Estimar, descomponer y comparar el error de mala clasificación

Evaluando y analizando el comportamiento de algoritmos de inducción de clasificadores

Aritz Pérez, Pedro Larrañaga e Iñaki Inza

Intelligent Systems Group
Ciencias de la Computación e Inteligencia Artificial
Universidad del Pais Vasco-Euskal Herriko Univertsitatea

14-IX-2005 / CEDI-TAMIDA'05

UPV - EHU

Índice

Índice

Notación

- $\boldsymbol{x}$ instanciación de las variables (predictoras) $\boldsymbol{X} \in \Re^n$
- $c \in \{1, ..., r\}$ es la clase real asociada a $\boldsymbol{x}$
- Densidad de los datos (real y desconocida) $f(\boldsymbol{x}, c)$
- $S_N = \{(\boldsymbol{x}_1, c_1), ..., (\boldsymbol{x}_N, c_N)\}$ es un conjunto de N casos
- El clasificador entrenado en S,

$$g(\boldsymbol{x}, S) : \Re^n \times \Re^{N \times n+1} \rightarrow \{1, ..., r\}$$

- Delta de Kronecker, $\delta(u, u') = 1 \leftrightarrow u = u'$

El error de clasificación $\epsilon(g|S_N) \in [0,1]$,

$$\epsilon(g|S_N) = P(g(\boldsymbol{X}, S_N) \neq C) = E_f(1 - \delta(c, g(\boldsymbol{x}, S_N)))$$

El error cuadrático medio $\epsilon_{ms}(g_r|S_N)$

$$\epsilon_{ms}(g_r|S_N) = E_f[(g_r(\boldsymbol{x}|S_N) - f(\boldsymbol{x}))^2]$$

- La clase $Y \in \Re$ es continua (regresión)
- Función de regresión (clase continua)

$$g_r(\boldsymbol{x}|S_N) : \Re^n \times \Re^{N \times n+1} \longrightarrow \Re$$

Índice

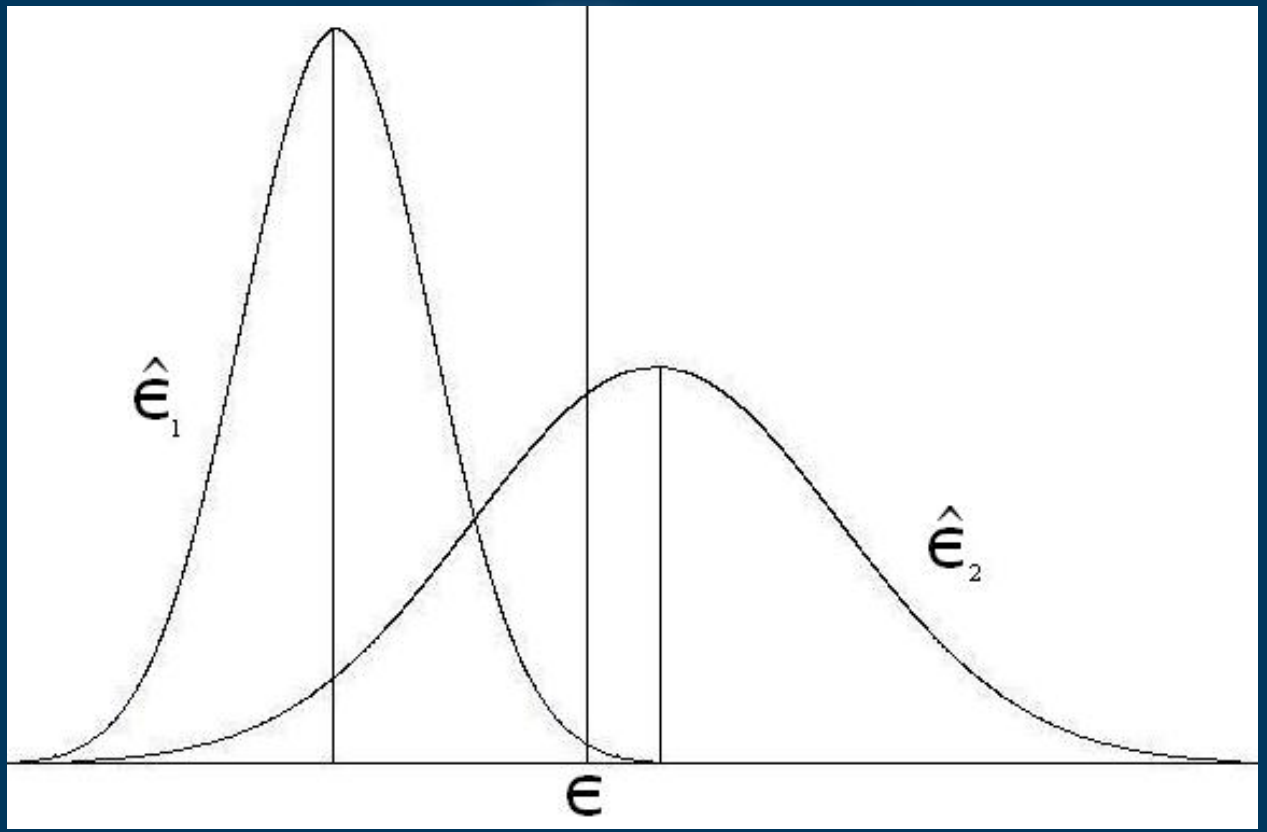
$$\hat{\epsilon}(g|S_N) = E_{f^*_{S_N}}(1 - \delta(c, g(\boldsymbol{x}, S_N)))$$

$\hat{\epsilon}(g|S_N)$ es el estimador que se obtiene al reemplazar f por la densidad empírica $f^*_{S_N}(\boldsymbol{x}, c) = 1/N \leftrightarrow (\boldsymbol{x}, c) \in S_N$, basada en la muestra S_N

$$f(\boldsymbol{X}, C) \longrightarrow S_N = \{(\boldsymbol{x}_1, c_1), ..., (\boldsymbol{x}_N, c_N)\} \longrightarrow f^*_{S_N}(\boldsymbol{X}, C)$$

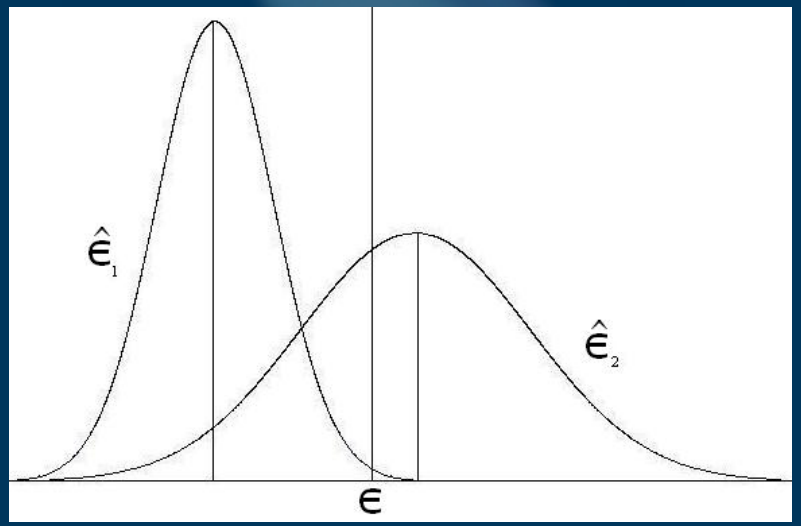
Sesgo y varianza de un estimador:

- $sesgo = \epsilon - E(\hat{\epsilon})$
- $varianza = E[(\hat{\epsilon} - E(\hat{\epsilon}))^2]$



Índice

Obtener una estimación del error de un clasificador entrenado en S_N, $\hat{\epsilon}(g|S_N)$, lo menos sesgada y variable posible



Índice

- Resustitución (resampling) (Smith 1947): Optimista por sobreajuste ($S_N = S^e = S^t$)

$$\hat{\epsilon}_r = \frac{1}{N}\sum_{i=1}^{N}(1 - \delta(c_i, g(\boldsymbol{x}_i, S_N)))$$

- Holdout: Pesimista (conj. entren $S^e_{N'}$ con $N' < N$).

$$S^e \cap S^t = \emptyset \text{ y } S_N = S^e \cup S^t$$

$$\hat{\epsilon}_h = \frac{1}{N - N'}\sum_{i=1}^{N-N'}(1 - \delta(c_{t:i}), g(\boldsymbol{x}_{t:i}, S^e_{N'}))$$

Índice

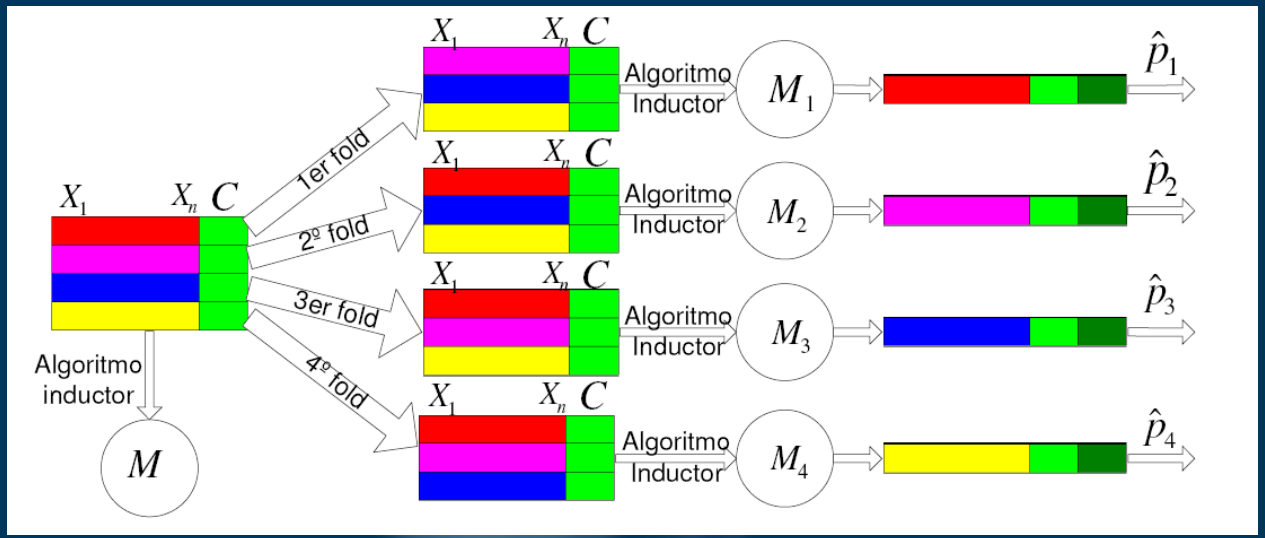
Repeated holdout

Repetir el proceso del holdout (l veces):

$$\hat{\epsilon}_{rh} = \frac{1}{l} \sum_{j=1}^{l} \hat{\epsilon}_h^{(j)}$$

- Estimación del error poco sesgada pero muy variable (muestra $\hat{\epsilon}_h^{(j)}, j = 1, ..., l$ dependiente)

K-fold cross-validation (Stone 1974)



$$\hat{\epsilon}_{kfcv} = \frac{1}{k}\sum_{i=1}^{k}\hat{\epsilon}_i$$

- Testar en conjuntos disjuntos (reducir dependencia)
- Estimación del error poco sesgada (menos que repeated holdout) pero muy variable.

UPV - EHU

Versiones K-fold cross-validation

- 10-fold cross-validation ($k = 10$): opción más difundida
- Leave-one-out ($k = N$): Algo menos variable pero más costoso
- Stratified cross-validation: menos variable
- Repeated cross validation: menos variable y más costoso

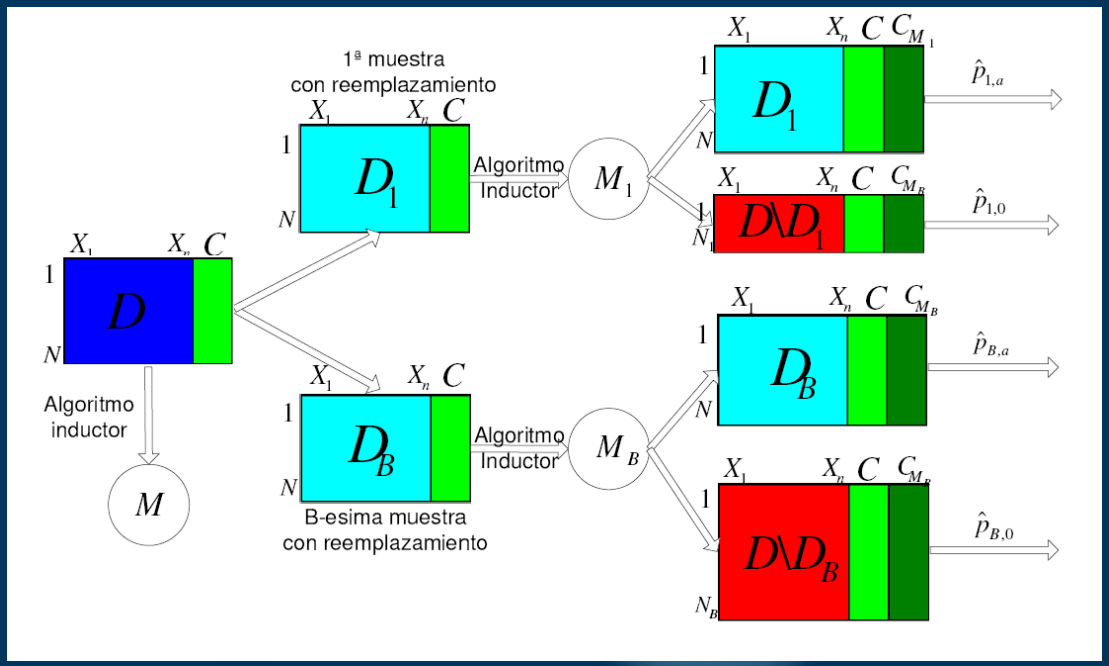
Jackknife (Quenouille 1956)

$$\begin{aligned}\hat{\epsilon}_J &= \frac{1}{n}\sum_{i=1}^{n}\delta(\hat{f}(x_i, S_n), y_i) \\ &+ (n-1)(\frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n}\delta(\hat{f}(x_i, S_n), y_i) \\ &- \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{k=1,k\neq j}^{n}\delta(\hat{f}(x_i, S_n), y_i)) \qquad (1)\end{aligned}$$

- Variante del leave-one-out que entrena con $S^i_{N-1} = S_N \setminus (\boldsymbol{x}, c)$ y testa en S_N
- Menos sesgo que la resustitución

0.632 Bootstrap estimator (Efron 1983)

Muestreo de la distribución empírica $F^*_{S_N}(\boldsymbol{X}, C)$



Estimador con poco sesgo (problemas con clasif. complejos) y menos varianza que k-fold c.v.

$$\hat{\epsilon}_b = 0.368\hat{\epsilon}_r + 0.632\hat{\epsilon}_0 \qquad \hat{\epsilon}_r = \frac{1}{B}\sum_{i=1}^{B}\hat{\epsilon}_r^{(i)} \qquad \hat{\epsilon}_0 = \frac{1}{B}\sum_{i=1}^{B}\hat{\epsilon}_0^{(i)}$$

Versiones de bootstrap

- Zero bootstrap estimator, $\hat{\epsilon}_0$: sesgado
- Balanced bootstrap resampling (Chernick 1999): menos variable
- Parametric bootstrap: muestrear un modelo probabilista
- Smoothed bootstrap: muestrear una densidad basada en kernels

Bolstered (Braga-Neto 2004)

Emplear una densidad basada en kernels $f^{\diamondsuit}$ en lugar de la densidad empírica f^* para computar la esperanza del error.

$$f^{\diamondsuit}(\boldsymbol{x}, c) = \frac{1}{N} \sum_{i=1}^{N} {f_i}^{\diamondsuit}(\boldsymbol{x} - \boldsymbol{x}_i)\delta(c = c_i)$$

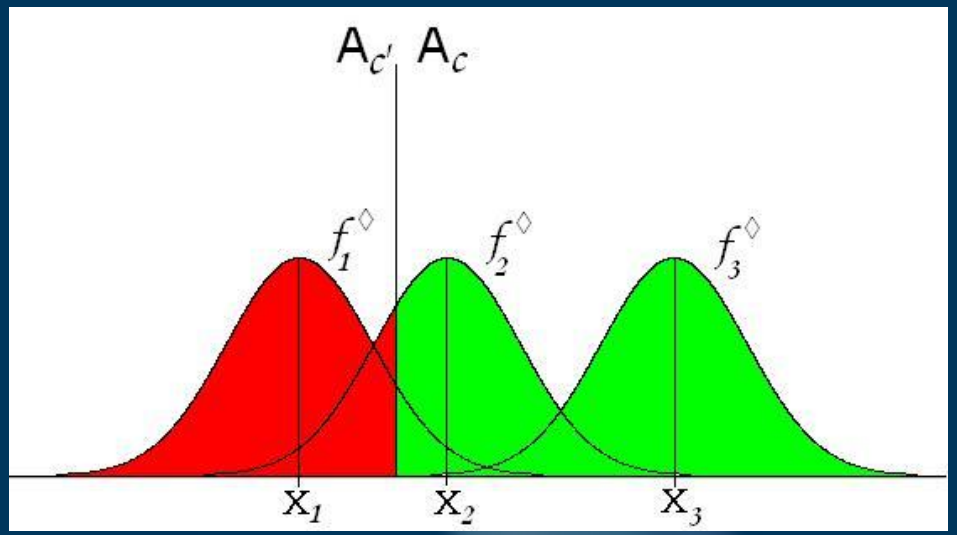
- Esparcir la masa $(1/N)$ de cada caso $\boldsymbol{x}$ en su entorno empleando
- Permitir que un caso aporte un error $\epsilon_i^{\diamondsuit} \in [0, 1/N]$ en lugar de $\epsilon_i^* \in \{0, 1/N\}$

Bolstered resubstitution (Braga-Neto'04)

$$\hat{\epsilon}_{br} = \frac{1}{N}\sum_{i=1}^{N}\left(\int_{A_1} f_i^{\diamondsuit}(\boldsymbol{x}-\boldsymbol{x}_i)d\boldsymbol{x}\,\delta(c_i,0) + \int_{A_0} f_i^{\diamondsuit}(\boldsymbol{x}-\boldsymbol{x}_i)d\boldsymbol{x}\,\delta(c_i,1)\right)$$

- Comparable al bootstrap pero más eficiente
- Computar las integrales mediante el muestreo de Monte-Carlo
- La mejor opción para clasificadores lineales (formulación cerrada)
- Peor comportamiento (optimista) con clasificadores complejos (sobreajuste)

Bolstered resubstitution (Braga-Neto 2004)



$$A_c = \{x / g(x, S_N) = c\}$$

$$c_1, c_2, c_3 = c \quad \{x_1\} \in A_{c'} \quad \{x_2, x_3\} \in A_c$$

Alternativas bolstered (Braga-Neto'04)

- Semi-bolstered resubstitution: más optimista pero más variable (clasificadores complejos).
- Bolstered leave-one-out: poco sesgo y menos variable que leave-one-out y repeated cross validation

Índice

Analizar el comportamiento de un clasificador empleando la descomposición en sesgo y varianza de su error.

- Introducida por German y col. (1992) para la esperanza del error cuadrático (mean squared error) al mundo del aprendizaje automático (regresión).
- La descomposición propuesta por Kohavi y Wolpert (1996) para funciones funciones de perdida 0-1 (zero-one-loss functions) es la más empleada.
- Analizar el dominio de un problema.
- Equilibrio entre sesgo y varianza (bias variance trade-off)

Descomposición del error cuadrático medio (regresión) $\epsilon_{ms}(\hat{f}_{g|S_N})$

$$
\begin{aligned}
\epsilon_{ms}(g_r|S_N) &= E_f[(g_r(\boldsymbol{x}|S_N) - f(\boldsymbol{x}))^2] \\
&= E_f[g_r(\boldsymbol{x}|S_N) - E(g_r(\boldsymbol{x}|S_N)) - f(\boldsymbol{x},c) + E(g_r(\boldsymbol{x}))]^2 \\
&= E_f[(g_r(\boldsymbol{x}|S_N) - E(g_r(\boldsymbol{x}))^2] \\
&\quad + E_f[(f(\boldsymbol{x}) - E(g_r(\boldsymbol{x}|S_N)))^2] \\
&\quad - E_f[2(g_r(\boldsymbol{x}|S_N) - E(g_r(\boldsymbol{x}|S_N)))(f(\boldsymbol{x},c) - E(g_r(\boldsymbol{x})))]
\end{aligned}
$$

- $variance = E_f[g_r(\boldsymbol{x}, S_N) - E(g_r(\boldsymbol{x}, S_N))]^2$
- $bias^2 = E_f[f(\boldsymbol{x}, c) - E(g_r(\boldsymbol{x}, S_N))]^2$
- El tercer termino es cero

UPV - EHU

Índice

Error de clasificación para funciones de perdida 0-1

$$\epsilon_{0-1} = \sum_{i=1}^{N} p(\boldsymbol{x}_i) \sum_{c_h=1}^{r} \sum_{c_t=1}^{r} (1 - \delta(c_h, c_t)) p(c_t|\boldsymbol{x}_i) \hat{p}(c_h|\boldsymbol{x}_i)$$

$$= \sum_{i=1}^{N} p(\boldsymbol{x}_i)(1 - \sum_{c=1}^{r} p(c|\boldsymbol{x}_i) \hat{p}(c|\boldsymbol{x}_i))$$

Descomposición de ϵ_{0-1}

$$\epsilon_{0-1} = \sum_{i=1}^{N} p(\boldsymbol{x}_i)(\sigma^2_{\boldsymbol{x}_i} + bias^2_{\boldsymbol{x}_i} + variance^2_{\boldsymbol{x}_i})$$

- El objetivo no consiste en estimar el error de forma insesgada e invariante
- Comportamiento aditivo de los términos
- Existe un equilibrio entre sesgo y varianza (bias variance trade-off).
- La incorporación de información *a priori* es una buena opción para tratar de reducir ambos términos.

Interpretación de la descomposición de ϵ_{0-1}

Ruido implícito:

$$\sigma^2 \equiv \frac{1}{2}\sum_{i=1}^{N}(1 - \sum_{c=1}^{r} f(\boldsymbol{x}_i, c)^2)$$

- Expresa el ruido de la distribución real de los datos
- Relacionado con el error de Bayes ϵ_{0-1}^{B} (mínimo error)
- En la práctica es cero a no ser que

$$(\boldsymbol{x}, c) \in S \wedge (\boldsymbol{x}, c') \in S \wedge c \neq c'$$

Interpretación de la descomposición de ϵ_{0-1}

Sesgo:

$$bias^2 \equiv \frac{1}{2}\sum_{i=1}^{N}\sum_{c=1}^{r}[f(\boldsymbol{x}_i, c) - \hat{f}(\boldsymbol{x}_i, c)]^2$$

- El sesgo mide el error debido al desajuste entre la densidad estimada $\hat{f}(\boldsymbol{x}, c)$ y la real $f(\boldsymbol{x}, c)$ (distancia)
- El sesgo tiende a ser mayor en clasificadores simples (con pocos parámetros)

Interpretación de la descomposición de ϵ_{0-1}

Varianza:

$$variance \equiv \frac{1}{2}\sum_{i=1}^{N}(1 - \sum_{c=1}^{r}\hat{f}(\boldsymbol{x}_i, c)^2)$$

- La varianza mide el error fruto de la variabilidad de la densidad estimada $\hat{f}(\boldsymbol{x}, c)$ a los cambios en el conjunto de entrenamiento.
- Puede considerarse una medida de sensibilidad a los cambios en el conjunto de entrenamiento.
- La varianza tiende a ser mayor en clasificadores complejos (con muchos parámetros)

Descomposiciones alternativas

- Mean squared error: German y col 1992.
- zero-one-loss: Kong and Dietterich 1995, Friedman 1997, Domingos 2000 y James 2003

Índice

Dados dos algoritmos de inducción de clasificadores A y B, poder establecer de forma fiable, si se comportan de manera similar o si uno es superior al otro

Herramienta matemática: Test estadístico

- Hipótesis nula H_0: los algoritmos A y B obtienen el mismo error

$$H_0 : \epsilon(g_A|S_N) = \epsilon(g_B|S_N) \tag{2}$$

Tests estadísticos

- Mann-Whitney (suma de rangos): no paramétrico, no apareada.
- Wilcoxon (diferencias): no paramétrico, apareada.
- T-test de Student: paramétrico (supone normalidad en las diferencias), apareado/no apareado, distribución t con $l-1$/l_A+l_B-2.

$$t = \frac{d^{(\cdot)}}{\sqrt{\frac{\sigma^2_{d^{(i)}}}{l}}} \qquad t = \frac{\epsilon_A^{(\cdot)} - \epsilon_B^{(\cdot)}}{\sqrt{\frac{\sigma^2_{\hat{\epsilon}_A^{(i)}}}{l_A} + \frac{\sigma^2_{\hat{\epsilon}_B^{(i)}}}{l_B}}}$$

Suponen independencia entre las muestras de un clasificador $\hat{\epsilon}_A^{(i)}$ y $\hat{\epsilon}_A^{(j)}$ $\forall i, j / i \neq j$. Los métodos de comparación que presentamos las incumplen

Criterios de evaluación del método

- Error Tipo I: probabilidad de rechazar la hipótesis nula cuando esta es cierta (falsa diferencia)
- Error Tipo II: probabilidad de aceptar la hipótesis nula cuando esta es falsa (falsa igualdad)
- Replicabilidad: probabilidad de que dos ejecuciones de un mismo método de comparación produzca los mismos resultados (estabilidad)

Índice

k-fold cross-validation + t-test pareado

$$t = \frac{d^{(\cdot)}}{\sqrt{\frac{1}{k}\sigma^2_{d^{(i)}}}}$$

- t-test pareado con $k-1$ grados de libertad
- Posibilidad de emplear otros tests
- Infraestima la varianza (dependencia train-train)
- Error Tipo I alto (llega a doblar la significatividad), Tipo II bajo y baja replicabilidad
- Comportamiento parecido a repeated holdout + t-test pareado.
- Casos particulares: $k = 10$ y $k = N$

5x2 cross validation (Dietterich 1998)

$$t = \frac{d^{(1,1)}}{\sqrt{\frac{1}{5}\sum_{i=1}^{5}\sigma^2_{d^{(i,\cdot)}}}}$$

- 5 ($l = 5$) ejecuciones de 2-fold c.v
- Sigue una distribución t con 5 grados de libertad
- Aceptable error Tipo I (mejor que 10 fold cv) y bajo error Tipo II
- Falla cuando la muestra de los errores estimados es heterogénea

Combined 5x2 cv F-test (Alpaydin 1999)

Emplear toda la muestra y emplear la media en el denominador

$$t = \frac{\sum_{i=1}^{5}\sum_{j=1}^{2} d^{(i,j)}}{2\sqrt{\sum_{i=1}^{5}\sigma^2_{d^{(i,\cdot)}}}}$$

- Sigue una F de Snedecor con 10 y 5 grados de libertad
- Menor error Tipo I y Tipo II que 10-fold cross-validation y 5x2 cross validation.

Corrected resampled t-test (Nadeau y Bengio 2003)

$$t = \frac{d^{(\cdot)}}{\sqrt{(\frac{1}{l} + \frac{N_t}{N_e})\sigma^2_{d^{(i)}}}}$$

- Muestreo aleatorio sin reemplazamiento (repeated holdout)
- Corrección sobre el estimador de la varianza del t-test pareado (modelando correlación de $\hat{\epsilon}^{(i)}$) para reducir el error Tipo I
- Distribución t con $l - 1$ grados de libertad
- Error Tipo I aceptable y error Tipo II bajo

Corrected repeated k-fold cross validation (Bouckaert y Frank 2004)

$$t = \frac{d^{(\cdot,\cdot)}}{\sqrt{(\frac{1}{k \cdot r} + \frac{1}{k-1})\sigma^2_{d^{(i)}}}}$$

- Obtiene $k \times l$ diferentes $d^{(i,j)}_{\hat{\epsilon}}$ (i-ésimo fold de la j-ésima ejecución).
- Estadístico equivalente al corrected resampled t-test (misma corrección) con $k \cdot l - 1$ grados de libertad.
- Errores Tipo I y Tipo II apropiados y mayor replicabilidad que corrected resampled t-test

Shorted runs sampling (Bouckaert 2004)

- Emplea los errores estimados mediante l times repeated k-cross validation (alto coste computacional)
- Dada la ejecución j-ésima $j \in \{1, ..., l\}$, ordena las diferencias $d^{(i,j)}$ obtenidas en cada fold $i \in \{1, ..., k\}$
- Una vez ordenadas las diferencias las promedia en las ejecuciones para obtener $d^{(i,\cdot)} = \frac{1}{l} \sum_{j=1}^{l} d^{(i,j)}$
- Errores Tipo I y Tipo II apropiados y alta replicabilidad (con t-test sin corrección y Wilcoxon)

Se han mostrado:

- Alguno métodos para estimar el error de un clasificador
- La descomposición en sesgo y varianza del error de clasificación para funciones de perdida 0-1 (Kohavi y Wolpert 1996)
- Varias herramientas que permiten comparar dos clasificadores en términos del error que cometen

aritz@si.ehu.es

UPV - EHU