Métodos estadísticos prerdictivos (o supervisados): Machine Learning

Profesor: Rodrigo Assar

El Problema

- **Ajustar (/predecir, /modelar) el valor de una variable (explicada) respecto (/en función) de los valores de variables explicativas.**

Buscando modelos generales

- Datos de entrenamiento.

- Datos de validación.

- **Para medir la calidad de un modelo (precisión, exactitud, …) este se debe testear (o validar) sobre datos no usados para el entrenamiento.**

Diferentes enfoques según tipo de variable explicada

- **Clasificación:** la variable explicada toma valores discretos (generalmente finitos) llamados categorías o clases.
  - Ejemplo simple: Análisis discriminante lineal

- **Regresión:** variable explicada modelada pudiendo tomar "una continuidad de valores".
  - Ejemplo simple: Regresión lineal (múltiple)

Muchos métodos de clasificación

- Análisis discriminante lineal, cuadrático, knn, CART, SVM, …

Muchos métodos de regresión

- Lineal (múltiple), lineal generalizada, logística, lineal por ventanas, …

Machine Learning

- Método estadístico que aprende de los datos de entrenamiento.
- Unos aprenden mejor que otros.
- Se puede crear flujo de modo de que el método se vaya alimentando iterativamente de nuevos datos para mejorar su calidad “general”.

Boosting

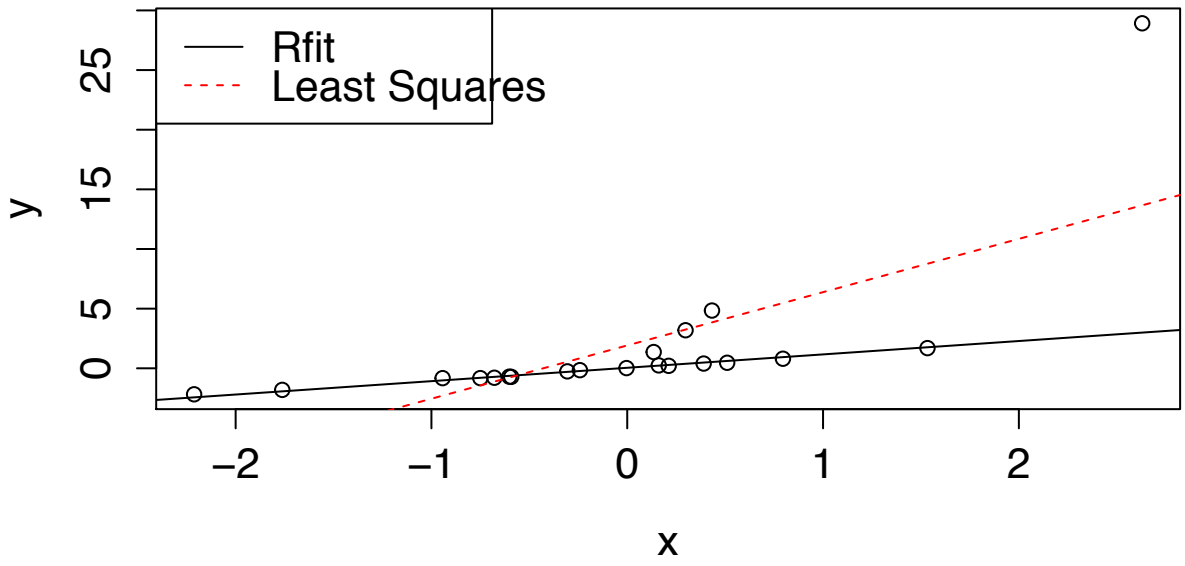
– learner=predictor trained from data, a learner is **weak** if it is slightly better than “throwing a coin”.

– **Q:** Can a set of **weak learners** create a single **strong learner**? Kearns, 1988.

- **Answer: Yes, boosting!** Schapire, 1990. Schapire and Freund Gödel Prize (2003)
- **AdaBoost** algorithm and others: LogitBoost.
- Idea: method focuses iteratively in correcting miss-predicted data.

Buscando robustez

- Detectando outliers:
  - Mediante “exploración” previa,
  - Modificando modelos que automáticamente pesen menos a outliers. Ejemplo: paquete rfit de R



Buscando robustez (cont.)

- Bootstrap: estimación por remuestreo.
- Efron, B.; Tibshirani, R. (1993). An Introduction to the Bootstrap. Boca Raton, FL: Chapman & Hall/CRC

Buscando robustez (cont.)

Modelos semi-supervisados o mixtos:

- Clusterizando +
- en cada grupo construyendo un modelo.

Ejemplo:

```
library(party)
#Ejemplo de Fisher iris, prediciendo con variables de petalo y #particionando con las de sepalo.

logistictree.iris <- mob(versicolor~iris$Petal.Width+iris$Petal.Length|
iris$Sepal.Width+iris$Sepal.Length,model = glinearModel,family = binomial())

logistictree.iris

coef(logistictree.iris)

summary(logistictree.iris)

#Graficando

plot(logistictree.iris)

plot(logistictree.iris,tp_args = list(cdplot = TRUE))
```

Resultado de ejemplo: árbol logístico