Técnicas escalables de análisis de datos en entornos Big Data: Clasificadores

Metodología experimental de evaluación y selección de modelos

Metodología experimental: creación y evaluación de clasificadores mediante la tasa de error

Contenido

1. Motivación.
2. Error.
3. Estimación del error.
4. Metodología experimental para estimar la tasa de error de una hipótesis.
5. Evaluación y selección de modelos.
6. Sobre la tasa de error.
7. Referencias.

1. Motivación

- Nos limitamos al ámbito de los clasificadores
  - Aprendizaje inductivo basado en ejemplos, supervisado, predicción valor categórico.
- Recordar: cada modelo inducido es una posible hipótesis.

- ¿Cuál es la calidad de una hipótesis?
  - Múltiples criterios
    - Conocimiento aprendido: calidad, novedad, utilidad…
    - Capacidad para describir ejemplos no utilizados para entrenar: calidad como clasificador, coste aplicación,…

- Primera aproximación: tasa de error.

- ¿Cómo utilizar los datos para crear la hipótesis y al mismo tiempo evaluarla?

2. Error

- Medida natural del rendimiento de un clasificador.
- Acierto: el clasificador predice correctamente la clase.
- Error: el clasificador predice incorrectamente la clase.
- Tasa de error: proporción del número de errores cometidos sobre un conjunto de instancias.

Error verdadero, $e_D(h)$

- Def. El error verdadero, $e_D(h)$, de una hipótesis h respecto a un concepto objetivo c y distribución de probabilidad del espacio de instancias D es:

$$e_D(h) = Pr_{x \in D}[c(x) \neq h(x)]$$

Tasa de error, $e_S(h)$

- Def. La tasa de error, $e_S(h)$, de una hipótesis h respecto a un concepto objetivo c y una muestra $S \subset X$ es:

$$e_S(h) = 1/card(S) \times \sum_{x \in S} \delta(c(x), h(x))$$

$\delta(c(x), h(x)) = 1$ *sii* $c(x) \neq h(x)$
$\delta(c(x), h(x)) = 0$ *sii* $c(x) = h(x)$

Error de resubstitución, e_r

- Tasa de error calculada sobre el conjunto de entrenamiento.

- Muy optimista.
- Estima tasas de error menores que el error verdadero.

Error verdadero y tasa de error

- La tasa de error es una estimación del error verdadero.

- ¿En qué circunstancias es una buena estimación?
- No basta con $card(S)$ "suficientemente grande".

- Cuando **la hipótesis h y el conjunto S** con el que se obtiene la tasa de error **son independientes**.

- El conjunto S debe seleccionarse con independencia de la hipótesis y sus elementos no deben utilizarse para la creación del clasificador de ninguna manera.

3. Estimación del error

- Proceso de Bernoulli y distribución binomial.
- Ejemplo: ¿probabilidad de obtener 3 caras lanzando una moneda 5 veces al aire?

- Ensayo de Bernoulli o Experimento base: lanzar moneda al aire.
  - La variable aleatoria solo toma uno de dos valores (distribución binomial)
- Proceso de Bernoulli: repetición del experimento base
  - Sucesos independientes, probabilidad constante.

- Problema: determinar la probabilidad del experimento base conocido el resultado de un Proceso de Bernoulli.

Tasa de error: Proceso de Bernoulli

- El problema de determinar el error verdadero, $e_D(h)$, conocida la tasa de error, $e_S(h)$, se puede plantear como un Proceso de Bernoulli:
  - Experimento base: determinar el error verdadero de h sobre una instancia cualquiera.
    - La variable aleatoria solo toma uno de dos valores.
  - Proceso de Bernoulli: repetición sobre los elementos de S
    - Sucesos independientes, probabilidad constante.

- Resultado del Proceso de Bernoulli: Errores cometidos sobre S (a partir del cual obtenemos $e_S(h)$)
- Probabilidad de error al realizar experimento base: error verdadero, $e_D(h)$

- Siempre que los sucesos sean **independientes**: h independiente de S

Estimación de la tasa de error con suficientes ejemplos

- Si S contiene n ejemplos seleccionados de forma aleatoria según probabilidad D, no utilizados de ninguna manera para la creación de la hipótesis h que clasifica mal r ejemplos de S.
- Si $n >= 30$ $(np(1-p) >= 5)$, aproximamos binomial por normal.

1. Tasa de error: $e_S(h) = r/n$
2. Error verdadero: $e_D(h)$
3. $E[e_D(h)] = E[e_S(h)]$
4. Desviación estándar $\sigma_{e_s(h)} \approx \left(\frac{e_{S(h)}(1-e_{S(h)})}{n}\right)^{1/2}$
5. Con probabilidad $N\%$, $e_D(h)$ está en el intervalo

$$e_{S(h)} \pm z_N \times \left(\frac{e_S(h)(1-e_S(h))}{n}\right)^{1/2}$$

   - Donde z_N lo obténtenos de las tablas (de dos colas) de la distribución Normal para la confianza deseada.

4. Metodología experimental para estimar la tasa de error de una hipótesis

Suficientes datos

- T: conjunto de entrenamiento.
  - Con el que se crea la hipótesis.
- V: conjunto de validación.
  - Para ajustar la hipótesis (parámetros, estructura,…)
- P: conjunto de prueba.
  - Con el que se calcula la tasa de error.

- Selección aleatoria de T, V y P. **Los elementos de P no pueden utilizarse en T y V de ninguna forma**.
- Cuanto más grandes, mejor.
- Construir clasificador con T, V (V no es necesario si validación cruzada interna)
- Tasa de error con P (según resumen pag. 11).

- Una vez evaluada la tasa de error, **TODOS** los datos pueden utilizarse para construir el clasificador. No necesario en un contexto Big Data.

Todos los datos para construir el clasificador

- Suposición: el error verdadero disminuye con el tamaño del conjunto de entrenamiento
  - Hipótesis de trabajo.
  - Razonable si los ejemplos se eligen de forma aleatoria.
  - Contrastada experimentalmente
    - Menor disminución con mayor tamaño del conjunto.
- Consecuencia: la tasa de error es una estimación pesimista
  - El error verdadero con más datos será menor que el error verdadero con el conjunto de entrenamiento.
- No son admisibles las estimaciones optimistas.

Insuficientes datos /ajuste de parámetros

- Es necesario reservar algunos datos para la evaluación.
- Que no se hayan utilizado para el ajuste de parámetros.

- Big data: no es problemático, pues dispone de suficientes datos.
- Problemático si pocos datos.
  - Normalmente, cuanto más grande sea el conjunto de entrenamiento mejor será el clasificador (mejoras cada vez más pequeñas).
  - Cuanto más grande sea el conjunto de test, más precisa será la estimación del error.

- Una vez que la evaluación se ha completado, se pueden usar **todos los datos** disponibles para construir el clasificador final.
- De nuevo, si la escala es Big Data, esto puede no ser necesario (ni posible).

Holdout (reserva)

- Dividir los datos originales en entrenamiento y prueba.
  - Dilema: idealmente, ambos conjuntos deberían ser grandes.
  - Buen clasificador o buena estimación del error.

- Típicamente $1/3$: $2/3\ T$, $1/3\ P$, de forma aleatoria.
- Big data 80%, 20% es razonable pues el conjunto de prueba es suficientemente grande.

- Inconveniente: las muestras podrían no ser representativas.
  - E.g., una clase podría no estar presente.
  - Incluso en big data.

- **Estratificación**: asegura que cada clase está representada con aproximadamente las mismas proporciones en los dos subconjuntos.

Holdout repetido

- La estimación de *hold out* depende de la partición aleatoria y tiene mucha variabilidad.
- Solución: repetir el proceso y promediar la tasa de error.

- Repetir el proceso k veces con diferentes muestras
  - En cada iteración se elijen nuevos T y P de forma aleatoria, manteniendo la proporción.

- Tasa de error: promedio, $e(h)=\sum_{i=1,k} e_i(h)/k$

- Varianza, estimación normal: optimista, no recomendable pues *h* y *P* no son tan independientes (peor aún si *n<30*).

Holdout repetido, varianza

- Más realista: estimar varianza muestral, S^2:

$$S^2_{e(h)} = 1/(k-1) \times \sum_{i=1,k} (e_i(h) - e(h))^2$$

- Intervalos de confianza: t-student.

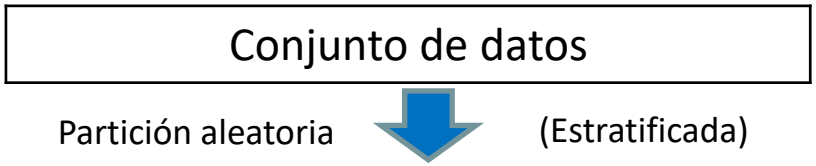
- Con probabilidad $N\%$, $e_D(h)$ está en el intervalo:

$$e(h) \pm t_{N,k-1} \times S_{e(h)}/\sqrt{k}$$

- $t_{N,k-1}$: *t-student* de $k-1$ grados de libertad para confianza $N\%$
  - Si $k \gg$, t tiende a la distribución normal.

Validación cruzada

- *Holdout* repetido no es óptimo: los conjuntos de prueba se solapan
- Validación cruzada evita el solapamiento de los conjuntos de prueba
- **Validación cruzada de k particiones** (k-*fold cross validation*, $k - XV$)
  - Primer paso: repartir los datos en k subconjuntos del mismo tamaño.
  - Segundo paso: usar cada subconjunto como prueba, el resto para entrenamiento.



Iteración	Entrenamiento	Prueba
1	S_2, S_3	S_1
2	S_1, S_3	S_2
3	S_1, S_2	S_3

- Variante estratificada: se estratifican los conjuntos en el primer paso.

Validación cruzada: estimación del error

- Tasa de error, promedio: $e(h) = \sum_{i=1,k} e_i(h)/k$

- Estimar varianza muestral, S^2:

$$S^2_{e(h)} = 1/(k-1) \times \sum_{i=1,k} (e_i(h) - e(h))^2$$

- Intervalos de confianza: *t-student.*
- Con probabilidad $N\%, e_D(h)$ está en el intervalo:

$$e(h) \pm t_{N,k-1} \times S_{e(h)}/\sqrt{k}$$

Validación cruzada

- Estándar: 10-fold *stratified cross validation*.
  - Habitual, pero nada especial con el 10.
  - En un contexto Big Data, 5, incluso 3 *folds*, pueden ser suficientes.
- La estratificación reduce la varianza del estimador.
- Ni la estratificación ni la división tienen que ser exactas.
- La estima del error se ve afectada por la partición aleatoria.

- Validación cruzada repetida.
  - Para paliar la influencia de la partición aleatoria.
  - E.g.: $10 \times 10, 5 \times 2 \cdots$

Validación cruzada repetida: estimación del error

- Validación cruzada repetida $R \times k$ *(Repetitions* $\times$ *k-folds).*
- Tasa de error, promedio: $e(h)=[\sum_{i=1,R\times k} e_i(h)]/(R \times k)$

- Estimar varianza muestral, S^2:

$$S^2_{e(h)} = 1/(R \times k-1) \times \sum_{i=1,R\times k} (e_i(h)- e(h))^2$$

- Intervalos de confianza: *t-student.*
- Con probabilidad $N\%$, $e_D(h)$ está en el intervalo:

$$e(h) \pm t_{N,R\times k-1} \times S_{e(h)}/\sqrt{R \times k}$$

- Ligeramente optimista: las repeticiones no son totalmente independientes.

Comparación de métodos de estimación del error.

Método	Características
Error de resubstitución	Optimista
Holdoud	Pesimista, muy muy variable
Holdoud repetido	Pesimista, muy variable
Validación cruzada	Menos sesgado, variable
Validación cruzada repetida	Menos variable, más costoso

5. Evaluación y selección de modelos

- Necesario para el ajuste de parámetros de un clasificador.

- *D*: conjunto de datos disponibles.
- Realizamos una partición aleatoria: ***T, P***.

- Utilizamos **T** para **entrenar y seleccionar el modelo.**
- Por lo tanto, hay que realizar evaluaciones honestas sobre T para determinar buenos valores de los parámetros del modelo
  - ***Entrenamiento y validación***.
  - ***Validación cruzada interna***.

- Una vez fijados los parámetros, crear clasificador con ***T***.
- **Evaluar con *P*.**
- Crear clasificador final con *D (quizás no en Big data).*

Entrenamiento y validación (y evaluación)

- En Big Data, de utilidad si no tenemos suficientes recursos

- *D*: conjunto de datos disponibles.
- Realizamos una participan aleatoria: ***T, P.***

- Utilizamos **T** para **entrenar y seleccionar el modelo**
- Mediante un proceso de ***Entrenamiento y validación***
  - Partición aleatoria de ***T*** *en* ***T'*** *y* ***V***
  - Creamos modelos con T'
  - Los evaluamos sobre V, para encontrar unos buenos valores de los parámetros del modelo

- Una vez fijados los parámetros**, crear clasificador con *T***.
- **Evaluar con *P***.
- Crear clasificador final con *D (quizás no en Big data)*.

Validación cruzada interna (y evaluación)

- Habitual para el ajuste de parámetros de un clasificador.

- *D*: conjunto de datos disponibles.
- Realizamos una participan aleatoria: ***T, P***.

- Utilizamos **T** para **entrenar y seleccionar el modelo**
- Mediante un proceso de ***Validación cruzada interna***
  - Creamos y evaluamos modelos mediante validación cruzada sobre ***T,*** para encontrar unos buenos valores de los parámetros del modelo.

- Una vez fijados los parámetros, **crear clasificador con *T***.
- **Evaluar con *P*.**
- Crear clasificador final con *D (quizás no en Big Data)*

6. Sobre la tasa de error

- La tasa de error de un clasificador permite evaluar la calidad de una hipótesis (un modelo).
- No proporciona información sobre la calidad de un algoritmo de aprendizaje.
- El mismo algoritmo podría generar hipótesis con mayor/menor tasa de error
  - Misma hipótesis sobre otro conjunto de prueba.
  - Otro conjunto de aprendizaje.
  - Otro dominio de aplicación.
- La comparación de algoritmos requiere test de hipótesis.
- **Ningún algoritmo es mejor que otro sobre cualquier dominio de aplicación.**

Limitaciones de la tasa de error

- Hasta ahora, hemos utilizado la tasa de error como criterio para evaluar hipótesis.
- Suposiciones:
  - Distribuciones de clases no muy desequilibradas.
  - Los costes de los errores son los mismos.
- No siempre es así
  - 99,99% de la población no es terrorista
    - «No terrorista» cierto 99,99%
  - El 97% de los días del mes las vacas no están en celo
    - «No celo» cierto 97,00%
- ¿Coste de diagnosticar erróneamente la presencia de una enfermedad frente a no diagnosticar una enfermedad real?

7. Referencias

- Tom M. Mitchell. Machine Learning. McGraw-Hill, 1997.
- Ian H. Witten, Eibe Frank, Mark A. Hall, Christopher Pal. Data Mining: practical machine learning tools and techniques (4th Edition). Morgan Kaufmann, 2016. ISBN: 9780128042915,
- David Page. Evaluating Machine Learning Methods. https://pages.cs.wisc.edu/~dpage/cs760/evaluating.pdf. Último acceso: octubre 2025.

Para los más teóricos:

- Arlot, S. and Celisse, A. A survey of cross-validation procedures for model selection. Statistics Surveys, 4, 40-79, 2010. DOI:10.1214/09-SS054. Disponible en: https://projecteuclid.org/journals/statistics-surveys/volume-4/issue-none/A-survey-of-cross-validation-procedures-for-model-selection/10.1214/09-SS054.full. Último acceso: octubre 2025.