PROGRAMA EDUCATIVO
MAESTRIA EN CIENCIAS DE LA COMPUTACIÓN

UNIDAD DE APRENDIZAJE
MINERIA DE DATOS

UNIDAD DE COMPETENCIA IV
Métodos estimadores de error

ELABORACION
ADRIAN TRUEBA ESPINOSA

Fecha de elaboración Mayo de 2018

PRESENTACIÓN DEL CURSO

Una de las principales actividades del maestrante en la Maestría en Ciencias de la Computación , en la área terminal de “Sistemas de Información”, es la búsqueda de patones que puedan conformarse a partir de datos. La minería de datos es esencial, sin embargo existe la posibilidad de tener errores por lo que es necesario conocer algunos estimadores del error. Con esta unidad el estudiante adquirirá el conocimiento necesario para estimar el grado de error que hay en una investigación.

Al finalizar el curso el alumno será capaz de aplicar técnicas de estimación de errores en la búsqueda de patrones con técnicas de minería de datos

CONTENIDO DEL CURSO

Unidad 1: Métodos para el tratamiento y análisis de datos

Unidad 2: Procesos de análisis supervisado

Unidad 3:Proceso de análisis no supervisado

Unidad 4: Métodos estimadores de error

Unidad 5:Método para análisis del índice de aciertos

METAS A ALCANZAR

Que el alumno conozca los elementos teóricos y prácticos de los estimadores de error, que se emplean en la evaluación de las técnicas de minería de datos.

OBJETIVO DEL MATERIAL DIDÁCTICO

Conocer las técnicas para estimar los errores en la minería de datos

UNIDAD DE COMPETENCIA IV

Métodos estimadores de error

ESTIMADORES

Un estimador es un estadístico (una función de la muestra) utilizado para estimar un parámetro desconocido de la población. De la Fuente F. S.(s/f)

El valor de un estimador proporciona una estimación puntual del valor del parámetro en estudio. En general, se realiza la *estimación* mediante un intervalo, es decir, se obtiene *un intervalo* [parámetro muestral $\pm$ error muestral] dentro del cual se espera se encuentre el valor poblacional dentro de un cierto nivel de confianza. El nivel de confianza es la probabilidad de que a priori el valor poblacional se encuentre contenido en el intervalo. De la Fuente F. S.(s/f)

Sesgo

Se denomina sesgo de un estimador a la diferencia entre la esperanza (valor esperado) del estimador y el verdadero valor del parámetro a estimar. Es deseable que un estimador sea insesgado o centrado, esto es, que el sesgo sea nulo para que la esperanza del estimador sea igual al valor del parámetro que se desea estimar.

Por ejemplo, si se desea estimar la media de una población, la media aritmética de la muestra es un estimador insesgado de la misma, ya que la esperanza (valor esperado) es igual a la media poblacional. De la Fuente F. S.(s/f)

Aspectos de la Evaluación

- Fiabilidad de las diferencias estimadas en el rendimiento.

- Elección de la medidas del rendimiento.
  - Número de clasificaciones correctas.
  - Precisión de las estimaciones de probabilidad.
  - Error en predicción numérica.

- Costes asignados a distintos tipos de error.
  - En muchas aplicaciones prácticas.

Rodríguez D. J. J. (S/f).

Entrenamiento y Test

• En clasificaciones la medida natural del rendimiento es la tasa de error.

Acierto: la clase se predice correctamente.

Error: la clase se predice incorrectamente.

Tasa de error: proporción del número de errores cometidos sobre

todo el conjunto de ejemplos.

• Error de resubstitución : tasa de error obtenida sobre el conjunto de entrenamiento.

Inevitablemente optimista.

Rodríguez D. J. J. (S/f).

Evaluación

• Cómo de bueno es prediciendo el modelo que hemos aprendido.

• El error en el conjunto de entrenamiento no es un buen indicador del error sobre datos nuevos.

Almacenar los datos sería el clasificador óptimo.

• Rendimiento futuro sobre nuevos datos.

• Conjunto independiente de los datos de entrenamiento: datos de test.

• Normalmente solo se dispone de un conjunto de datos etiquetado.

• Si tenemos muchos datos etiquetados, dividir en entrenamiento y test.

• A menudo los datos etiquetados son limitados.

Técnicas más sofisticadas.

Concentración (asociado a la precisión del estimador). Criterios: ECM e insesgamiento)

EFICIENCIA Un estimador es más eficiente o más preciso que otro estimador, si la varianza del primero es menor que la del segundo.

CONSISTENCIA Si no es posible emplear estimadores de mínima varianza, el requisito mínimo deseable para un estimador es que a medida que el tamaño de la muestra crece, el valor del estimador tienda a ser el valor del parámetro poblacional, propiedad que se denomina consistencia.

SUFICIENCIA Un estimador θ es suficiente cuando no da lugar a una pérdida de información. Es decir, cuando la información basada en θ es tan buena como la que hiciera uso de toda la muestra

De la Fuente F. S.(s/f)

Evaluación de un clasificador

Random subsampling" (muestreo aleatorio)

Repite el método *holdout muchas veces y* calcular estadísticos sobre dicho proceso

La exactitud del modelo es dada por el promedio

Se sugiere repetir como mínimo 30 veces

No hay control sobre los ejemplos que ya han sido usados para entrenamiento

León G. E. (s/f).

ERROR CUADRÁTICO MEDIO DE LOS ESTIMADORES (ECM)

La utilización de la estimación puntual como si fuera el verdadero valor del parámetro conduce a que se pueda cometer un error más o menos grande.

El Error Cuadrático Medio (ECM) de un estimador $\hat{\theta}$ viene definido:

$$ECM(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = V(\hat{\theta}) + \left[\underbrace{E(\hat{\theta}) - \theta}_{\text{sesgo}}\right]^2 \quad \text{siendo el sesgo } b(\hat{\theta}) = E(\hat{\theta}) - \theta$$

Cuando el estimador es centrado, el sesgo $b(\hat{\theta}) = 0 \rightarrow ECM(\hat{\theta}) = V(\hat{\theta})$

Un error cuadrático medio pequeño indicará que en media el estimador $\hat{\theta}$ no se encuentra lejos del parámetro θ

De la Fuente F. S.(s/f)

Consistencia

- Un estimador es consistente si al aumentar el tamaño de muestra el estimador se acerca más al parámetro.

$$T \text{ estimador de } \theta$$

$$T_1, \cdots, T_n \equiv T \text{ muestras tamaño } 1,2,\cdots,n$$

$$T \text{ es consistente si}$$

$$\lim_{n\to\infty} P(|T_n - \theta| < \varepsilon) = 1 \text{ (converge en probabilidad)}$$

Eficiencia

- Si T_1 y T_2 son dos estimadores insesgados de θ T_1 es más eficiente que T_2 si

$$V(T_1) < V(T_2),$$

$$Eficiencia \quad relativa < 1, con$$

$$\frac{V(T_1)}{V(T_2)} \equiv Eficiencia\ relativa$$

- Dentro de los estimadores insesgados de θ , el que tiene la varianza más pequeña se llama UMVUE (estimador insesgado de varianza mínima)

Suficiencia

- Una estadística es suficiente si utiliza toda la información de la muestra respecto al parámetro

$$X_1, X_2, \cdots X_n \; ma$$

$$T_1 = \frac{X_1 + X_3 + \cdots + X_{n-1}}{n/2}$$

$$T_2 = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

- T_2 suficiente y T_1 no

Error de clasificación

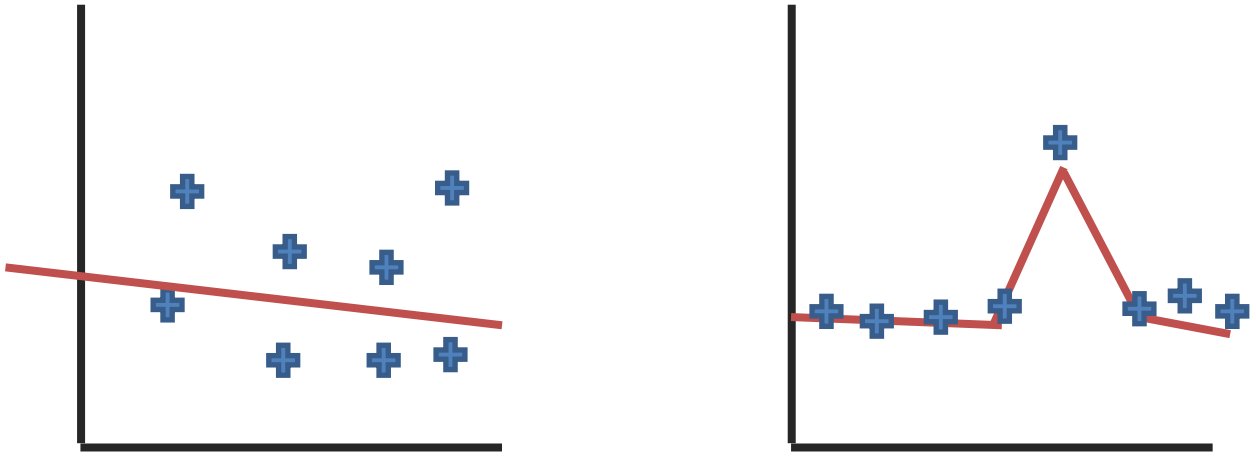
- **Error de entrenamiento:**
  e(modelo, datos)
  Número de ejemplos de entrenamiento
  clasificados incorrectamente
  Conocido como error de re-substitución o error aparente
- **Error de generalización:**
  e'(modelo, datos)
  Error esperado del modelo en ejemplos no usados en el entrenamiento.

Un buen modelo debe tener errores de entrenamiento y generalización bajos

León G. E. (s/f).

Errores de clasificación

Se dice que cuando un algoritmo de aprendizaje se ajusta a los datos de entrenamiento pierde su capacidad de generalización, y con ello deja de ser útil.
A esto se le conoce como **Sobre ajuste o (Overfitting)**



Sobre-ajuste: Bajo error de entrenamiento pero error de generalización alto

Sub-ajuste (underfitting): Errores de entrenamiento y generalización altos

Métodos de Estimación

Holdout

El conjunto de datos original es dividido en la muestra de entrenamiento (para entrenar al clasificador) y en la muestra de prueba (que valida al clasificador).

El método de Hold Out se puede usar para comparar la eficiencia entre algoritmos de clasificación

$$E_{Ho} = \frac{1}{|R_2|} \sum_{X_i \in R_2} E(X_i, Y)$$

$|R2| =$ conjunto de muestra

$E(Xi,Y) =$ estimador de error

$$E(X_i, Y) = \begin{cases} 0 \text{ si } T(X_i) = T(Y) \\ 1 \text{ si } T(X_i) \neq T(Y) \end{cases}$$

$T(Xi) =$ etiqueta de entrenamiento

$T(Y) =$ etiqueta de prueba

Métodos de Estimación

Holdout

Técnicas para evitar sesgo en Holdout

• **Holdout estratificado**: – Clases ocurren con la misma frecuencia en partición entrenamiento/prueba. – Salvaguarda básica para sesgo.

• **Holdout repetitivo:** – Repetir la prueba varias veces pero cambiando la partición entrenamiento/prueba. – Error estimado: promedio de errores de cada iteración

Hurtado L.C. (S/F)

Métodos de Estimación

Holdout

- Si la cantidad de datos es limitada.

- Holdout: reserva una cantidad para test, el resto para entrenamiento.
  E.g., un tercio para test.
- Problema: las muestras podrían no ser representativas.
  E.g., una clase podría no estar presente.
- Estratificación: asegura que cada clase está representada con aproximadamente las mismas proporciones en los dos subconjuntos.

Rodríguez D. J. J. (S/f).

Métodos de Estimación

Holdout repetid

Más fiable si repetimos el proceso varias veces con diferentes muestras.
En cada iteración se selecciona aleatoriamente una proporción para entrenamiento (posiblemente con estratificación).

Las tasas de error de las diferentes iteraciones se promedian para obtener la tasa de error global.

- No es óptimo, los diferentes conjuntos de test se solapan.
  Cómo prevenir el solapamiento.

Rodríguez D. J. J. (S/f).

Métodos de Estimación

Coss Validation

Es un método derivado de Hold Out, el cual divide la muestra de entrenamiento original en cierto número de particiones fijas disjuntas. Estos subconjuntos se van alternando quedando uno fuera, el resto son usados para el entrenamiento del clasificador

$$E_{cv} = \frac{1}{|Rv|}\sum_{v=1}^{v} E(Xi, Y)$$

$|R_v|$ = subconjuntos

V =1,2,3,4,…….,v.

Métodos de Estimación

Coss Validation

- Evita el solapamiento de los conjuntos de test.

  Primer paso: repartir los datos en k subconjuntos del mismo tamaño.

  Segundo paso: usar cada subconjunto como test, el resto para entrenamiento.

- k-fold cross-validation.

- A menudo los subconjuntos se estratifican antes de realizar la validación cruzada. • Se promedian las tasas de error.

Rodríguez D. J. J. (S/f).

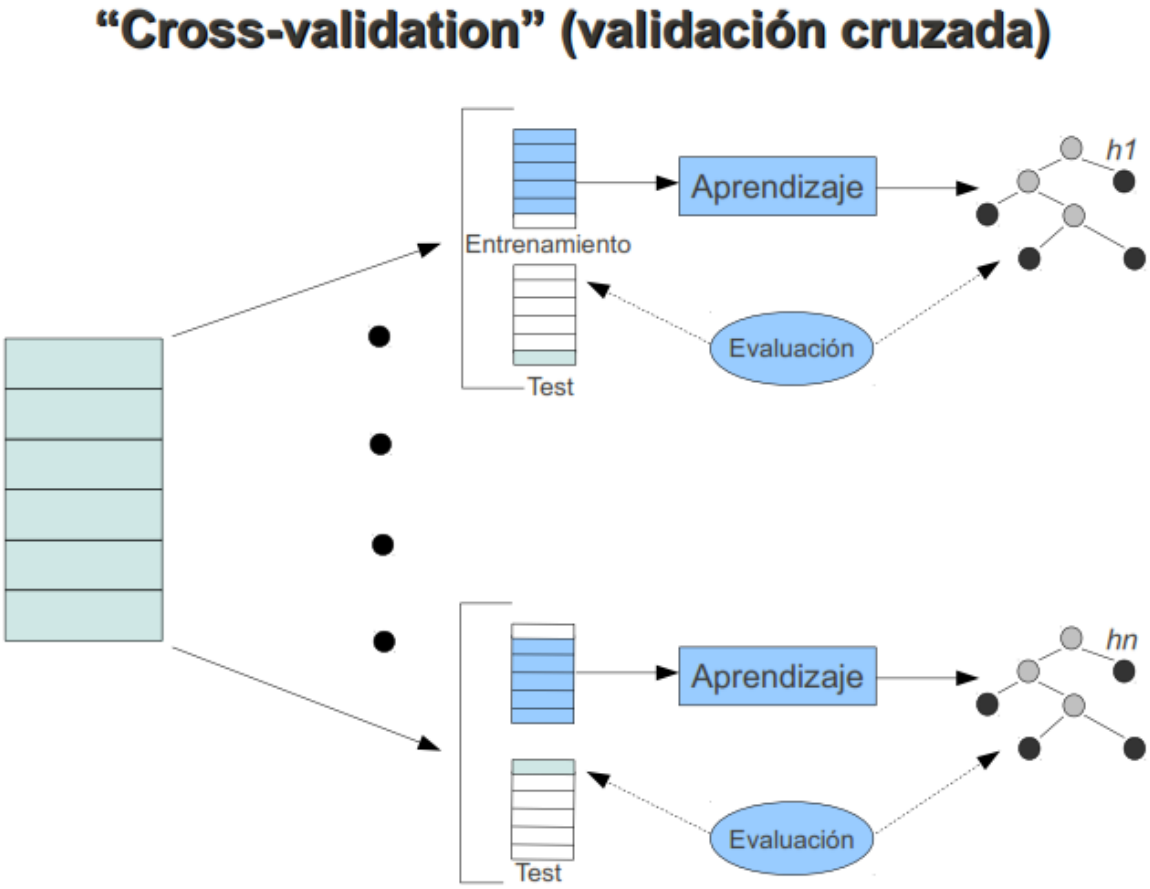
Métodos de Estimación

Coss Validation

- Estándar: 10 fold stratified cross validation.
  Apoyado por experimentación exhaustiva.
- La estratificación reduce la varianza del estimador.
- Ni la estratificación ni la división tienen que ser exactas.
- Validación cruzada repetida.
  Para paliar la influencia de la partición aleatoria.
  E.g.: 10×10, 5×2...

Hurtado L.C. (S/F)

Coss Validation



León G. E. (s/f).

Métodos de Estimación

Validación cruzada (“cross validation”)

- Forma de Hold-out repetitivo
- Validación cruzada de “n-fold”

– Datos se dividen en un número n fijo de subconjuntos

– Dado un subconjunto s, se usa s como prueba y los datos restantes como entrenamiento.

– Esto se repite para cada subconjunto

Hurtado L.C. (S/F)

Métodos de Estimación

Validación cruzada (“cross validation”)

• Error estimado: promedio de los errores en cada iteración

• Se puede usar estratificación

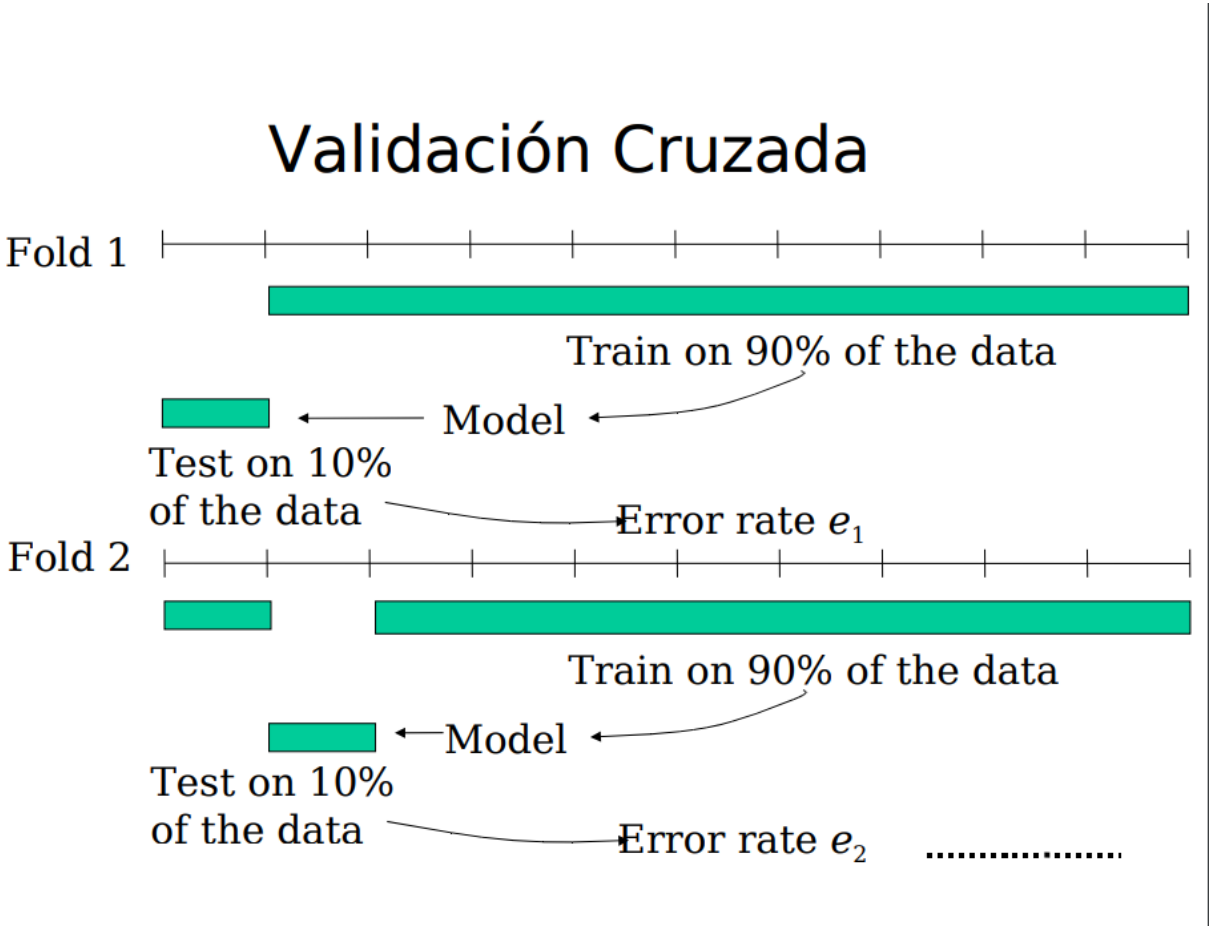
• Desventaja: costo computacional. Se debe inducir el modelo n veces. – No es factible para conjuntos de datos grandes.

Uso típico: 10 veces validación cruzada de 10-fold

• “leave-one-out”: caso particular, donde n es número de datos – útil cuando se tienen pocos datos.

León G. E. (s/f).

Métodos de Estimación



León G. E. (s/f).

Métodos de Estimación

Estimación del error

$$\bar{Y} = \frac{1}{k} \sum_{i=1}^{k} \bar{X}_k$$

La variable $\bar{Y}$ tiene media

Desviación estándar de $\bar{Y}$

$$s_{\bar{Y}^2} = \frac{1}{(k-1)} \sum_{i=1}^{k} (\bar{X}_k - \bar{Y})^2$$

León G. E. (s/f).

Métodos de Estimación

- Tenemos $$\frac{\bar{Y}-\pi}{\sqrt{\frac{s_{\bar{Y}^2}}{k}}} \sim t_{k-1}$$

- Obtenemos el intervalo (95% de confianza):

$$\pi = \bar{Y} \pm t_{.025} \frac{s_{\bar{Y}}}{\sqrt{k}}$$

León G. E. (s/f).

Métodos de Estimación

Método de Leave On Put (Deja uno fuera)

Se lleva acabo cuando cada patrón del conjunto de datos es alternado para validar el clasificador y el resto es usado para entrenamiento

$$E_L = \sum E(Xi, Y)$$

$E(X_i, Y)$ = estimador de error
m 0 número de particiones
X_i = etiqueta de entrenamiento
Y = etiqueta de prueba

Métodos de Estimación

Método de Leave On Put (Deja uno fuera)

Validación cruzada con tantos grupos como ejemplos.

- Ventajas:

  Cantidad máxima de datos para entrenamiento.

  Determinista.

- Inconveniente: muy costoso computacionalmente.

  Excepciones, e.g., vecino más cercano.

Rodríguez D. J. J. (S/f).

Métodos de Estimación

- No es posible estratificar.

  El conjunto de test solo tiene un ejemplo.

- Ejemplo artificial: conjunto completamente aleatorio con el mismo número de ejemplos de las dos clases.

  Mejor clasificador: predecir la mayoría.
  Sobre un conjunto nuevo de datos, acierto del 50%.
  De acuerdo a LOO, 100% de error.

BIBLIOGRAFIA

De la Fuente F. S.(s/f) Estimadores. Gestión Aeronáutica: Estadística Teórica Facultad de Ciencias Económicas y Empresariales . Departamento de Economía Aplicada
http://www.fuenterrebollo.com/Aeronautica2016/estimadores.pdf

Gómez F.W (s/f) Análisis de datos. Validación de clasificadores. CINVESTAV
https://www.tamps.cinvestav.mx/~wgomez/diapositivas/RP/Clase14.pdf

Hurtado L.C. (S/F).Evaluación de modelos (II). Departamento de Ciencias de la Computación. U de Chlie.
http://docplayer.es/60756150-Evaluacion-de-modelos-ii-carlos-hurtado-l-departamento-de-ciencias-de-la-computacion-u-de-chile.html

León G. E. (s/f). Minería de datos. Ingeniería de sistemas. Grupo Investigación MIDAS. Universidad Nacional de Colombia.
http://disi.unal.edu.co/~eleonguz/cursos/mda/presentaciones/Sesion10_Evaluacion.pdf

Perote P. J. (s/f). Especificación de Modelos Econométricos
http://campus.usal.es/~ehe/perote/documentos/TEMA%201%20MODELIZACI%C3%93N%20ECON%C3%93MICA%20II.pdf

Rodríguez D. J. J. (S/f). Metodología Experimental. Doctorado en Informática . Métodos y Técnicas de Minería de datos
https://www.infor.uva.es/~calonso/MUI-TIC/MineriaDatos/tr_metodologia.pdf

BIBLIOGRAFIA COMPLEMENTARIA

J. Demšar. Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research, 7:1–30, 2006.

Thomas G. Dietterich. Approximate statistical test for comparing supervised classification learning algorithms. Neural Computation, 10(7):1895–1923, 1998.

Tom Fawcett. An introduction to roc analysis. Pattern Recognition Letters, 27(8):861–874, June 2006.

C. Ferri, J. Hernandez-Orallo, and R. Modroiu. An experimental comparison of performance measures for classification. Pattern Recognition Letters, September 2008.

SalvadorGarcíaandFranciscoHerrera. Anextensionon“statisticalcomparisonsof classifiersovermultipledatasets”forallpairwisecomparisons. JournalofMachine Learning Research, 9:2677–2694, December 2008.

Rodríguez D. J. J. (S/f).

BIBLIOGRAFIA COMPLEMENTARIA

T. Mitchell. Machine Learning. McGraw Hill, 1997. [NB03]
ClaudeNadeauandYoshuaBengio. Inferenceforthegeneralizationerror. Machine Learning, 52(239–281), 2003.

Steven L. Salzberg. On comparing classifiers: Pitfalls toavoid and a recommended approach. Data Min. Knowl. Discov., 1(3):317–328, 1997.

I. H. Witten and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, 2nd edition, 2005.

Rodríguez D. J. J. (S/f).