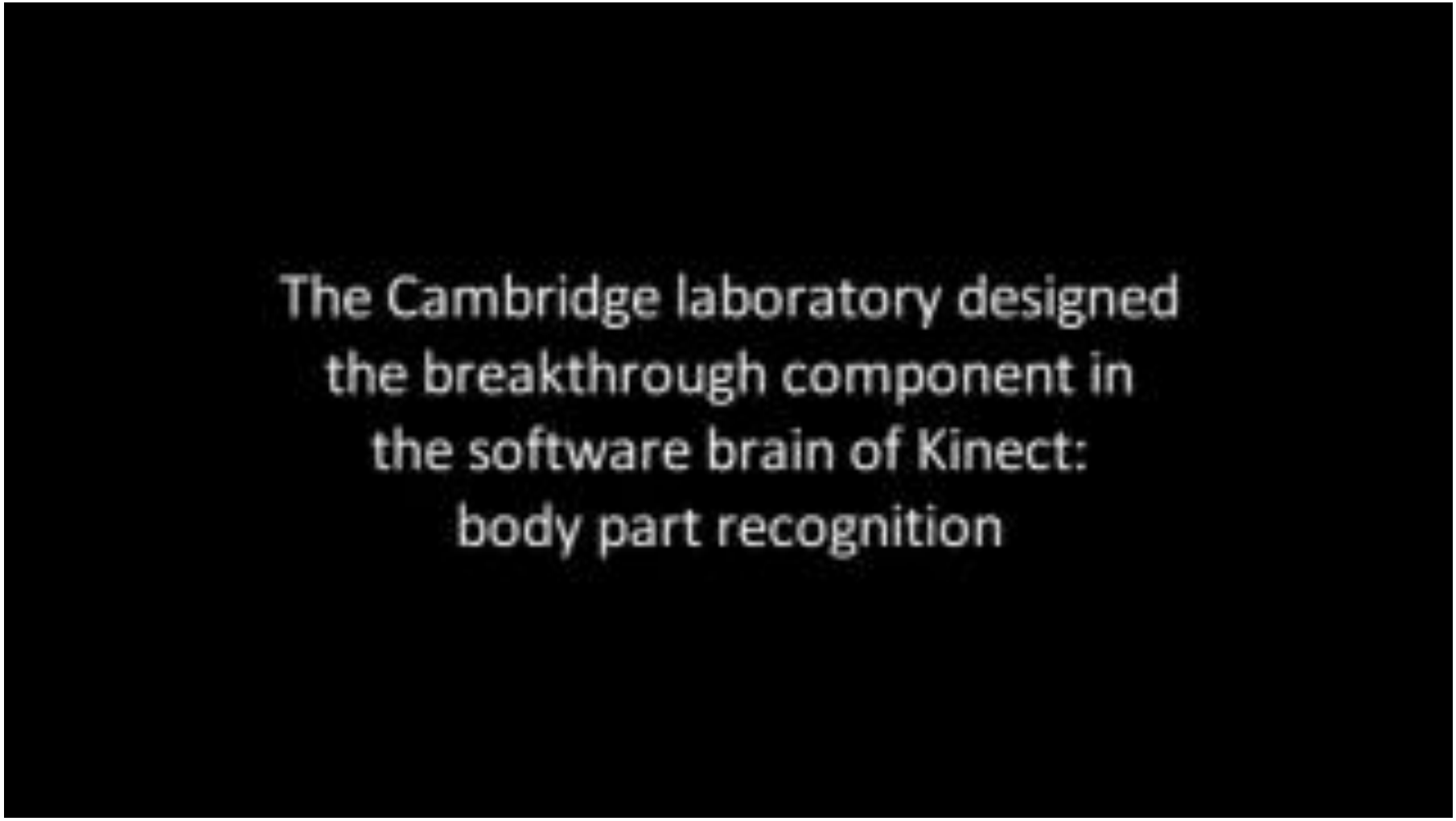


Figure 4. **Randomized Decision Forests.** A forest is an ensemble of trees. Each tree consists of split nodes (blue) and leaf nodes (green). The red arrows indicate the different paths that might be taken by different trees for a particular input.



Shotton et al – Real-Time Human Pose Recognition in Parts from Single Depth Images

Caso histórico: Xbox Kinect

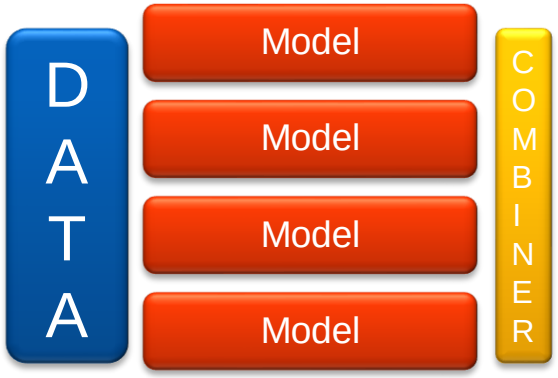


Shotton et al – Real-Time Human Pose Recognition in Parts from Single Depth Images

2. Técnicas de construcción de ensembles

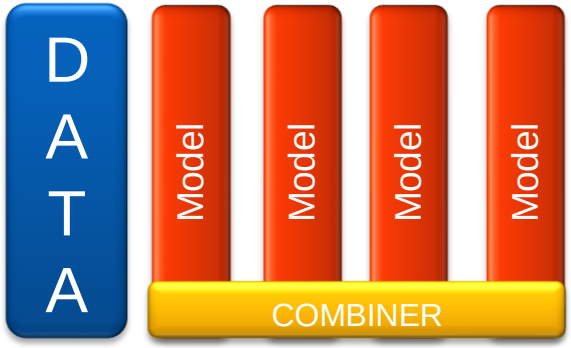
Principales paradigmas de construcción

Ensembles paralelos



Cada modelo trabaja en paralelo sobre los datos, y no colaboran directamente entre ellos. La predicción de todos los modelos se combina.

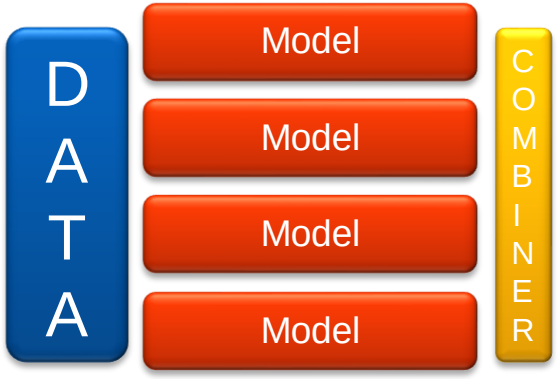
Ensembles secuenciales



Cada modelo trata de corregir los defectos del modelo anterior. La predicción de todos los modelos también se combina.

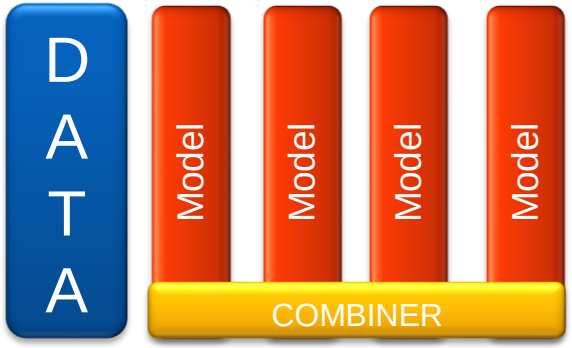
Ventajas de paradigmas

Ensembles paralelos



- ✓ Explotan la independencia entre modelos, fomentando la diversidad
- ✓ Permite el cálculo simultáneo de todos los modelos, ideal para arquitecturas paralelas

Ensembles secuenciales



- ✓ Explotan la dependencia entre modelos, colaborando explícitamente para reducir el error global
- ✓ Se demuestra teóricamente que puede construirse un ensemble fuerte a partir de modelos base débiles

Cotas de error a la combinación

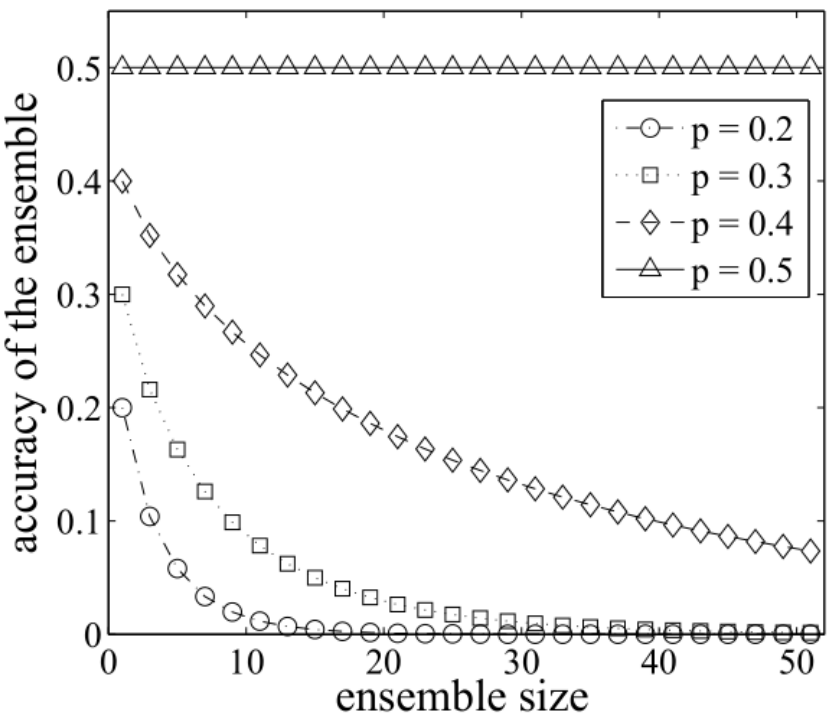
Suponer un problema de clasificación binaria $y_i = \{-1, +1\}$, y T clasificadores base en un ensemble que combina sus predicciones de la forma $F(x) = sign(\sum_{i=1}^{T} f_i(x))$, esto es, sumando las predicciones de todos los modelos base y tomando el signo del resultado.

Suponer además que la probabilidad de error de un clasificador base es $P(f_i(x) \neq y_i) = \epsilon$, y que esta probabilidad es independiente entre clasificadores.

Usando argumentos de combinatoria y que la probabilidad de fallo del ensemble es la probabilidad de que la mitad o más de los modelos base se equivoquen:

$$P(F(x) \neq y) = \sum_{k=0}^{\lfloor T/2 \rfloor} \binom{T}{k} (1-\epsilon)^k \epsilon^{T-k} \leq exp\left(-\frac{1}{2} T(2\epsilon - 1)^2\right)$$

Cotas de error a la combinación



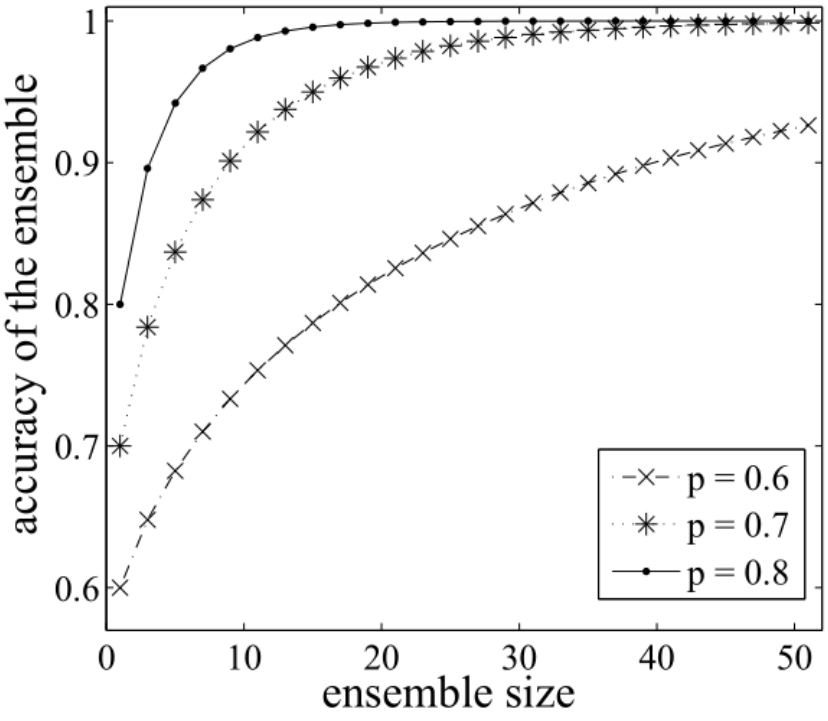


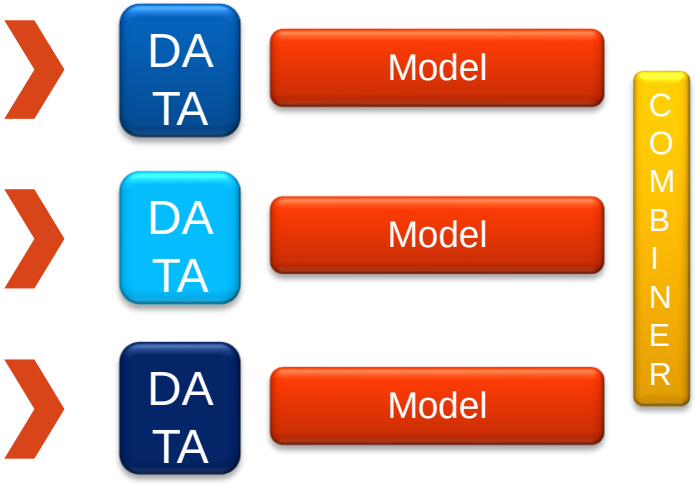
FIGURE 4.2: Ensemble accuracy of majority voting of T independent classifiers with accuracy p for binary classification.

Fomentando la diversidad

!!! Asumir que todos los modelos base de un ensemble van ser independientes entre ellos no es realista.

✓ Pero pueden diseñarse medidas para fomentar esta diversidad.

Suponiendo que contamos con un suministro infinito de datos, una manera sencilla de fomentar la diversidad es tomar muestras diferentes de este suministro, de forma aleatoria, para entrenar cada modelo base:

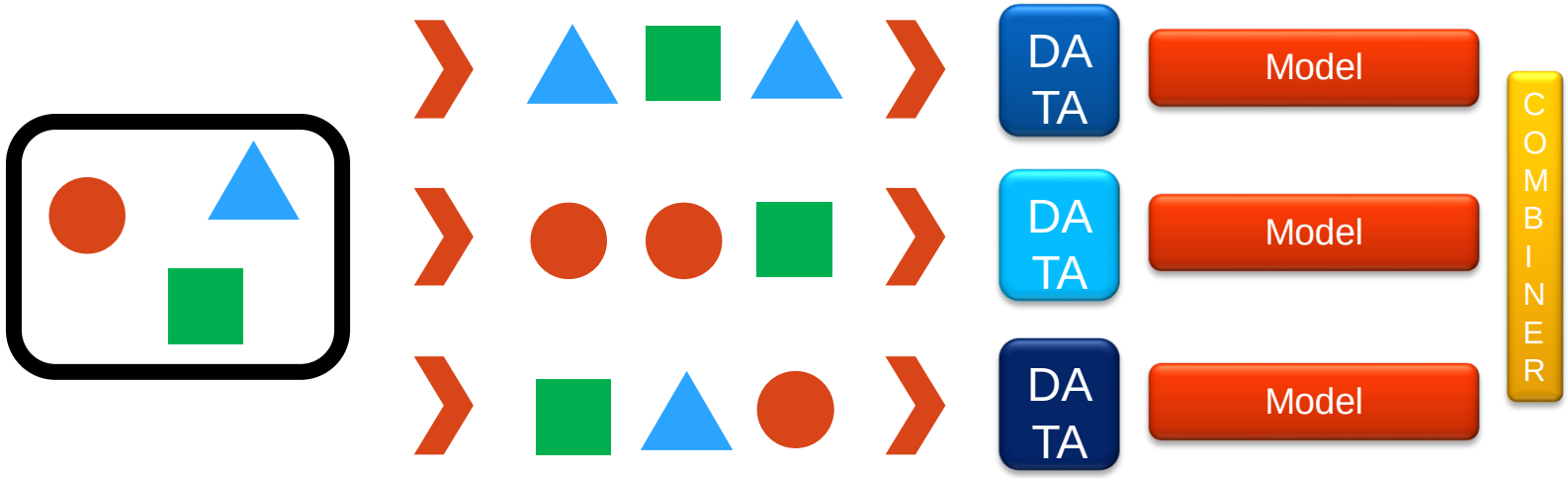


Fomentando la diversidad - Bagging

No es habitual tener una fuente de datos ilimitada, pero el mismo efecto puede simularse con Bagging (Bootstrap AGGregatING).

Para cada modelo, tomar una muestra aleatoria con reemplazamiento del conjunto de datos, del mismo tamaño que el conjunto original. (Bootstrap Sampling)

Para combinar los clasificadores se agregan sus resultados (Aggregating)



Fomentando la diversidad - Bagging

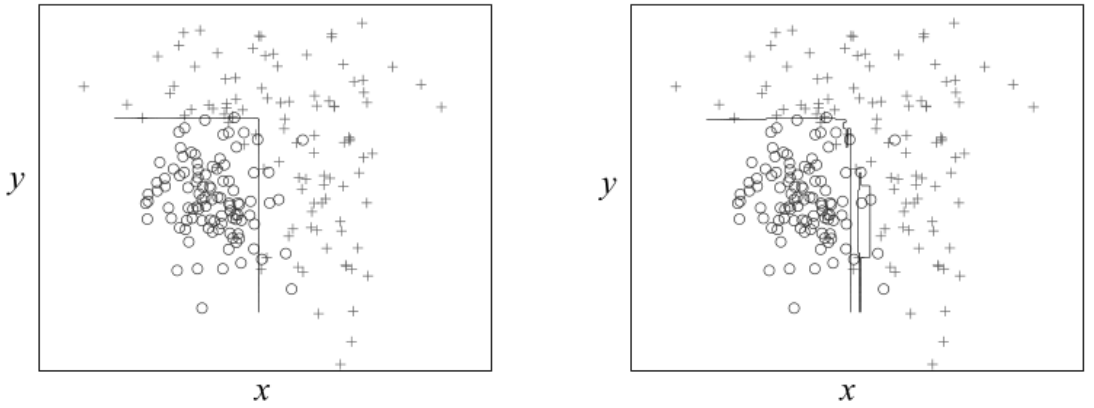


FIGURE 3.2: Decision boundaries of (a) a single decision tree, (b) Bagging and (c) the 10 decision trees used by Bagging, on the *three-Gaussians* data set.

Zhou – Ensemble methods – Foundations and Algorithms

Fomentando la diversidad - Bagging

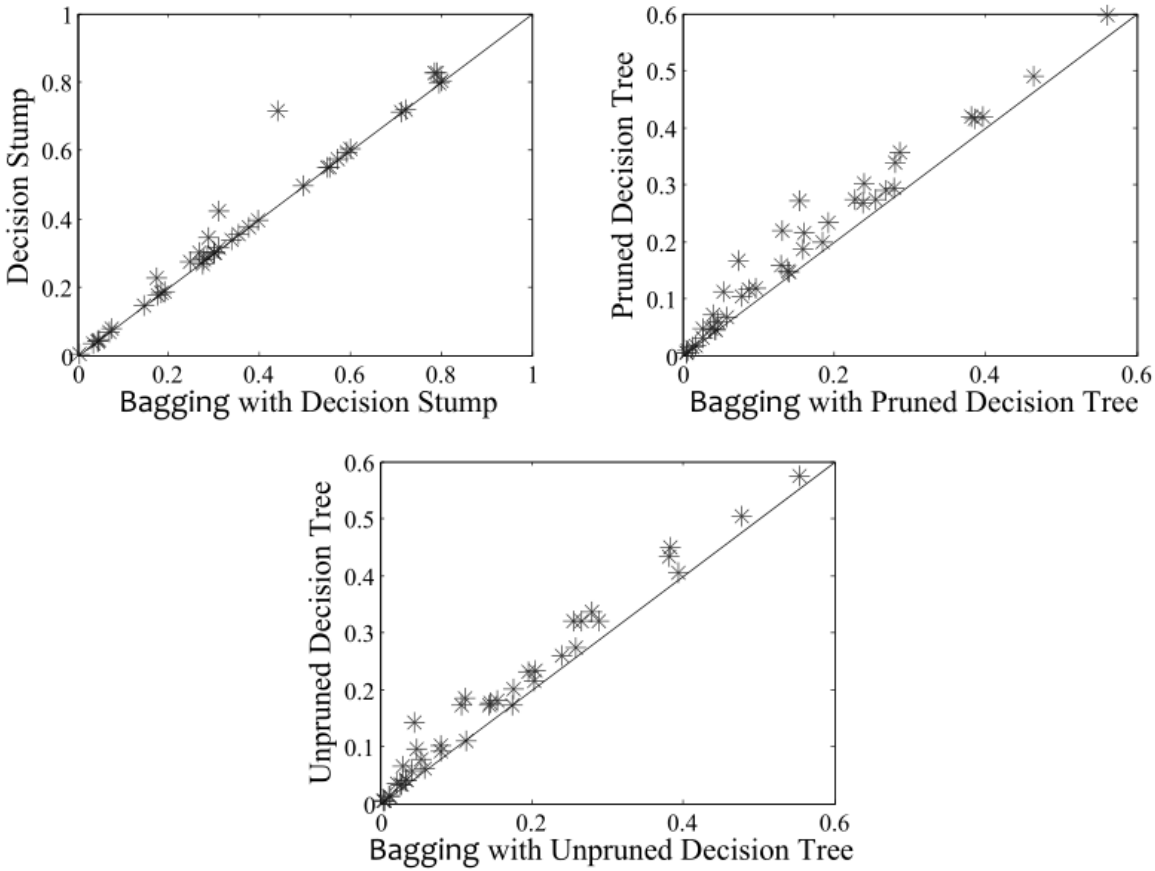


FIGURE 3.3: Comparison of predictive errors of Bagging against single base learners on 40 UCI data sets. Each point represents a data set and locates according to the predictive error of the two compared algorithms. The diagonal line indicates where the two compared algorithms have identical errors.

Zhou – Ensemble methods – Foundations and Algorithms

Clasificadores inestables

Cuando se utiliza bagging para conseguir diversidad es muy importante que los modelos base utilizados sean inestables.

Un clasificador es estable si tiene poca sensibilidad ante perturbaciones en los datos de entrenamiento.

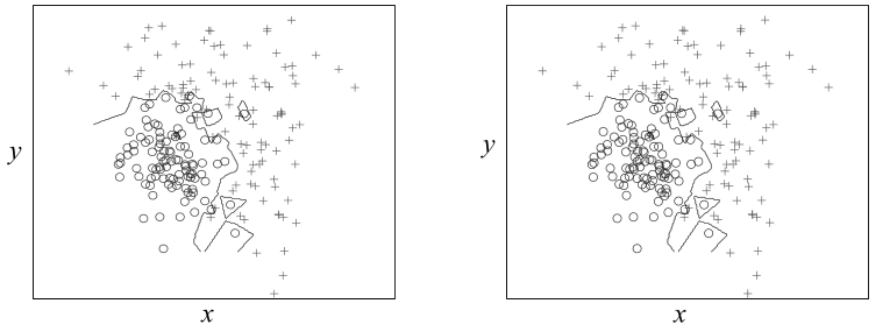


FIGURE 3.5: Decision boundaries of (a) 1-nearest neighbor classifier, and (b) Bagging of 1-nearest neighbor classifiers, on the *three-Gaussians* data set.

Modelos estables

- Vecinos próximos
- SVMs
- Regresión lineal
- Modelos con regularización l_2

Modelos inestables

- Árboles de decisión
- Redes neuronales (algunas)

Zhou – Ensemble methods – Foundations and Algorithms

Convergencia en el tamaño del ensemble

Aunque con bagging no se generan modelos base totalmente independientes, sí que se observa que la precisión del ensemble tiende a mejorar a medida que crece el número de modelos, hasta un cierto punto de convergencia. (Ley de grandes números + distribución de bootstrap).

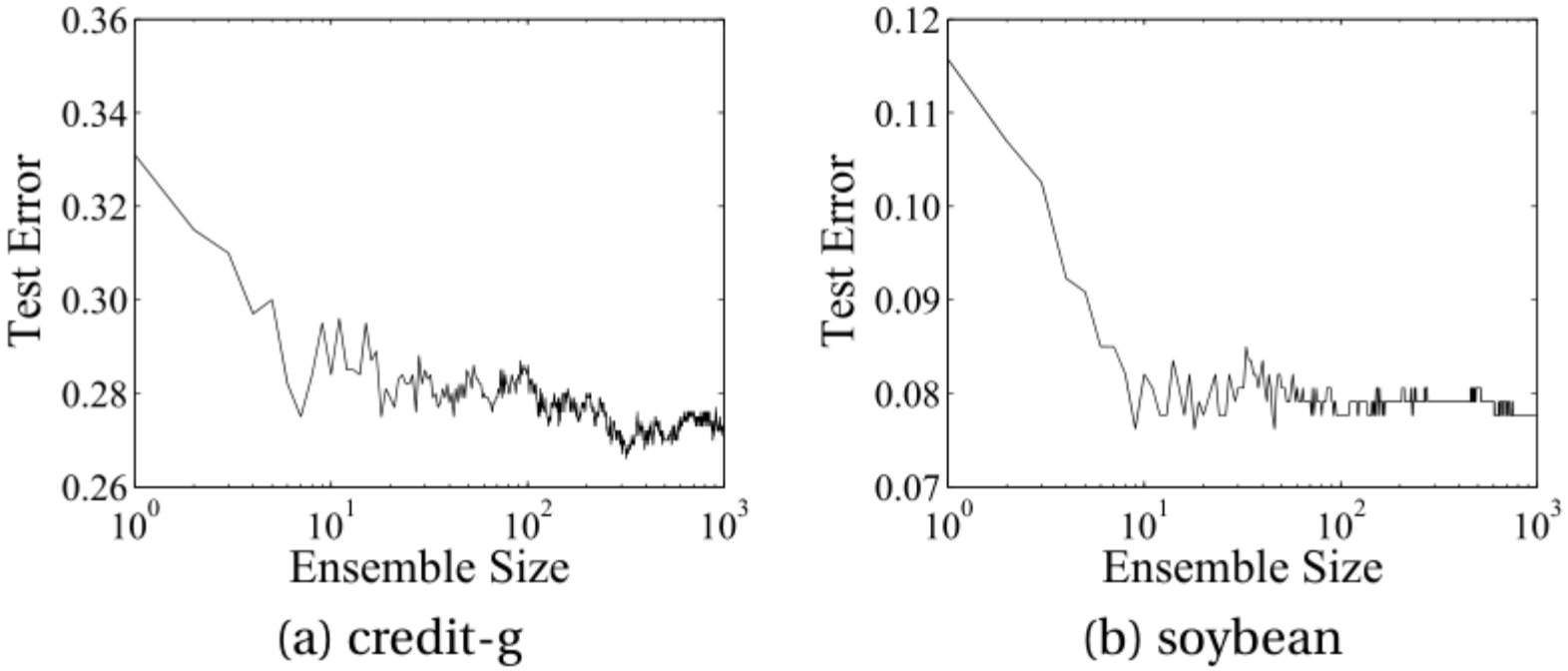


FIGURE 3.6: Impact of ensemble size on Bagging on two UCI data sets.

Zhou – Ensemble methods – Foundations and Algorithms

3. Fundamentos teóricos

Motivación teórica de bagging

Suponer f la función que intentamos aproximar con un ensemble H de clasificadores base h. Denotando D_{bs} la distribución de datos dada por bootstrap, el error del ensemble viene dado por:

Error del ensemble

$$(f(x) - H(x))^2 = (f(x) - \mathbb{E}_{\mathcal{D}_{bs}}[h(x)])^2 = (\mathbb{E}_{\mathcal{D}_{bs}}[f(x) - h(x)])^2$$

$$= \mathbb{E}_{\mathcal{D}_{bs}}\left[(f(x) - h(x))^2\right] - \mathrm{Var}_{\mathcal{D}_{bs}}[f(x) - h(x)]$$

$(\mathbb{E}[x^2] = \mathbb{E}[x]^2 + \mathrm{Var}[x])$

Error medio de cada modelo base

Varianza de errores en los modelos base (diversidad)

Por tanto el error del ensemble se reduce si

- ✓ El error de cada modelo base disminuye
- ✓ La diversidad entre los modelos base aumenta

(Moraleja: la diversidad es buena, pero solo si cada individuo hace cosas razonables)

Beneficios de la combinación

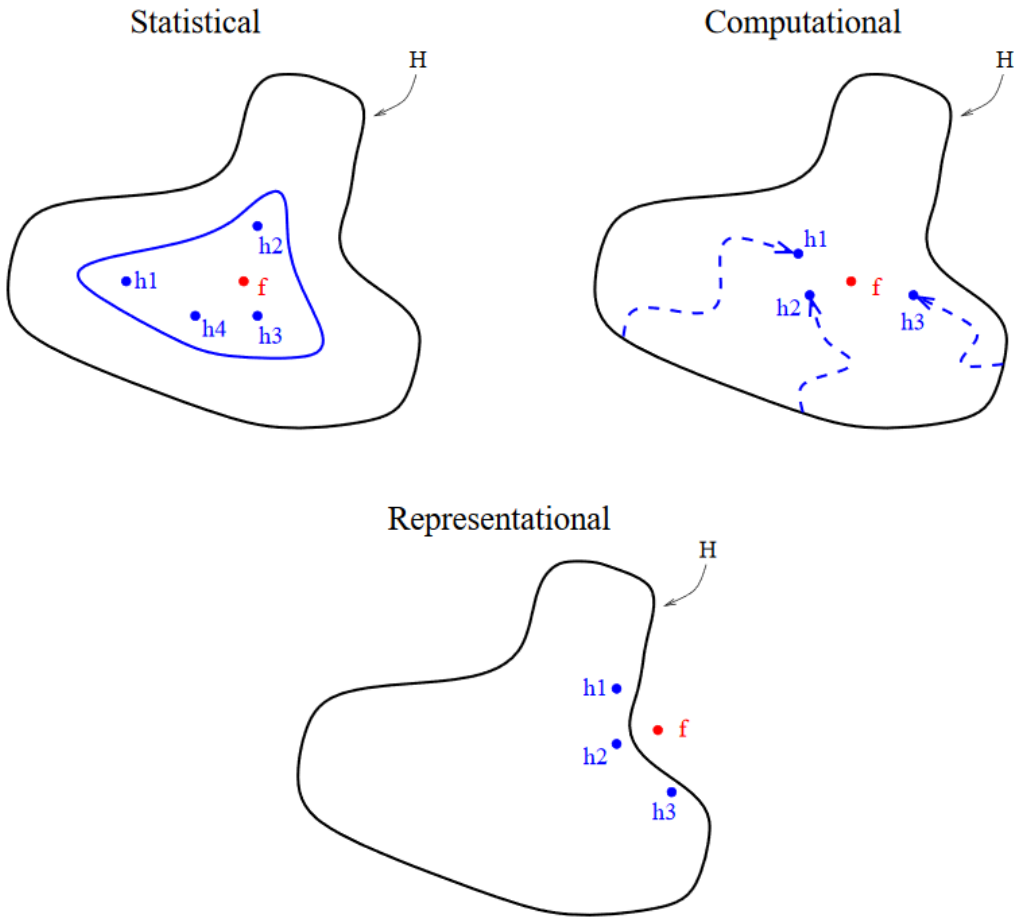


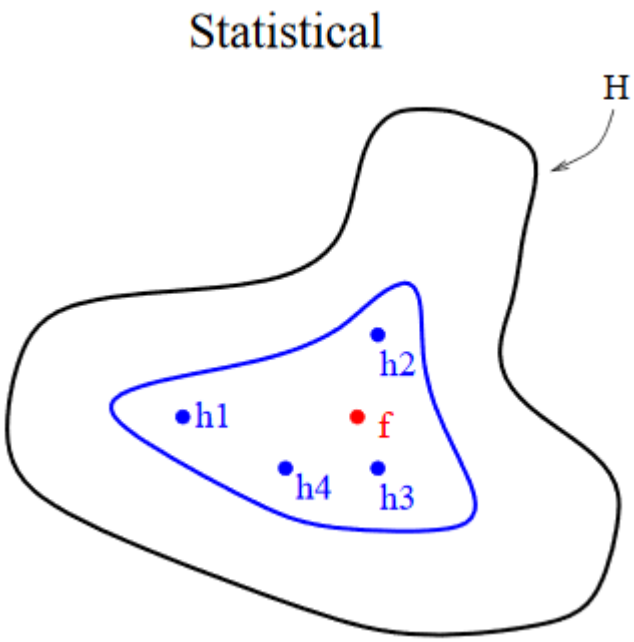
Fig. 2. Three fundamental reasons why an ensemble may work better than a single classifier

En un esquema general de combinación de modelos de aprendizaje automático, existen 3 razones principales por las que un ensemble puede ser superior a un modelo individual:

- ✓ Estadísticas
- ✓ Computacionales
- ✓ Representacionales

Dietterich – Ensemble methods in Machine Learning

Beneficios de la combinación

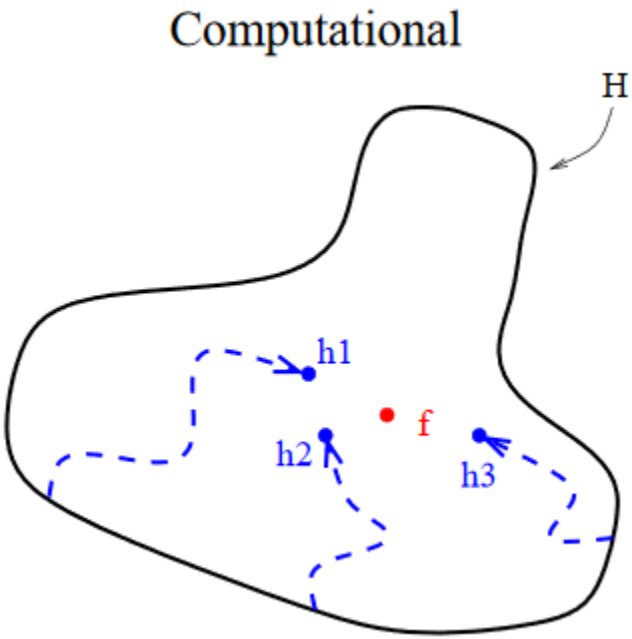


Razones estadísticas

El proceso de entrenamiento de un modelo puede verse como un búsqueda de la hipótesis que mejor explica los datos dentro de un espacio de posibles hipótesis H. Sin embargo, cuando los datos de entrenamiento son pocos en comparación al tamaño de H puede ocurrir que varias hipótesis (modelos) expliquen adecuadamente los datos. En este caso tomar varias de las hipótesis posibles $h_1, h_2, \ldots$ y promediarlas disminuye el riesgo de elegir una hipótesis muy alejada de la realidad f. (Epicuro)

Dietterich – Ensemble methods in Machine Learning

Beneficios de la combinación



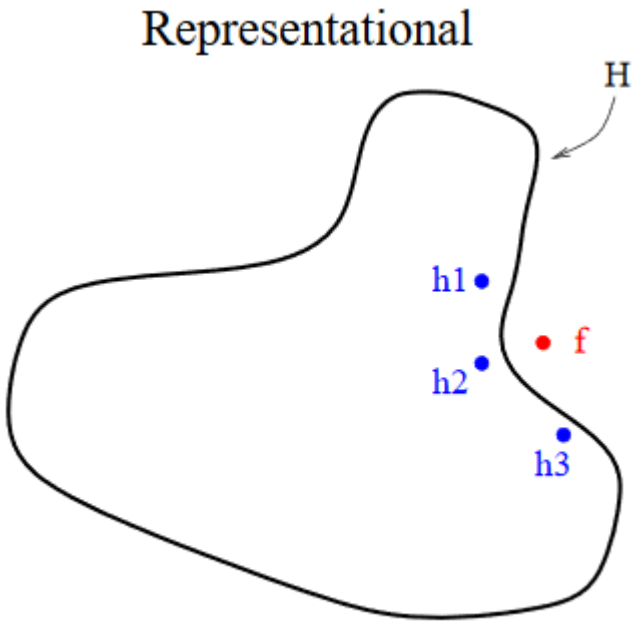
Razones computacionales

El entrenamiento en la mayoría de modelos no lineales implica resolver un problema de optimización no convexo, donde encontrar la configuración óptima de parámetros f puede ser un problema NP. El entrenar diversas variantes del modelo $h_1, h_2, \ldots$ en paralelo nos permite explorar diversas regiones del espacio de parámetros, teniendo así mayor probabilidad de alcanzar soluciones mejores.

Dietterich – Ensemble methods in Machine Learning

Beneficios de la combinación

Razones representacionales



En algunos problemas de aprendizaje la realidad de los datos f no puede representarse con la familia de modelos elegida H. Sin embargo, al combinar varios modelos de la misma clase en un ensemble se pueden lograr representaciones que con un solo modelo no serían posibles, logrando así extender el espacio de hipótesis.

4. Estrategias de combinación simples

Estrategias de combinación

Existente diferentes estrategias de combinación de las predicciones de un ensemble, en función del tipo del problema a resolver y de la naturaleza de los modelos base:

Regresión

- Media simple
- Media ponderada

Clasificación

- Votación por mayoría
- Votación por pluralidad
- Votación ponderada

Media
simple

$$H(x) = \frac{1}{T}\sum_{i=1}^{T} h_i(x)$$

Suponiendo que cada modelo base genera predicciones con cierto error ϵ en la forma $h_i(x) = f(x) + \epsilon_i(x)$, tenemos lo siguiente:

Error de un modelo base: $$\int (h_i(x) - f(x))^2 \, p(x)dx = \int \epsilon_i(x)^2 p(x)dx$$

Error medio entre modelos: $$\overline{err}(h) = \frac{1}{T}\sum_{i=1}^{T} \int \epsilon_i(x)^2 p(x)dx$$

Error medio de la combinación:

$$err(H) = \int \left(\frac{1}{T}\sum_{i=1}^{T} h_i(x) - f(x) \right)^2 p(x)dx = \int \left(\frac{1}{T}\sum_{i=1}^{T} \epsilon_i(x) \right)^2 p(x)dx$$

Media simple

Continuando con la media simple, supongamos además que los errores de cada modelo base tienen media cero y no están correlados, esto es:

$$\int \epsilon_i(x)p(x)dx = 0, \int \epsilon_i(x)\epsilon_j(x)p(x)dx = 0 \quad \text{for } i \neq j$$

Se tiene entonces que el error esperado del ensemble es:

$$err(H) = \int \left(\frac{1}{T} \sum_{i=1}^{T} \epsilon_i(x) \right)^2 p(x)dx$$

$$= \int \frac{1}{T^2} \left(\epsilon_1(x) \sum_{i=1}^{T} \epsilon_i(x) + \epsilon_2(x) \sum_{i=1}^{T} \epsilon_i(x) + \ldots + \epsilon_T(x) \sum_{i=1}^{T} \epsilon_i(x) \right) p(x)dx$$

$$= \int \frac{1}{T^2} \sum_{i=1}^{T} \epsilon_i(x)^2 p(x)dx = \frac{1}{T} \overline{err}(h)$$

$$\overline{err}(h) = \frac{1}{T} \sum_{i=1}^{T} \int \epsilon_i(x)^2 p(x)dx$$

Media
ponderada

$$H(x) = \sum_{i=1}^{T} w_i h_i(x), \quad w_i \geq 0, \sum_{i=1}^{T} w_i = 1$$

Se asignan pesos diferentes a cada miembro del ensemble, en función de su importancia. El error de esta combinación toma la forma:

$$err(H) = \int \left(\sum_{i=1}^{T} w_i h_i(x) - f(x) \right)^2 p(x)dx = \int \left(\sum_{i=1}^{T} w_i (f(x) + \epsilon_i(x)) - f(x) \right)^2 p(x)dx$$

$$= \int \left(\sum_{i=1}^{T} w_i \epsilon_i(x) \right)^2 p(x)dx = \int \sum_{i=1}^{T} \sum_{j=1}^{T} w_i w_j \epsilon_i(x) \epsilon_j(x) p(x)dx = \sum_{i=1}^{T} \sum_{j=1}^{T} w_i w_j C_{ij}$$

con correlaciones de errores $C_{ij} = \int \epsilon_i(x)\epsilon_j(x)p(x)dx$

Media ponderada

Los pesos óptimos para la media ponderada pueden obtenerse como la combinación que minimiza el error del ensemble

$$w^* = \min_w err(H) = \min_w \sum_{i=1}^{T}\sum_{j=1}^{T} w_i w_j C_{ij} \quad \text{s.t. } w_i \geq 0, \sum_{i=1}^{T} w_i = 1$$

Como resultado aproximado, se ignora la restricción de positividad, y aplicando coeficientes de Lagrange:

$$w^* = \arg\min_w \max_\lambda w^T C w - \lambda(1 - e^T w)$$

(e = vector de todo valores 1)

$$\nabla_{(w,\lambda)} w^T C w - \lambda(1 - e^T w) = 0 \implies \begin{cases} 2Cw + \lambda e = 0 \implies w = -\frac{\lambda}{2} C^{-1} e \\ -1 + e^T w = 0 \implies e^T w = 1 \end{cases}$$

Se obtiene un sistema de ecuaciones que incluye la inversa de C, la cual es necesario que exista.

Media ponderada

$$\begin{cases} w = -\frac{\lambda}{2} C^{-1} e \\ e^T w = 1 \end{cases}$$

Resolver para λ

$$w = -\frac{\lambda}{2} C^{-1} e$$

$$\implies e^T w = -\frac{\lambda}{2} e^T C^{-1} e$$

$$\implies \lambda = -\frac{2}{e^T C^{-1} e}$$

Resolver para w

$$w = -\frac{\lambda}{2} C^{-1} e$$

$$\implies w = -\frac{-\frac{2}{e^T C^{-1} e}}{2} C^{-1} e$$

$$\implies w = \frac{C^{-1} e}{e^T C^{-1} e}$$

Por tanto los pesos óptimos son: $w_i^* = \frac{\sum_{j=1}^{T} C_{ij}^{-1}}{\sum_{k=1}^{T} \sum_{j=1}^{T} C_{kj}^{-1}}$

En la práctica, esta forma de combinación es superior a la media simple solo si los modelos base tienen precisiones muy diferentes entre sí.

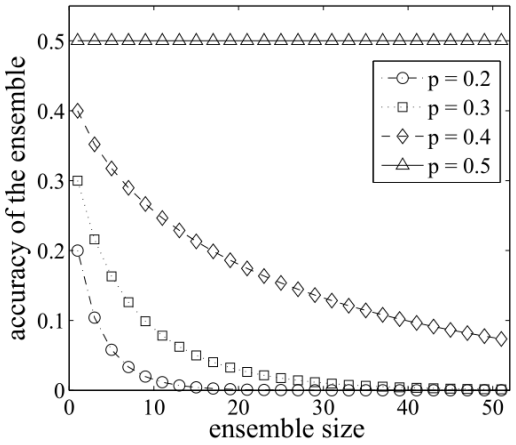
Votación
por mayoría

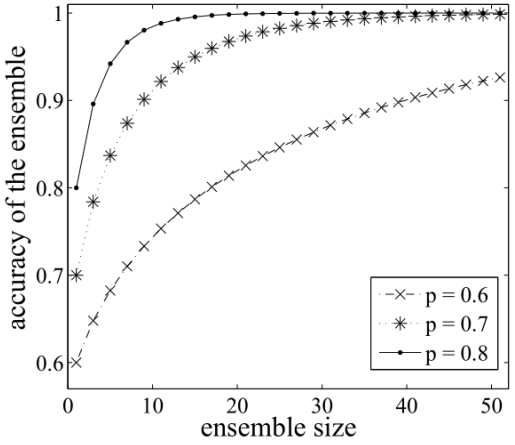
$$H(x) = \begin{cases} c_j & \text{if } \sum_{i=1}^{T} h_i^j(x) > \frac{1}{2} \sum_{k=1}^{l} \sum_{i=1}^{T} h_i^k(x) \\ \text{rejection} & \text{otherwise} \end{cases}$$

En problemas de clasificación, cada modelo base h_i es un clasificador que para un patrón x genera predicciones h_i^j, indicando la probabilidad de que x pertenezca a la clase c_j según ese clasificador.

El ensemble predice la clase j solo si al menos la mitad de las probabilidades de clase de todo el ensemble se han asignado a la clase j. En otro caso se genera un valor especial *rejection* que indica que el ensemble no tiene suficiente evidencia para tomar una decisión.

Votación por mayoría





$$\sum_{k=\lfloor T/2+1 \rfloor}^{T} \binom{T}{k} p^k (1-p)^{T-k}$$

FIGURE 4.2: Ensemble accuracy of majority voting of T independent classifiers with accuracy p for binary classification.

Si la probabilidad de acierto p de cada clasificador base es > 0.5, y los clasificadores son independientes, la precisión del ensemble tiende a 1 cuando el tamaño del ensemble tiende a infinito. Si $p < 0.5$, la precisión del ensemble tiende a 0.

(Moraleja: el voto por mayoría solo funciona si cada votante es suficientemente inteligente para escoger la opción correcta con mayor probabilidad.)

Votación
por pluralidad

$$H(x) = c_{\arg\max_j \sum_{i=1}^{T} h_i^j(x)}$$

Modificación de la votación por mayoría en la que la clase con mayor probabilidad acumulada es la predicha por el ensemble, incluso si no alcanza la mayoría.

No existe opción de rechazo.

En problemas de clasificación binaria es un sistema idéntico a la votación por mayoría.

Votación ponderada

$$H(x) = c_{\arg\max_j \sum_{i=1}^{T} w_i h_i^j(x)}$$

Versión de la votación por pluralidad en la que se incluyen pesos en función de la importancia de cada clasificador.

Los pesos óptimos de combinación pueden obtenerse usando el teorema de Bayes. Tomando l como el vector de las clases predichas por cada clasificador base, y suponiendo que la clase correcta a predecir es c_j, la combinación óptima debe seguir:

$$H^j(x)^* = P(c_j|l) = \frac{P(c_j)P(l|c_j)}{P(l)} \propto \log\left(P(c_j)P(l|c_j)\right)$$

Votación ponderada

Asumiendo independencia condicionada entre los clasificadores,

$$H^j(x)^* \propto \log\left(P(c_j)P(l|c_j)\right) = \log P(c_j) + \log \prod_{i=1}^{T} P(l_i|c_j)$$

Separar en clasificaciones correctas/incorrectas

Clasificadores que aciertan

Clasificadores que fallan

$$= \log P(c_j) + \log \left(\prod_{i=1, l_i=c_j}^{T} P(l_i|c_j) \prod_{i=1, l_i \neq c_j}^{T} P(l_i|c_j) \right)$$

Prob. clasificación correcta es p_i

$$= \log P(c_j) + \log \left(\prod_{i=1, l_i=c_j}^{T} p_i \prod_{i=1, l_i \neq c_j}^{T} (1-p_i) \right)$$

Reordenar términos

$$= \log P(c_j) + \sum_{i=1, l_i=c_j}^{T} \log \frac{p_i}{1-p_i} + \sum_{i=1}^{T} \log(1-p_i)$$

No depende de la clase

$$\propto \log P(c_j) + \sum_{i=1}^{T} h_i^j(x) \log \frac{p_i}{1-p_i}$$

$h_i^j = (l_i = c_j)$

Votación ponderada

Hemos obtenido que

$$H^j(x)^* \propto \log P(c_j) + \sum_{i=1}^{T} h_i^j(x) \log \frac{p_i}{1-p_i}$$

Solo depende del problema

Indica la forma adecuada de ponderar los clasificadores

Por tanto los pesos óptimos para cada modelo deben escogerse de forma proporcional a su probabilidad de acierto:

$$w_i \propto \log \frac{p_i}{1-p_i}$$

(Obs: si $p_i < 0.5$, su w_i es negativo, ya que se acierta más contradiciendo su predicción)

(Moraleja: da mejores resultados que valgan más los votos de quienes tienen más probabilidades de elegir la opción correcta)

5. Estrategias de combinación avanzadas

Estrategias de combinación avanzadas

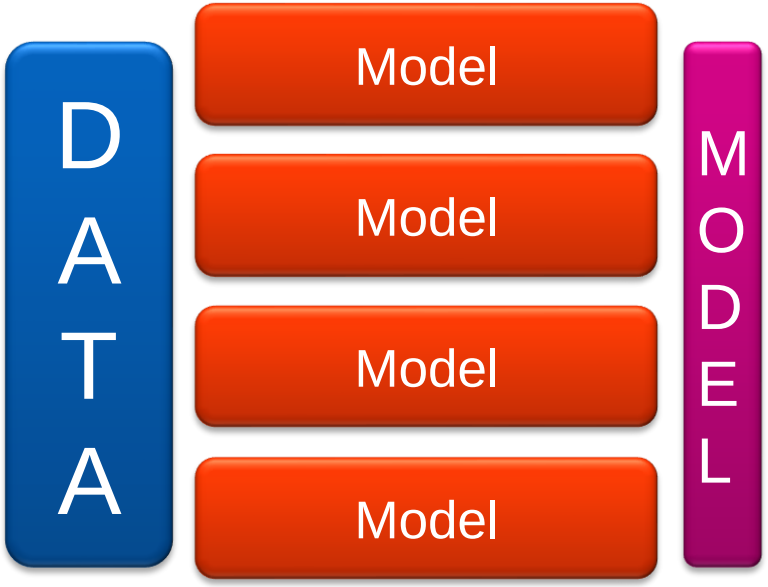
En el caso de contar con modelos muy diferentes, y de querer invertir tiempo en un ensemble más preciso, pueden utilizarse la siguientes estrategias de combinación:

Stacking

Mezcla de expertos

Stacking

Podemos considerar los pesos w como los parámetros de otro modelo de clasificación (**meta-learner**) que intenta combinar las predicciones de cada modelo base para obtener un consenso.



$$H(x) = m(h_1(x), \ldots, h_T(x))$$

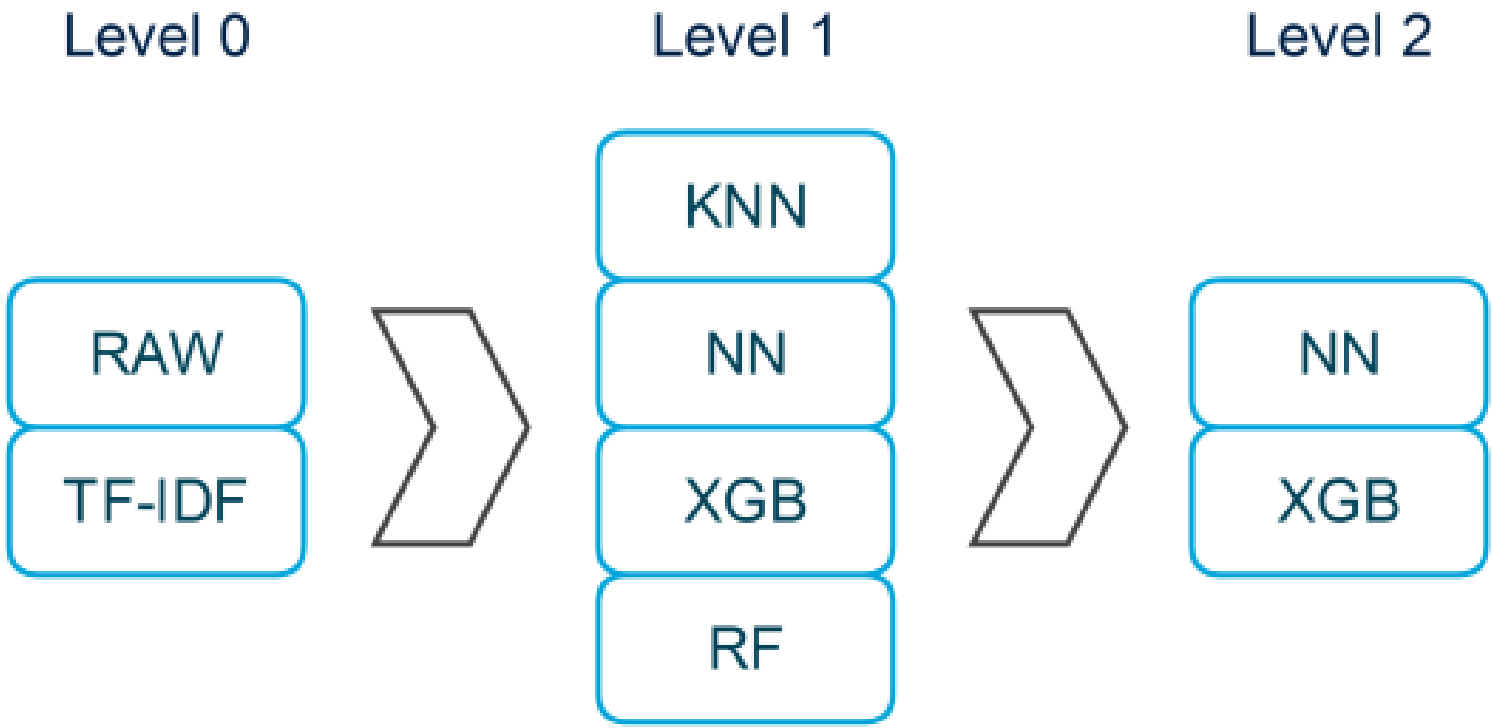
Entradas del meta-learner: **predicciones** de cada uno de los modelos base (+ datos de entrada)

Targets del meta-learner: **targets** del problema original

Puede usarse un modelo lineal como **regresión logística** o un modelo más complejo como otro ensemble. Pero mientras más complejo el modelo, más probabilidades de sobreajuste.

Stacking

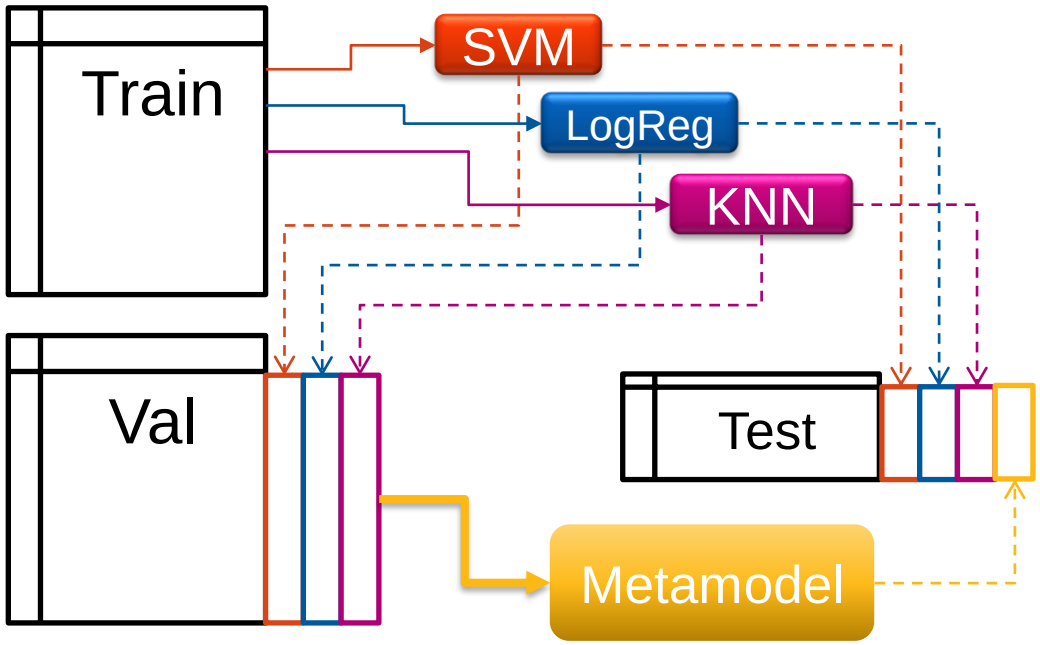
Otto Product Classification Kaggle Competition: 2nd place



http://blog.kaggle.com/2015/06/09/otto-product-classification-winners-interview-2nd-place-alexander-guschin/

Stacking honesto

Si tenemos muchos datos disponibles, una forma de evitar el sobreajuste al hacer stacking es usar una parte de los datos para entrenar los modelos base, y otra parte (validación) para entrenar el meta-modelo.



Entrenamiento

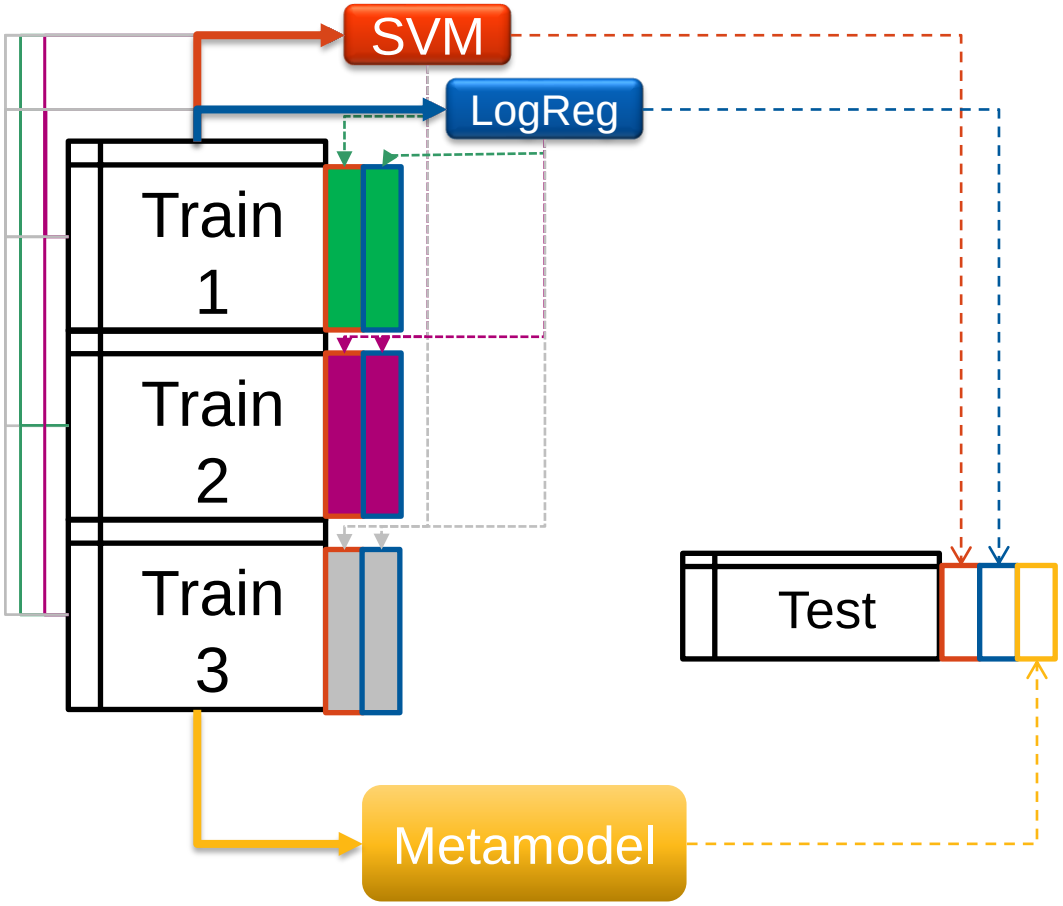
1. Entrenar cada modelo base con los datos de entrenamiento (+CV).
2. Hacer predicciones con cada modelo base para los datos de validación.
3. Entrenar el metamodelo con los datos de validación (+CV), que incluyen predicciones de modelos base.

Test

1. Hacer predicciones con cada modelo base para los datos de test.
2. Hacer predicciones con el metamodelo para los datos de test, usando predicciones de modelos base.

Stacking honesto con pocos datos

Aún si tenemos pocos datos podemos hacer un stacking que prevenga el overfitting partiendo el conjunto de entrenamiento en K hojas de validación, y haciendo un uso cuidadoso de ellas:



Entrenamiento

1. Para cada modelo base y para cada hoja k: entrenar el modelo con las otras K-1 hojas, y generar predicciones para la hoja k.

2. Entrenar el metamodelo con todo el conjunto de entrenamiento, usando también las predicciones de los modelos base.

3. Reentrenar cada modelo base con todo el conjunto de entrenamiento, sin usar las predicciones de los modelos base.

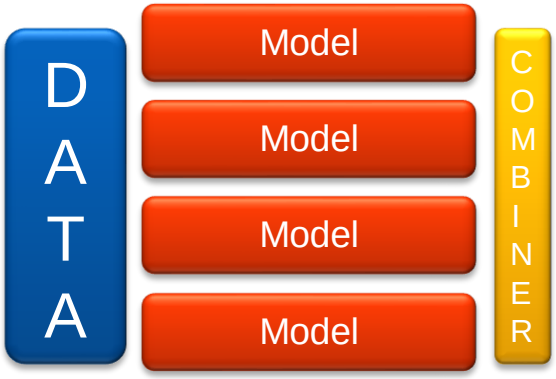
Test

1. Hacer predicciones con cada modelo base para los datos de test.

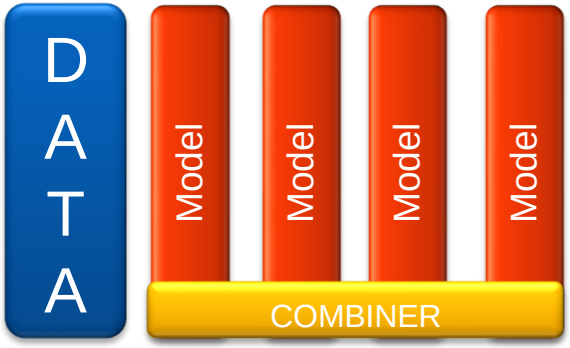
2. Hacer predicciones con el metamodelo para los datos de test, usando predicciones de modelos base.

Mezcla
de expertos

Los ensembles por lo general fomentan que cada modelo colabore en mejorar la capacidad predictiva global mediante dos estrategias:



Cada modelo intenta de forma independiente obtener la mejor predicción posible, y las predicciones se combinan colaborativamente.

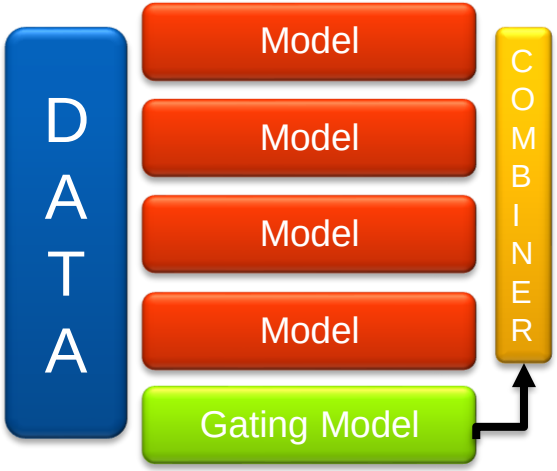


Cada modelo intenta corregir los errores que está cometiendo el ensemble hasta ese momento, colaborando así a una mejora global.

Existe una tercera vía: que cada modelo base se especialice en tratar cierto tipo de patrones.

Mezcla de expertos

La mezcla de expertos es un tipo de ensemble que fomenta la especialización en lugar de la colaboración.



$$H(x) = \sum_{i=1}^{T} \pi_i(x) \cdot h_i(x)$$

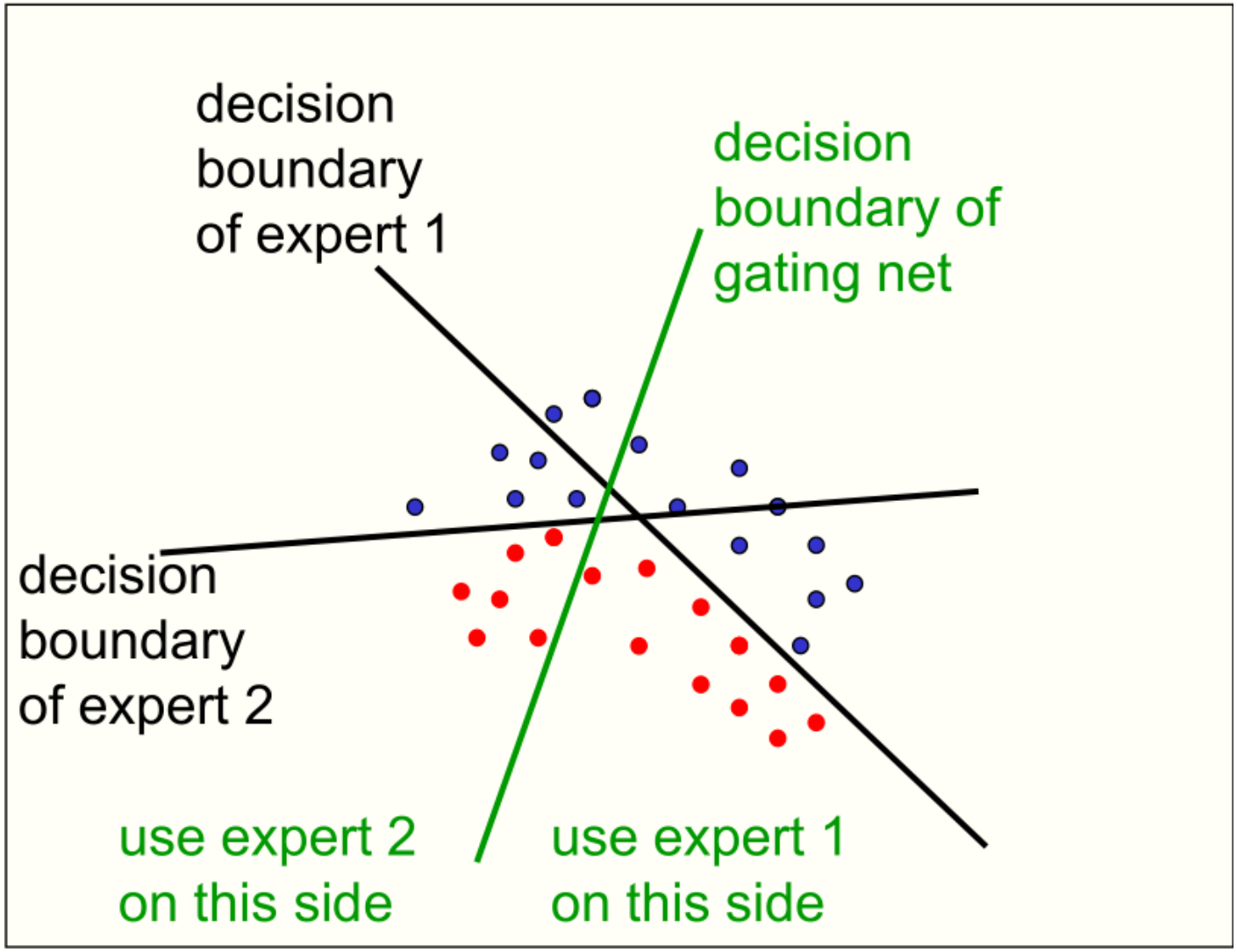
$\pi_i(x)$ es un modelo de puerta (gating) que decide para cada dato x la relevancia que tiene el modelo h_i en la combinación.

Se entrena alternando entre optimizar los modelos base h_i y el modelo de gating π, con una función de error que promueve la especialización:

$$E(x, y) = \sum_{i=1}^{T} \pi_i(x)(y - h_i(x))^2$$

Jacobs et al – Adaptive Mixtures of Local Experts

Mezcla de expertos

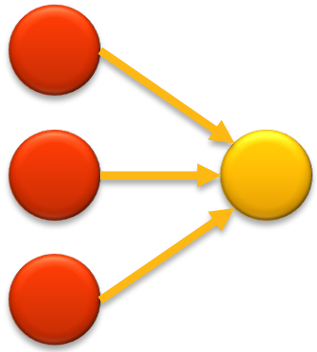


*. Relación con otros modelos de aprendizaje automático

Perceptrón multicapa

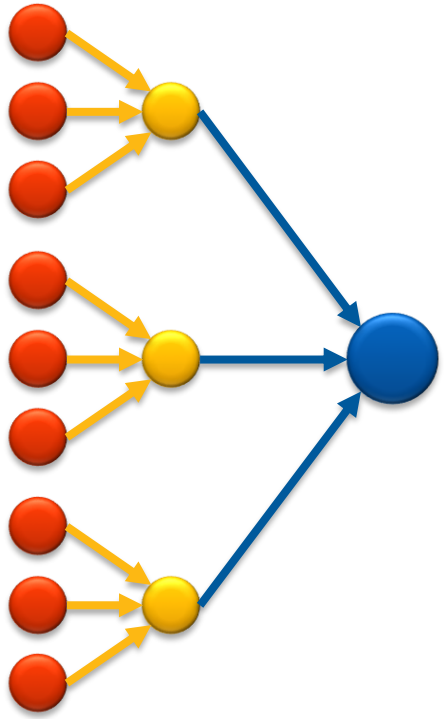
Un perceptrón o red neuronal sin capas ocultas puede definirse como un clasificador en la forma:

$$f(x) = \text{sign}(a^T x + b)$$



Un ensemble de perceptrones toma la forma:

$$H(x) = \sum_{i=1}^{T} w_i \, \text{sign}(a_i^T x + b_i)$$

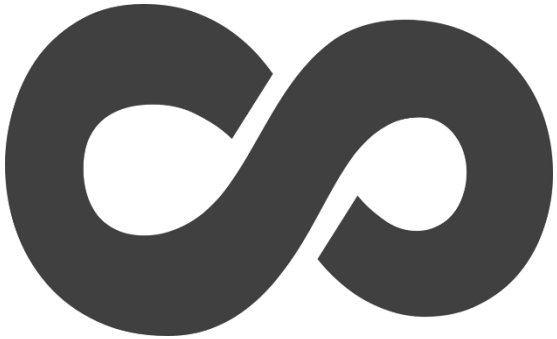


… lo cual es un perceptrón con una capa oculta que emplea la función signo como activación.

Un perceptrón con múltiples capas ocultas es un ensemble de ensembles de ensembles … de ensembles de perceptrones.

Ensemble infinito

Obs: si el ensemble tiende a mejorar cuantos más modelos base contiene… ¿por qué no tratar de hacer un ensemble con infinitos modelos?



Suponer $H = \{h_\alpha : \alpha \in C\}$ espacio de hipótesis, C un espacio de medida. Un ensemble infinito toma la forma:

$$g(x) = \text{sign}\left(\int_C w(\alpha) h_\alpha(x) d\alpha + b\right)$$

Que esencialmente es una suma infinita (integral) de las contribuciones de cada modelo base h_α ponderadas por pesos $w(\alpha)$.

x En la práctica no podemos calcular ni almacenar infinitos modelos.
✓ Existe un truco para hacerlo

Ensemble infinito

Se define la función kernel

$$K_{\mathcal{H},r}(x_i, x_j) = \int_{\mathcal{C}} \Phi_{x_i}(\alpha)\Phi_{x_j}(\alpha)d\alpha \qquad \Phi_x(\alpha) = r(\alpha)h_\alpha(x)$$

Con $r: C \to \mathbb{R}^+$ tal que la integral existe para todo (x_i, x_j) en el conjunto de modelos base H escogidos. Se puede demostrar que este kernel cumple la condición de Mercer, por lo que es un kernel que representa un producto interno en un espacio de infinitas dimensiones.

Por tanto un ensemble infinito es una SVM entrenada con un kernel especial que representa la suma de infinitos modelos base.

Entrenamiento

$$\min_{\alpha \in \mathbb{R}^N} \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j) - \sum_{i=1}^{N} \alpha_i$$

$$\text{s.t.} \begin{cases} 0 \leq \alpha \leq C \\ y^T \alpha = 0 \end{cases}$$

Predicción

$$g(x) = \text{sign}\left(\sum_{i=1}^{N} y_i \alpha_i K(x_i, x) + b \right)$$

Ensemble infinito: ejemplo

Tomar como espacio de hipótesis H el conjunto de todos los árboles de clasificación con un único corte (a.k.a. stumps).

Un stump puede definirse mediante la variable d por la que corta, el valor $\alpha \in [L_d, R_d]$ del corte en esa variable, y el signo de la predicción $q \in \{-1, +1\}$:

$$s_{q,d,\alpha}(x) = q \cdot \text{sign}((x)_d - \alpha)$$

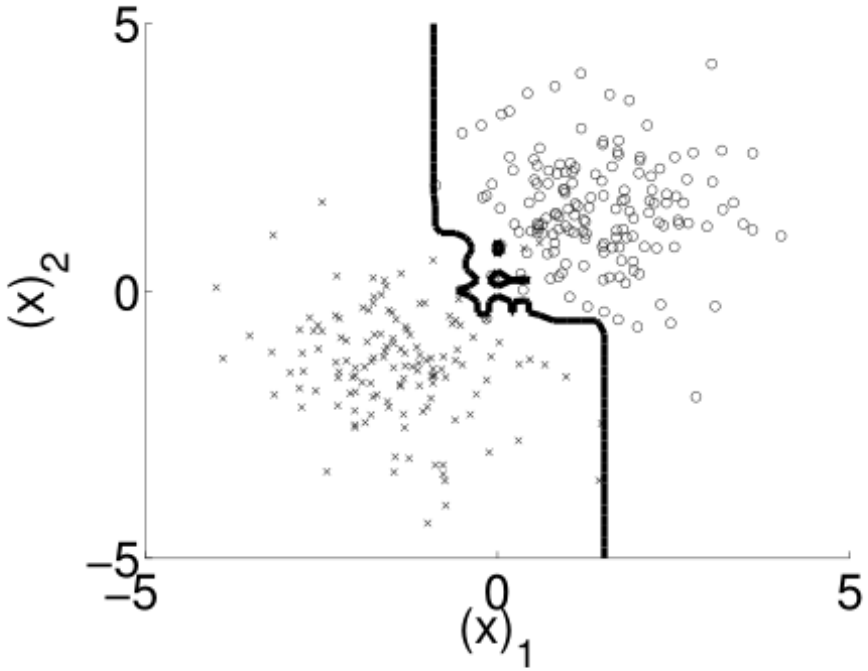
Puede demostrarse que el kernel que representa un ensemble infinito de stumps toma la forma:

$$K(x, x') = \Delta - \|x - x'\|_1 \qquad \text{con} \quad \Delta = \frac{1}{2}\sum_{d=1}^{D}(R_d - L_d)$$

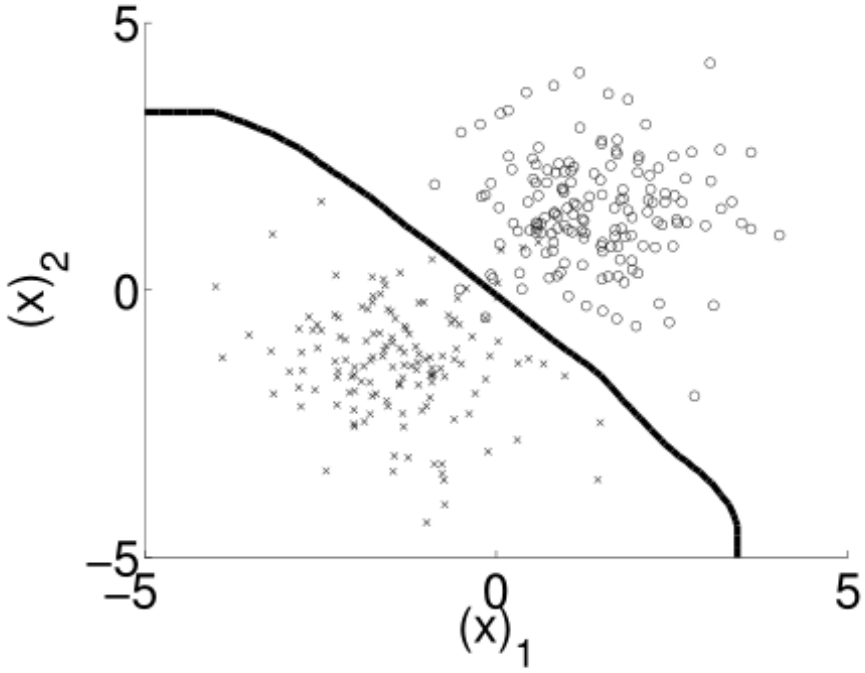
($||\cdot||_1$ es la norma en valor absoluto)

Lin & Li – Support Vector Machine for Infinite Ensemble Learning

Ensemble infinito: ejemplo



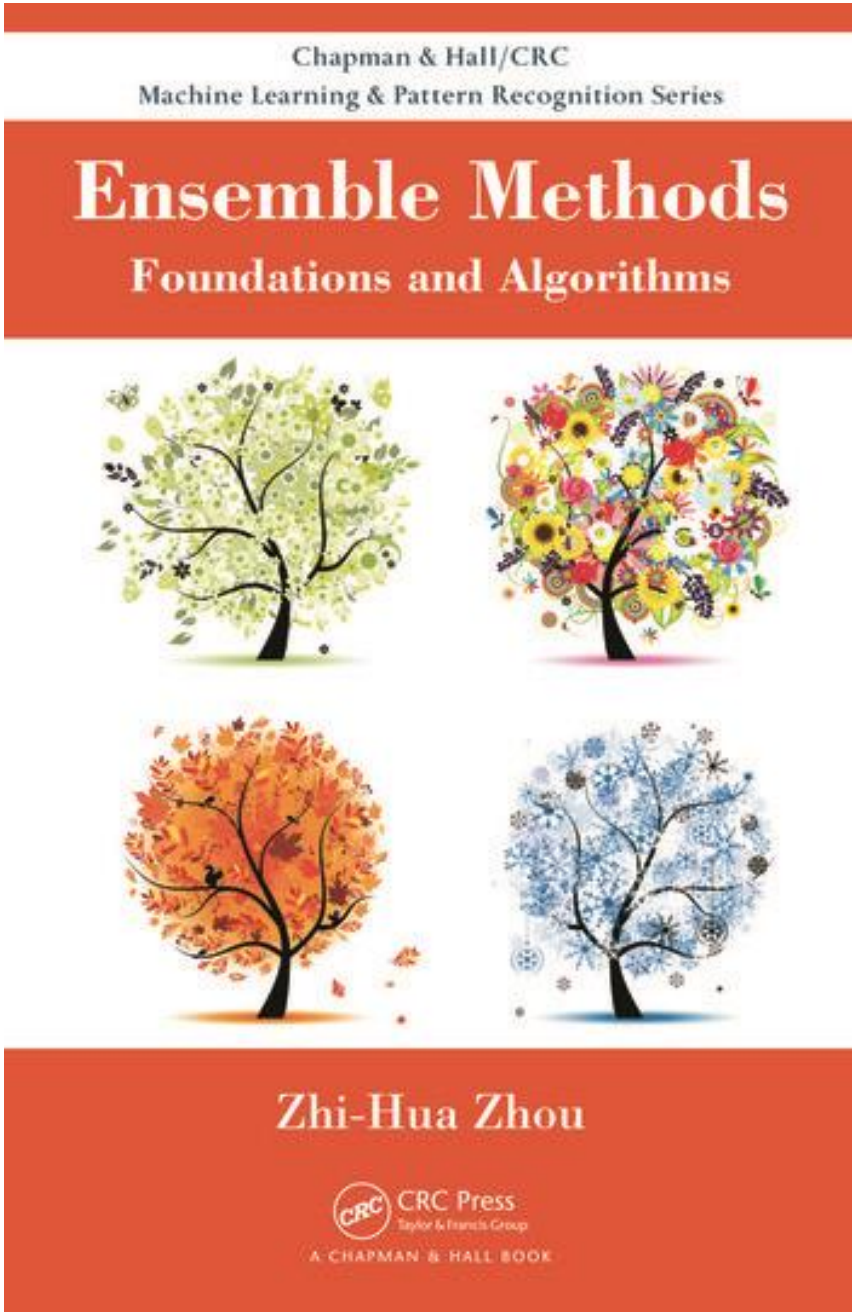
(c) AdaBoost-Stump



(a) SVM-Stump

Lin & Li – Support Vector Machine for Infinite Ensemble Learning

8. Bibliografía


Chapman & Hall/CRC
Machine Learning & Pattern Recognition Series
Ensemble Methods
Foundations and Algorithms
Zhi-Hua Zhou
CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Afi Escuela