Bagging

Un problema muy común es intentar predecir el comportamiento de una variable dependiente Y a partir de un conjunto de p variables independientes $X \in R^p$

A partir de una muestra de observaciones $(\vec{x}_i, y_i),\ i = 1, \ldots, n$ se busca una regla de clasificación, RC, mediante regresión logística, árboles de clasificación, etc.

Bagging

Si se conociera la distribución conjunta de $(\vec{X}, Y)$, es decir,

$$P[(\vec{X}, Y) = (\vec{x}, y)]$$

el procedimiento más eficiente sería utilizar como regla de predicción la probabilidad a posteriori

$$\mathrm{RC}(\vec{x}) = \arg\max_y P(Y = y \mid \vec{x}) = \arg\max_y P[(\vec{X}, Y) = (\vec{x}, y)]$$

El problema es que la distribución conjunta suele ser desconocida y es necesario estimarla.

Bagging. Bootstrap

El método bootstrap se basa en sustituir la función de distribución teórica $F_{(X,Y)}$ por la función de distribución empírica $\hat{F}(X,Y)$. Para ello se considera la distribución estimada de probabilidad

$$\hat{P}(X,Y) = \begin{cases} \frac{1}{n} & \text{si } (\vec{x},y) = (\vec{x}_i,y_i) \text{ para algún } i \\ 0 & \text{en otro caso} \end{cases}$$

(Teorema de Glivenko-Cantelli)

Una muestra bootstrap $\left\{(\vec{x}_i^*, y_i^*) \mid i = 1,\ldots,n\right\}$, se obtiene eligiendo al azar y con reposición n observaciones de la muestra original $(\vec{x}_1,y_1),\ldots,(\vec{x}_n,y_n)$.

Así se tiene una idea de la variabilidad que produciría repetir varias veces el proceso de muestreo inicial.

Bagging

A partir de una muestra $(\vec{x}_i, y_i),\ i = 1, \ldots, n$ el método bagging utiliza el siguiente algoritmo

- Se eligen b muestras bootstrap $\{(\vec{x}_i^{*b}, y_i^{*b}) \mid i = 1, \ldots, n; b = 1, \ldots, B\}$
- Se calcula una regla de clasificación RC^b basada en la correspondiente muestra bootstrap b-ésima.

La clasificación final de cada individuo $\vec{x}$ se hace como promedio de la B predicciones

$$\mathrm{RC}_{\mathrm{bag}}(\vec{x}) = \arg\max_k \sum_{b=1}^{B} I(\mathrm{RC}^b(\vec{x}) = k)$$

es decir, se asigna a la categoría más elegida.

Bagging

La estimación de la probabilidad de pertenencia a un categoría k se basa en las probabilidades de clasificación obtenidas en cada una de las B réplicas realizadas

$$\hat{P}_{\text{bag}}(k \mid \vec{x}) = \frac{1}{B} \sum_{b=1}^{B} \hat{P}^b(k \mid \vec{x})$$

Esta estimación sugiere asignar las observaciones a clases empleando un nuevo criterio:

$$\text{RC}_{\text{bag}}(\vec{x}) = \text{arg}\, \max_k \{\hat{P}_{\text{bag}}(k \mid \vec{x})\}$$

Bagging

Estimación del error out-of-bag (OOB)

- Usar validación cruzada para estimar el error en la predicción.
- En cada bootstrap se utiliza aproximadamente el 63 % de las muestras.
- Usar el 37 % restante para estimar el error.

$$e_i = y_i - \hat{y}_i^{\text{oob}}$$

- Calcular el Error Cuadrático Medio.

$$\text{ECM} = \frac{1}{n} \sum e_i^2$$

Bagging

VENTAJAS

- La tasa de error tiende a reducirse al promediar muchos criterios de clasificación.
- Permite estimar la importancia de las variables.
  1. Se estima la reducción del Error Cuadrático Medio cada vez que interviene una variable.
  2. Se calcula el promedio de las reducciones en todas las muestras bootstrap.

Bagging

DESVENTAJAS

- No se debe emplear cuando los clasificadores individuales funcionan mal.
- La interpretación del criterio de clasificación suele ser mucho más difícil.
- Requiere muchos más cálculos.

CÓDIGO DE R

```
## Ejemplo

library(ipred)                  # para bagging

library(rpart)                  # para rpart.control

library(mlbench)                # para BreastCancer

data(BreastCancer)              # cargar ejemplo

BC <- BreastCancer              # ahorrar pulsaciones

n <- nrow(BC)                   # tamaño muestral

entrena <- sample(n, 2*n/3) # índices entrenam.
```

CÓDIGO DE R

```
BC.bagg <- bagging(Class ~., data=BC[entrena,-1],
                   nbagg=10, coob=TRUE,
                   control=rpart.control(maxdepth=3))

BC.bagg.pred <- predict(BC.bagg, BC[-entrena,-1],
                  type="class") # por omisión

BC.bagg.pred <- predict(BC.bagg, BC[-entrena,-1],
                  type="prob")

names(BC.bagg)     # componentes de la salida
BC.bagg$mtrees     # árboles
BC.bagg$OOB        # si se usan los out-of-bag
BC.bagg$err        # estimación del error basada en OOB
```

CÓDIGO DE R

```
## bagging depende de las rutinas de rpart

## En cada bootstrap, rpart se controla con los
## parámetros habituales.

## Por defecto rpart.control tiene minsplit=2, cp=0
## y xval=0.

## Se puede utilizar la función prune para podar
## los árboles.
```