Boosting y Clasificación Funcional

Rogelio Ramos Quiroga

rramosq@cimat.mx

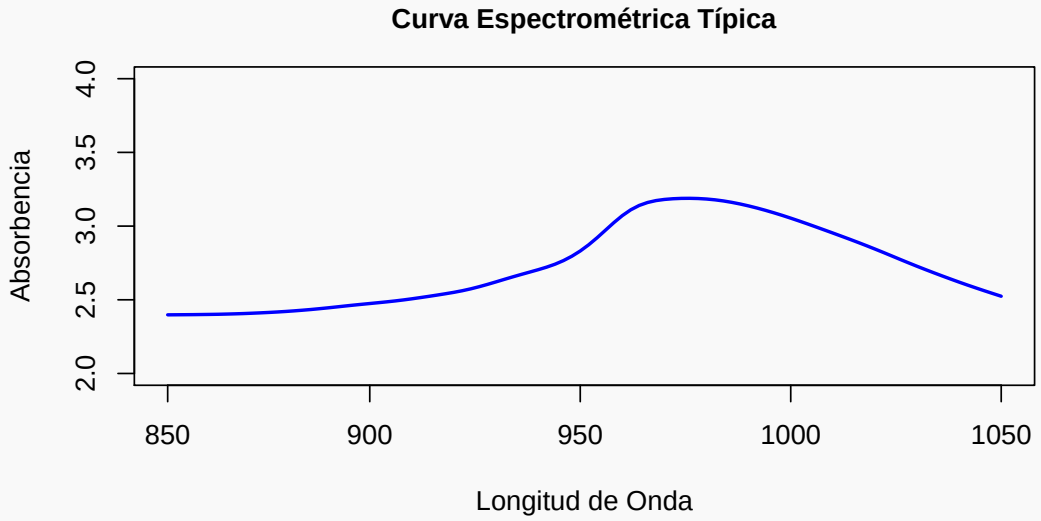
SEMINARIO ALEATORIO DEL ITAM
29 de Abril de 2016

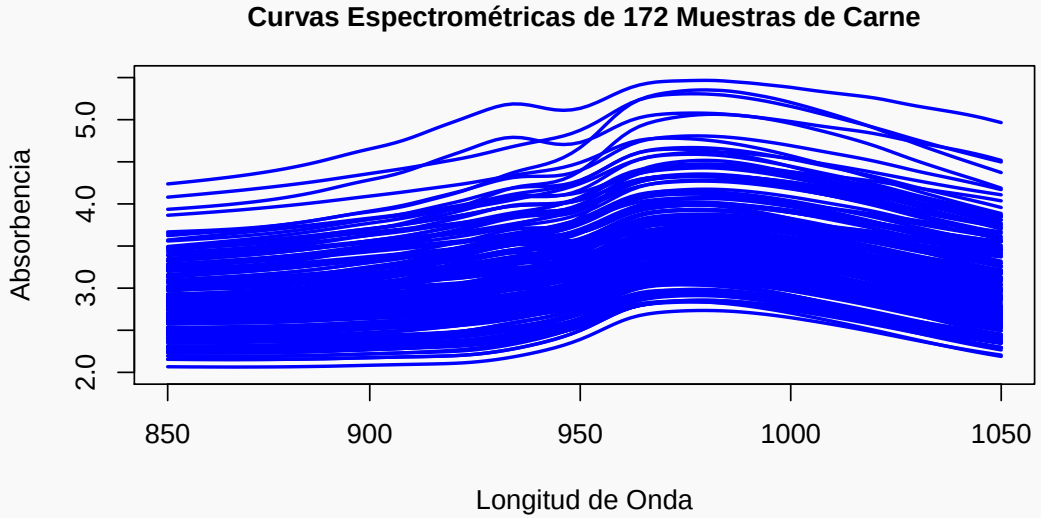
Boosting y Clasificación Funcional

- Análisis de datos funcionales
- Problemas de clasificación
- Boosting
- Adaptaciones al caso funcional

Análisis de Datos Funcionales

Estructura de Datos





(Datos: Ferraty & Vieu (2006))

Estructura de Datos

Características

- cantidad
- frecuencia (resolución)
- similaridad
- suavidad

Datos Funcionales

- FDA se refiere usualmente a aquellos problemas estadísticos en donde los datos consisten de una muestra de n funciones $x_1 = x_1(t), \ldots, x_n = x_n(t)$.
- **"Los datos funcionales son datos multivariados con un ordenamiento en las dimensiones"**. Muller (2006).

Aplicaciones importantes en:

- electrocardiografía, electroencefalografía
- monitoreo intensivo en sistemas de manufactura
- quimiometría
- espectrografía en general (astronomía, ciencias de materiales)

Algunos Problemas Básicos en Estadística

Tenemos observaciones $y_1, y_2, \cdots, y_n$, de alguna población P.

1. P es conceptualizada en Estadística como una distribución de probabilidad: ¿Quién es P? y ¿Cómo se estima?
2. Si observo $(x_1, y_1), \cdots, (x_n, y_n)$, ¿Cómo determino $P_{Y|X}$?

   (En realidad, observamos $\{(y_i|x_i)\}_i$, en el contexto de estudios controlados)
3. Si $y_1, \cdots, y_n$ son objetos complejos, ¿Es posible encontrar proxys de las y's, $z_1, \cdots, z_n$, en un espacio más sencillo (e.g. $\mathbb{R}^k$), tal que se tenga un mínimo de pérdida de información?
4. Si observo y, ¿Cómo determino la razonabilidad de que $y \in P$?

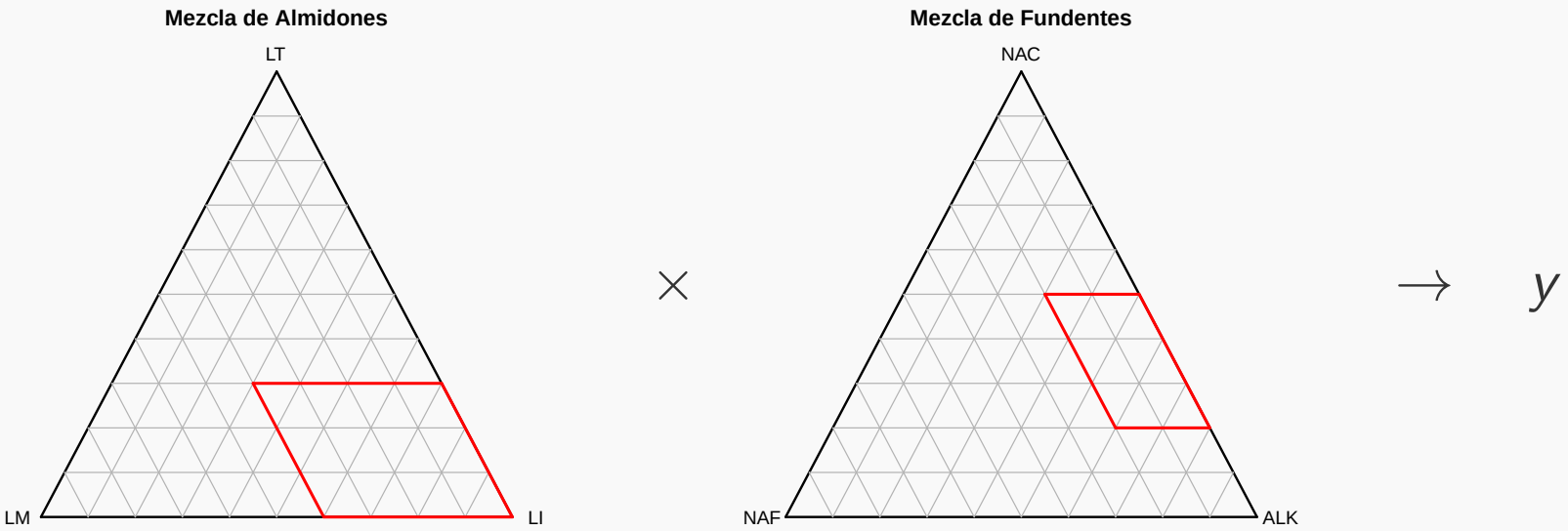
Problema 1

- Tenemos observaciones $y_1, y_2, \cdots, y_n$, de alguna población P. ¿Quién es la distribución P? y ¿Cómo se estima?
- Esto representa un problema básico en el análisis de datos funcionales: No hay distribuciones P's estándar.
- Procesos Gaussianos (Teorema Central del Límite)
- Expansión Karhunen-Loeve, $y(t) = \sum_{j=1}^{\infty} Z_j e_j(t)$, donde las Z_j's y e_j's, estan relacionadas con el operador de covarianza de y

Delaigle & Hall (2010) Defining probability density for a distribution for random functions, *AOS*.

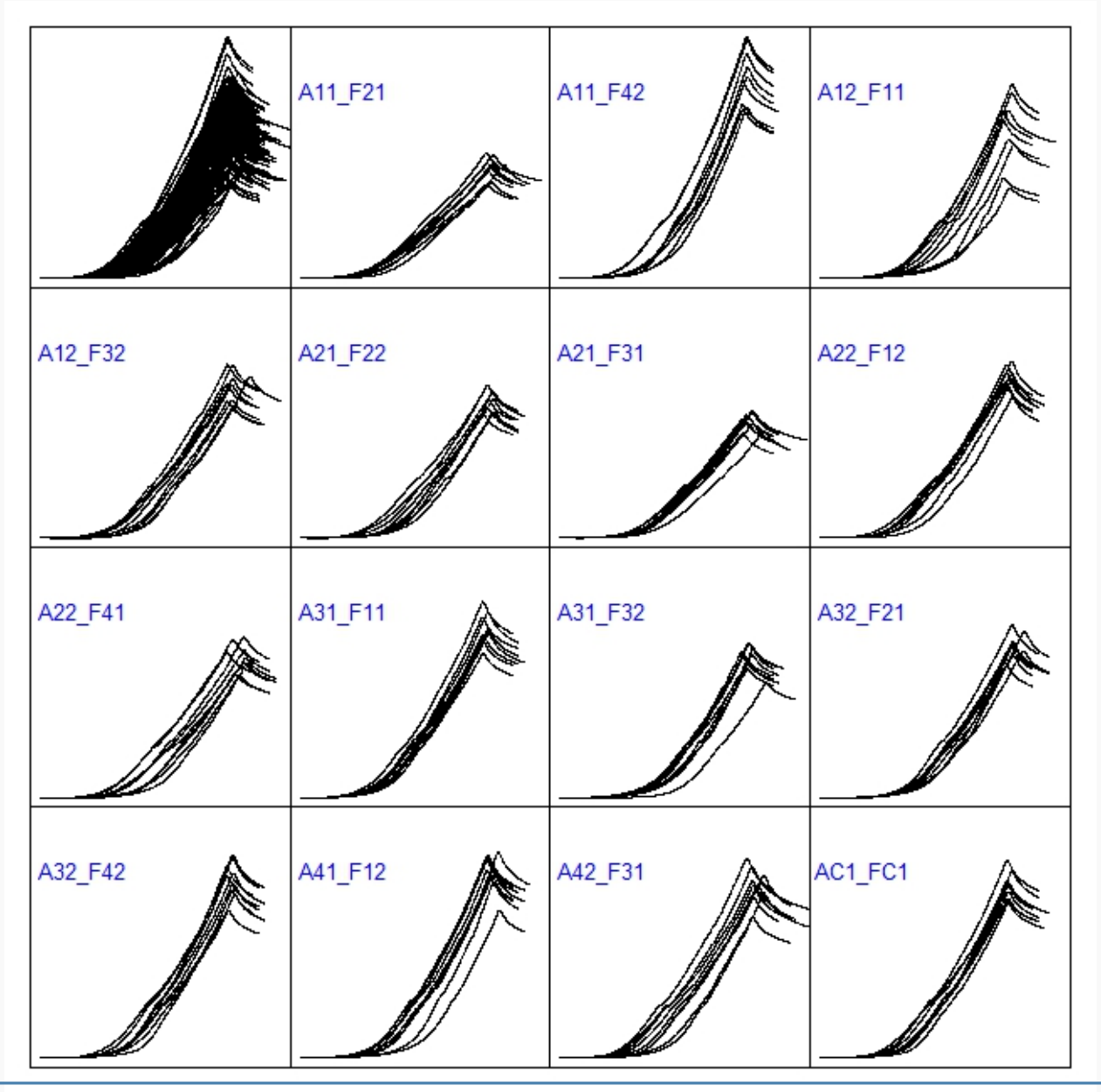
Problema 2

- Si observo $(y_1|x_1), \cdots, (y_n|x_n)$, ¿Cómo determino $P_{Y|X}$?
- Regresión funcional



- Ramsay & Silverman (1995) Functional Data Analysis. Springer

· · · Problema 2


A11_F21
A11_F42
A12_F11
A12_F32
A21_F22
A21_F31
A22_F12
A22_F41
A31_F11
A31_F32
A32_F21
A32_F42
A41_F12
A42_F31
AC1_FC1

Problema 3

- Si $y_1, \cdots, y_n$ son objetos complejos, ¿Es posible encontrar proxys de las y's, $z_1, \cdots, z_n$, en un espacio más sencillo (e.g. $\mathbb{R}^k$), tal que se tenga un mínimo de pérdida de información?
- Análisis de Componentes Principales

$$y_i \leftrightarrow (\langle \alpha_1, y_i \rangle, \cdots, \langle \alpha_k, y_i \rangle)$$

  donde las α_j's son ciertas "direcciones" (funciones) principales.
- Ramsay & Dalzell (1991) Some tools for functional data analysis. *JRSS-B*, 53, 539-572

Problema 4

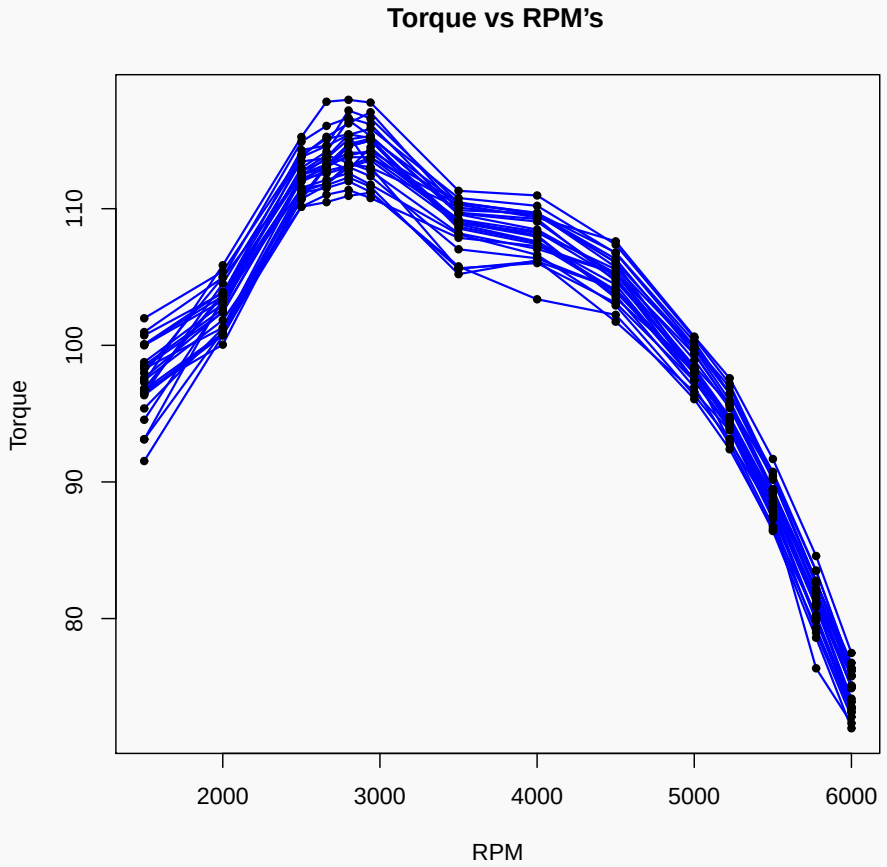
- Tenemos observaciones $y_1, y_2, \cdots, y_n$, de alguna población P. Si observo un nuevo dato y, ¿Cómo determino la razonabilidad de que $y \in P$?
- Problemas de Clasificación
  - Nociones de distancia
  - Medidas de profundidad
  - Esimadores "plug in" del Bayesiano óptimo
  - k-vecinos más cercanos

Baíllo, Cuevas & Fraiman (2011) Classification methods for functional data. In: Ferraty & Romain (Eds.), The Oxford Handbook of Functional Data Analysis.

Clasificación

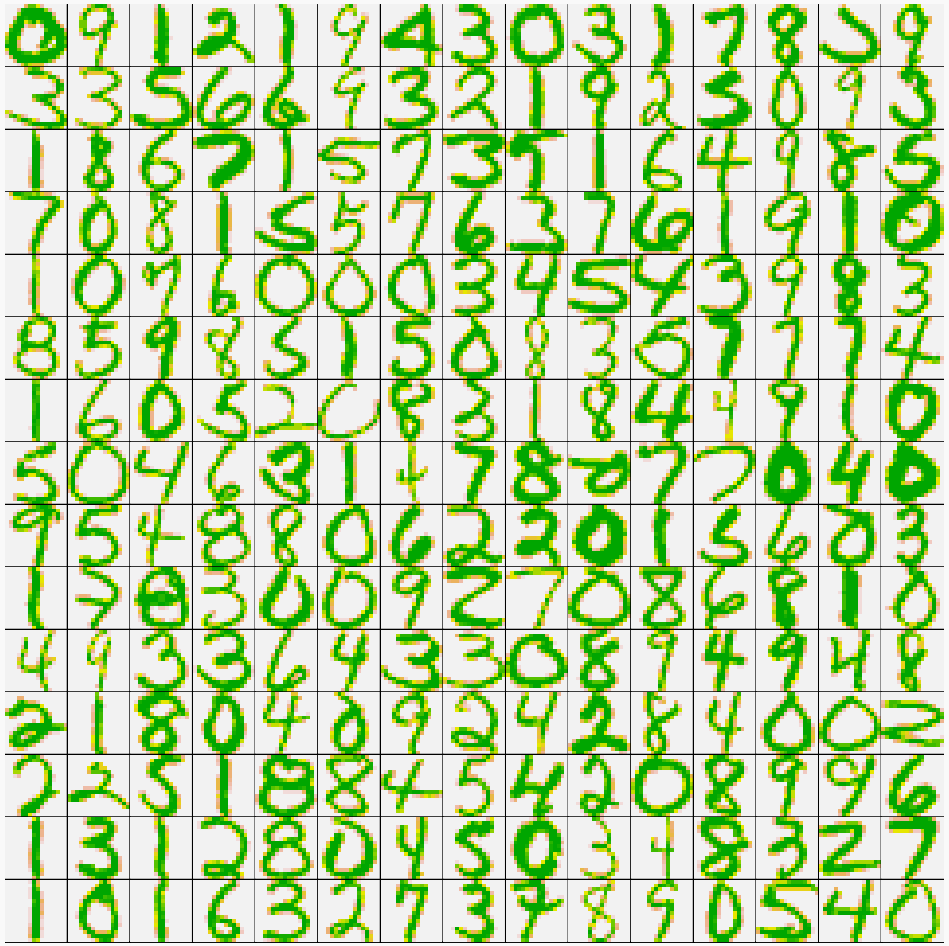
El problema de Clasificación

- primero, ¿cómo caracterizar una población "normal"?
- segundo, dado el perfil de un motor dado, ¿cómo decido si es normal?



Problemas de Clasificación

Problema: En base a la matriz 16×16 de intensidades de pixeles, **clasificar** cada imágen como $0, 1, \cdots, 9$.



Problemas de Clasificación

Queremos clasificar **objectos** en **categorías**. Los objectos, x, se describen mediante un vector de atributos.

$$\begin{aligned} x &= (x_1, \cdots, x_p) \in R^p \\ y &\in \mathcal{Y} = \{1, \cdots, k\} \end{aligned}$$

Deseamos construír $h : R^p \to \mathcal{Y}$ en base a los datos de entrenamiento $\{(x_i, y_i), i = 1, \cdots n\}$
Tal que una medida del riesgo sea mínima:

$$R(y, h(x)) \;=\; E[\, \mathcal{L}(y, h(x)) \,] \;=\; E[\, I(y \neq h(x)) \,]$$

Clasificadores

- Discriminantes de Fisher
- Regresión logística
- Máquinas de soporte vectorial
- Árboles de decisión
- k-Vecinos más cercanos
- $\cdots$ etc.

Mejora de Clasificadores Vía Boosting

Boosting

Boosting es un método general para mejorar la precisión de clasificadores.

- Kearns & Valiant, principios de los 90's, plantearon la pregunta de si era posible que un clasificador "débil" pudiera mejorarse ("boost") y convertirse en uno "fuerte".
- Freund & Schapire, a mediados de los 90's propusieron la metodología Boosting.

La idea básica es generar una sucesión de clasificadores débiles, construídos de forma tal que enfatizen los objetos mal clasificados y, al final, formar el "ensamble" de clasificadores, ponderados de acuerdo a su calidad. Hay razones teóricas que justifican que Boosting no adolece de sobreajuste.

Algoritmo AdaBoost

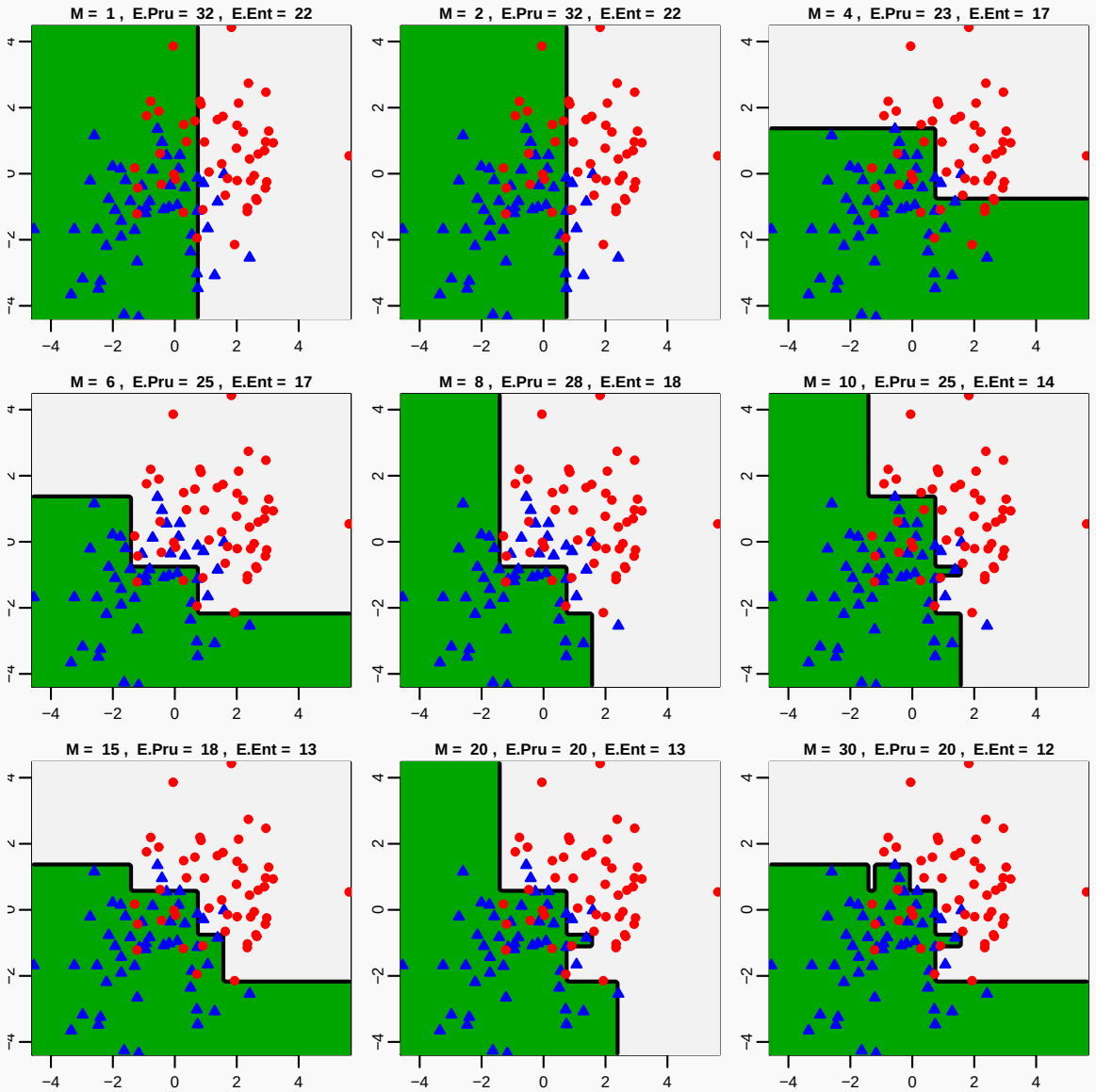
Freund & Schapire (1997): Construír distribución sobre mal clasificados.

1. Inicializar pesos $w_i = 1/n, \quad i = 1, 2, \cdots, n$
2. Para $m = 1, \cdots, M$:
   1. Ajustar clasificador $h_m(x)$ usando los pesos w_i
   2. Calcular
   $$e_m = \frac{\sum_{i=1}^{n} w_i I(y_i \neq h_m(x_i))}{\sum_{i=1}^{n} w_i}$$
   3. Calcular $\alpha_m = \log(\ (1 - e_m)/e_m\ )$
   4. Hacer $w_i \leftarrow w_i \exp[\alpha_m I(y_i \neq h_m(x_i))], \quad i = 1, 2, \cdots, n$
3. Construír clasificador final

$$H(x) = \text{signo}\left[\sum_{m=1}^{M} \alpha_m h_m(x)\right]$$

Rosett (2003) justifica que M es un parámetro de regularización.

Boosting en acción



Boosting

Breiman (1996) considera que los árboles de clasificación con AdaBoost, son los

“best off-the-shelf classifiers in the world”

con la propiedad extra de que “casi nunca” sobreajustan.

AdaBoost Alternativo

Es posible redefinir el algoritmo AdaBoost de modo que cambie pesos, no sólo a elementos mal clasificados, sino también a sus vecinos, con el fin de mejorar la probabilidad de buena clasificación (Ayala-Godoy, CLAPEM 2014). Los elementos para ello:

- Definición de similaridades:

$$D(x_q, x_i) = \frac{1}{(1 + \exp(-P_i^t))\ d(x_q, x_i)}$$

- Redefinición de pesos:

$$P_{t+1,x_q} = P_t - \frac{\lambda}{d(x_q, x_i)}, \quad g_t(x_i) \neq y_i$$

$$P_{t+1,x_q} = P_t + \frac{\lambda}{d(x_q, x_i)}, \quad g_t(x_i) = y_i$$

$$P_{t+1} = \text{máx}_{x_q}\ P_{t+1,x_q}$$

AdaBoost y Modelos Aditivos

Friedman, Hastie y Tibshirani (2000) consideraron la función de pérdida exponencial

$$L(F) = \mathrm{E}\left(e^{-y\,F(x)} \right)$$

y mostraron:

- Su minimizador poblacional es el clasificador óptimo

$$F(x) = \frac{1}{2}\log\frac{P(y=1|x)}{P(y=-1|x)}$$

· · · AdaBoost y Modelos Aditivos

FHT (2000) mostraron:

- Al minimizar $L(F)$ por pasos hacia adelante (construyendo $F(x) \leftarrow F(x) + c_m h_m(x)$) se tiene que AdaBoost no es más que un método para ajustar un modelo de regresión logística aditivo

$$\log \frac{P(y=1|x)}{P(y=-1|x)} = \sum_m f_m(x)$$

- Este resultado abrió la puerta a la consideración de otras funciones de pérdida, e.g. la basada en la logverosimilitud binomial

$$L_*(F) = \mathrm{E}\left(\log(1 + e^{2\,y\,F(x)})\right)$$

  La minimización de L_* da lugar al "'LogitBoost"'.

- Kramer (2006) es, hasta donde sabemos, el único esfuerzo de Boosting, usando estas ideas, en el contexto de clasificación funcional.

Clasificadores Funcionales

Clasificación Óptima

La regla óptima para clasificar una observación (curva) es asignarla a la clase con la probabilidad posterior más alta, i.e. clasificamos x en la clase 1 (en un problema con dos clases) si

$$\pi_1 f_1(x) > \pi_0 f_0(x)$$
$$\pi_1 P(x|y=1) > \pi_0 P(x|y=0)$$
$$\mathrm{E}(y|x) > \frac{1}{2}$$

- Hastie et al. (AoS 1995) estiman Σ usando regularización.
- Baíllo et al. (Scand JS 2011) hacen una revisión del área de clasificación funcional.
- James & Hastie (JRSS-B 2001), Shin (JMA 2008), Delaigle & Hall (JRSS-B 2012).
- Cuevas et al. (Computational Stat. 2007), López-Pintado & Romo (JASA 2009) usan el concepto de "profundidad" para definir métricas en espacios de funciones y lo usan para discriminación.

Discriminantes Lineales (C_1)

- Clasificador funcional de Alonso, Casado y Romo (2012)
- Transforman datos funcionales en multivariados
- Si $f_1, \cdots, f_m$ y $g_1, \cdots, g_n$ son observaciones de dos poblaciones, para cada f_j se construye $x_j = (x_{j1}, x_{j2}, x_{j3})$, donde

$$x_{jk} = d(f_j^{k-1}, \bar{f}^{k-1}) - d(f_j^{k-1}, \bar{g}^{k-1})$$

donde $f^(k-1)$ es la $(k-1)$-ésima derivada de f. De manera análoga se construyen $y_j = (y_{j1}, y_{j2}, y_{j3})$

- Discriminante lineal: $z(x) = a^T x$, con

$$a = \text{argmax} \left\{ \frac{a^T B a}{a^T W a} \right\}$$

Discriminantes Lineales (C_1)

- Discriminante lineal: $z(x) = a^T x$, con

$$a = \text{argmax} \left\{ \frac{a^T B a}{a^T W a} \right\}$$

y

$$B = m(\bar{x} - M)(\bar{x} - M)^T + n(\bar{y} - M)(\bar{y} - M)^T$$

$$W = \sum_{j=1}^{m} \frac{1}{P_t(x_j)} (x_j - \bar{x})(x_j - \bar{x})^T + \sum_{j=1}^{n} \frac{1}{P_t(y_j)} (y_j - \bar{y})(y_j - \bar{y})^T$$

los P_t's son los pesos definidos por AdaBoost al paso t.
Una observación, x, es clasificada en la población de las f's si $z(x) > \frac{1}{2}(\bar{x} - \bar{y})^T W^{-1} (\bar{x} - \bar{y})$.

Vecinos Más Cercanos (C_2)

- Clasificador funcional de Ferraty y Vieu (2006)
- Clasificador tipo "plug-in" del óptimo
- La probabilidad de pertenecer al grupo de etiquetas $\{y = 1\}$ es estimada por

$$\widehat{P_y(x)} = \mathrm{E}(I[Y = 1]\ x) = \frac{\sum P_t(x_i) I[y_i = 1] K(d(x, x_i)/h)}{\sum P_t(x_i) K(d(x, x_i)/h)}$$

donde K es el kernel gaussiano, d es la distancia en L^2 y h elegida vía validación cruzada.

Clasificador Basado en Profundidad (C_3)

- Clasificador funcional de López-Pintado y Romo (2006)
- Las medidas de profundidad permiten definir un orden para un conjunto de funciones.
- La distancia de banda generalizada (GBD) puede definirse como el promedio de la proporción del tiempo que una curva se encuentra entre pares de curvas, entre triadas de curvas, etc.
- La distancia de una curva, x, al grupo $x_1, \cdots, x_n$ se define como

$$DMT(x) = \frac{\sum \frac{1}{P_t(x_{(i)})} d(x, x_{(i)}) GBD(x_{(i)})}{\sum \frac{1}{P_t(x_{(i)})} GBD(x_{(i)})}$$

y se clasifica a x al grupo más cercano, de acuerdo a esta Distancia Media Truncada.

Experimentos

Experimentos

Se examinó el desempeño de los clasificadores C_1, C_2 y C_3, con AdaBoost, en sus versiones original y modificada. Así como con tres conjuntos de prueba: Puentes Brownianos simulados y dos conjuntos estándar de la literatura de FDA.

Datos Simulados

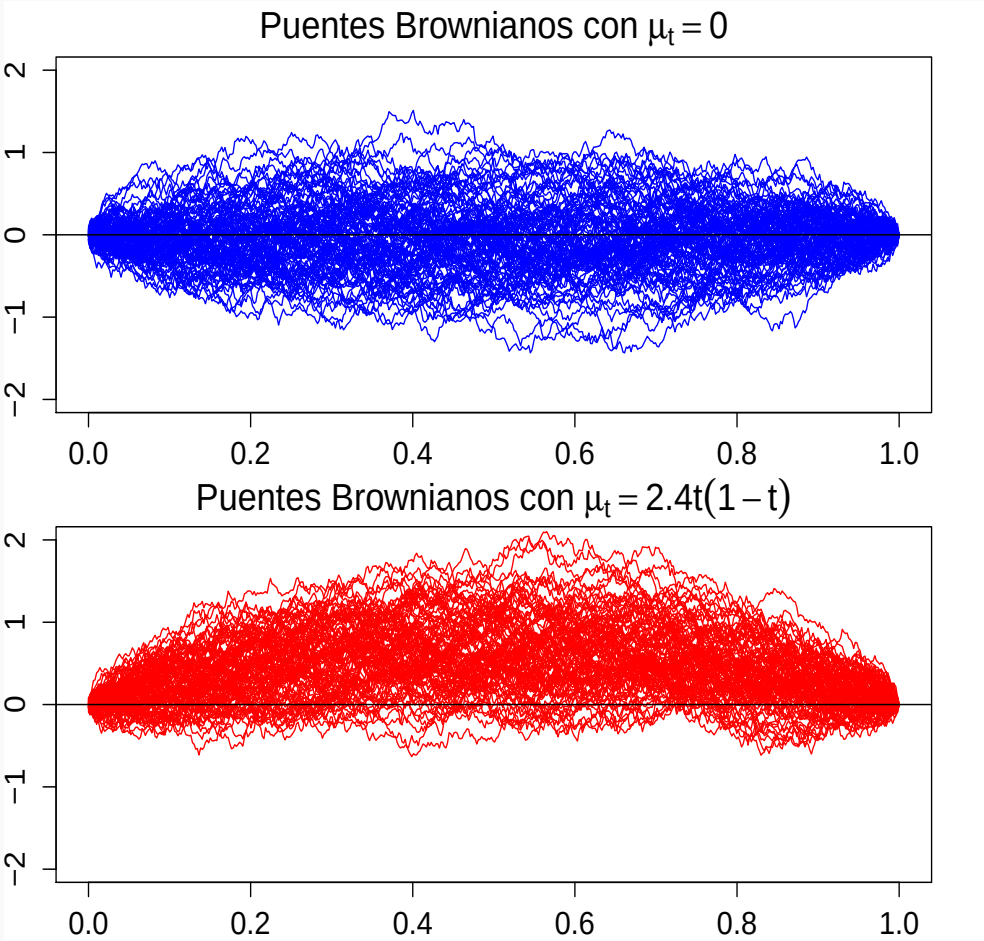


Figura 4.3: 100 puentes Brownianos con media cero en la parte superior, (color azul), y 100 puentes Brownianos con $\mu^*(t) = 2.4t(1-t)$ en la parte inferior, (color rojo).

Desempeño de Clasificadores

Datos Simulados

	Algoritmo		
Clasificador	Estándar	AdaBoost	Alternativo
Discriminante	52	76	76
kNN	52	80	82
Profundidad	56	82	80

Datos de Espectrometría

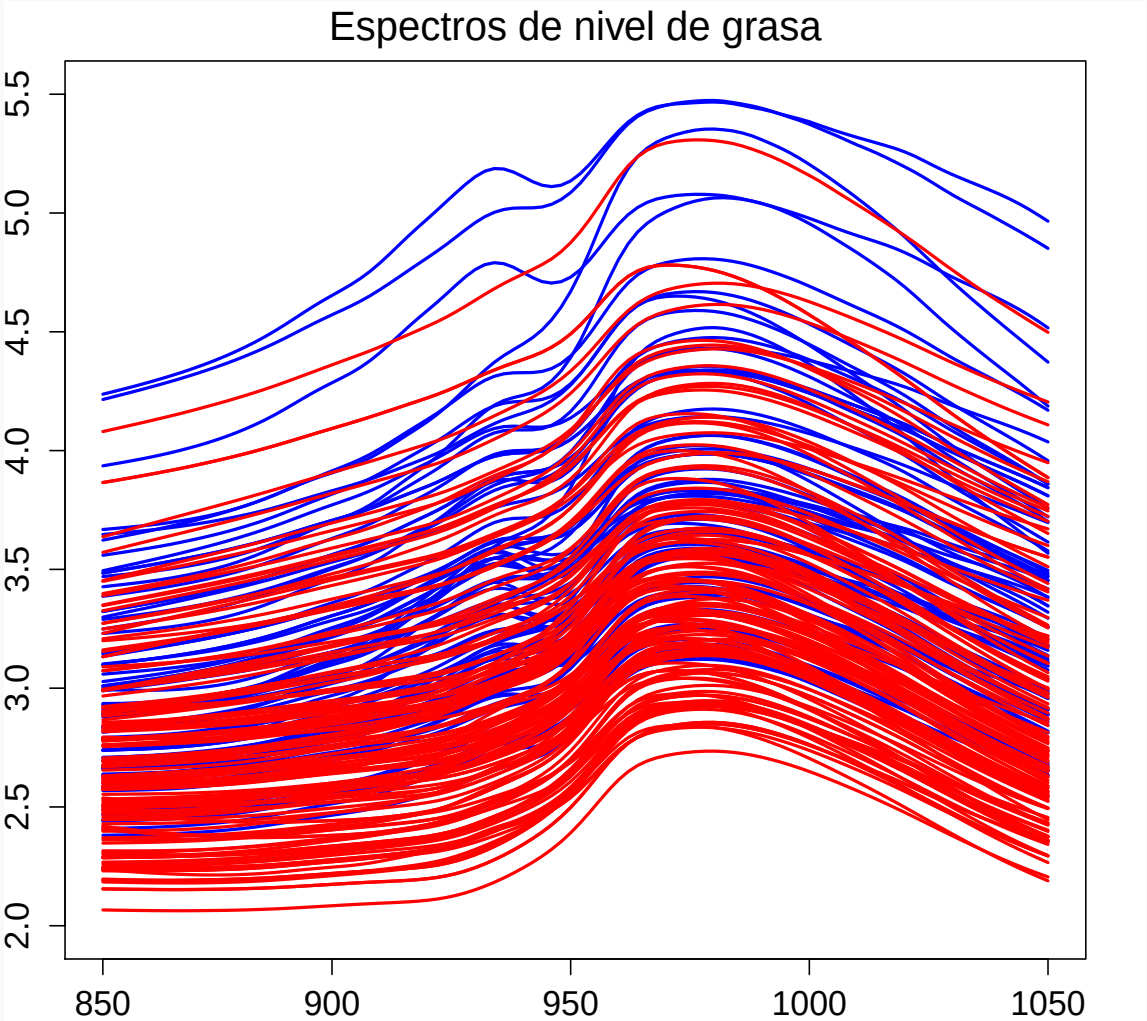


Figura 4.4: Datos de espectrometría provenientes de Tecator Infratec Food.

Desempeño de Clasificadores

Datos de Espectrometría

	Algoritmo		
Clasificador	Estándar	AdaBoost	Alternativo
Discriminante	97.93	100	98.94
kNN	95.88	97.89	97.89
Profundidad	96.91	96.91	97.89

Datos de Crecimiento

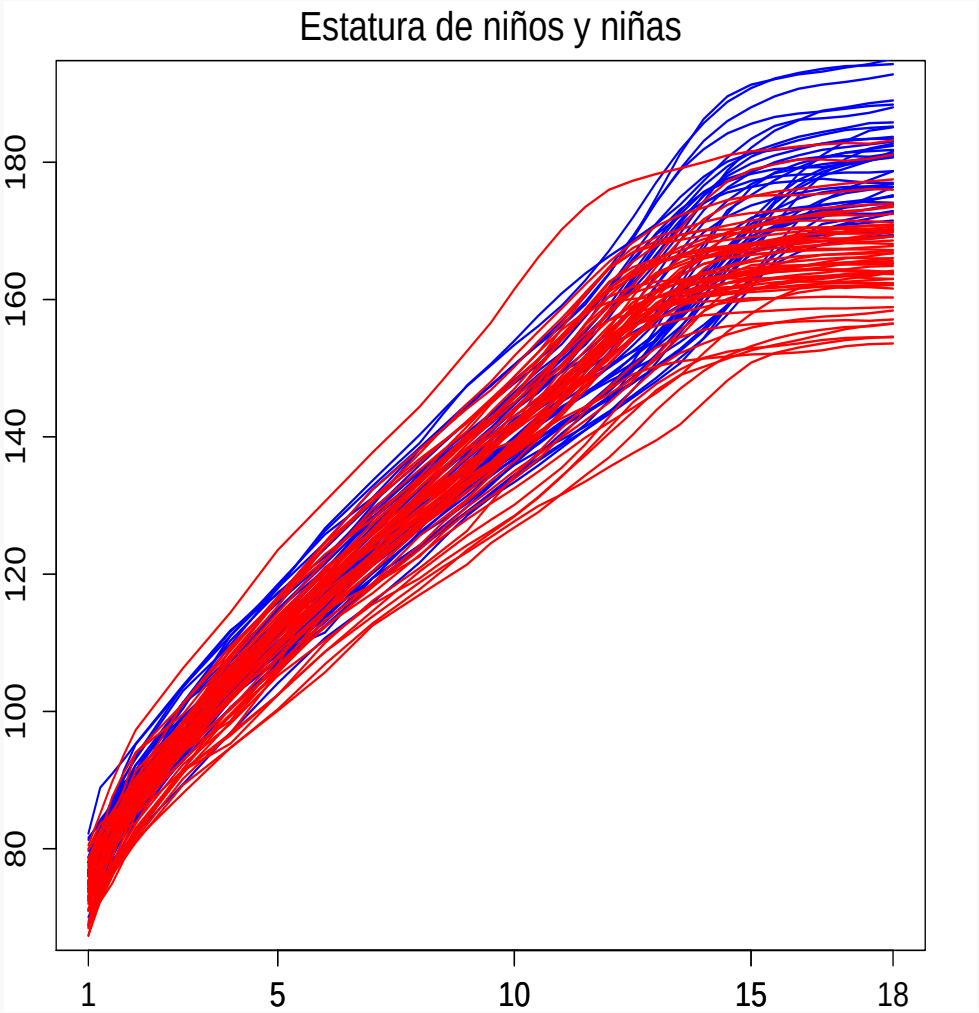


Figura 4.6: Datos de un estudio de crecimiento en Berkeley. Se tomaron 31 mediciones de 1 a 18 años.

Desempeño de Clasificadores

Datos de Crecimiento

	Algoritmo		
Clasificador	Estándar	AdaBoost	Alternativo
Discriminante	95.65	95.65	95.65
kNN	91.30	95.65	100
Profundidad	95.65	95.65	100

Conclusiones

- Boosting mejora la eficiencia de los clasificadores
- No hay una ganancia clara al usar el AdaBoost alternativo
- Las ganancias de Boosting son más relevantes en situaciones con poblaciones difíciles de separar.
- Trabajo pendiente: Evaluar las propiedades en general del procedimiento alternativo.

Algunas Referencias

- Alonso, A.M., Casado, D., & Romo, J. (2012). Supervised classification for functional data: A weighted distance approach. *Computational Statistics and Data Analysis*. **56**, 2334-2346.
- Ferraty, F. & Vieu P. (2006). *Nonparametric functional data analysis. Theory and practice*. Springer-Verlag, New York.
- Freund, Y. & Schapire, R.E. (1997). A decision-theoretic generalization of online learning and an application to boosting. *Journal of Computer and System Sciences*. **55**, 119-139.
- Friedman, J., Hastie, T. & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. *Annals of Statistics*. **28**, 2, 337-407.
- Kramer, N. (2006). Boosting for functional data. arXiv preprint math/0605751
- Lopez-Pintado, S. & Romo, J. (2006). Depth-based classification for functional data. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*. American Mathematical Society. **72**, 103-120.

GRACIAS!!