UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

ESCOLA TÉCNICA SUPERIOR DE ENXEÑARÍA

Plataforma Web para la Validación de Experimentación en Aprendizaje Automático y Minería de Datos

TRABAJO DE FIN DE GRADO

Realizado por:
Adrián Canosa Mouzo

Dirigido por:
Ismael Rodríguez Fernández
Alberto J. Bugarín Díz
Manuel Mucientes Molina

Índice general

Índice de figuras

Capítulo 1

Introducción

En el contexto tecnológico actual, en donde el Big Data es un recurso cada vez más utilizado, el rol del analista de datos (data scientist) se está convirtiendo en una profesión emergente y de elevada demanda. Un analista de datos es aquel que reúne, analiza e interpreta los datos obtenidos con el objetivo de sacar ciertas conclusiones de ellos y así tomar diferentes decisiones, con las que aumentar la productividad. El analista de datos combina diferentes habilidades, especialmente las técnicas de la minería de datos y del aprendizaje automático (DM&M).

Según Mitchell [1], una definición de aprendizaje automático sería la siguiente: Un programa de ordenador aprende a partir de una experiencia E a realizar una tarea T (de acuerdo con una medida de rendimiento P), si su rendimiento al realizar T, medido con P, mejora gracias a la experiencia E.

Una de las tareas más importantes que se deben llevar a cabo es la validación de resultados obtenidos por los algoritmos de aprendizaje. El método estándar más aceptado en la actualidad es el de la aplicación de test estadísticos sobre los experimentos, que, entre otras utilidades, apoyan la toma de decisiones (por ejemplo la elección del algoritmo más adecuado).

Este proyecto se centra en crear y desarrollar una interfaz web para asistir al analista de datos en el proceso de validación de resultados. En este proyecto se desarrollará una plataforma web y se extenderá una librería de tests estadísticos con el objetivo de que el analista de datos pueda introducir en la web los datos obtenidos y seleccionar el test estadístico que desea utilizar para que, de forma automática, el sistema devuelva los resultados obtenidos después de la aplicación del test. Así, el sistema permitirá de un modo fácil y centralizado la utilización de tests estadísticos.

La herramienta se incorporará en la lista de aplicaciones disponibles a través de la web del CiTIUS para su uso. El impacto y difusión del resultado del proyecto va a ser amplio, ya que en la actualidad no existe ninguna herramienta que centralice la aplicación de los test estadísticos de mayor utilidad y que resulte fácil de usar.

1.1. Objetivos del proyecto

El proyecto se centra en crear y desarrollar una plataforma web para asistir al analista de datos en el proceso de validación de resultados obtenidos de diferentes algoritmos de aprendizaje. Para ello, habrá que desarrollar tres componentes diferenciados:

1. Completar y extender una librería de test estadísticos, actualmente implementada en el lenguaje Python.

2. Crear los servicios web en Python, basados en REST, que hagan disponible el acceso a los tests estadísticos vía web.

3. Desarrollar una interfaz web (HTML + JavaScript) para facilitar el uso de los tests sobre los datos introducidos por el analista de datos.

1.2. Contraste de hipótesis

El contraste de hipótesis, o lo que se conoce como tests estadísticos se engloba dentro de la rama de la estadística: *Inferencia Estadística*, que es una parte de la estadística que estudia cómo sacar conclusiones generales (sujetas a un determinado grado de fiabilidad o significancia) para toda la población a partir del estudio de una muestra. En nuestro caso, se tratará de sacar conclusiones de los resultados obtenidos por diferentes algoritmos (muestra) para determinar por ejemplo si los algoritmos tienen un rendimiento similar y por lo tanto se pueden considerar iguales (población).

El contraste de hipótesis es uno de los problemas más comunes dentro de la inferencia estadística. En él se contrasta una hipótesis estadística. Por ejemplo:

Un ingeniero de software afirma que la media de los resultados obtenidos por un algoritmo de aprendizaje automático es 10. ¿Se podría desmentir la afirmación del ingeniero?

El planteamiento del contraste sería el siguiente:

$$\mu = 10$$
$$\mu \neq 10$$

Para tomar una decisión (desmentir o no la afirmación), habría que basarse en los datos de una muestra, y ver si en efecto la media de los resultados es 10. Se podría establecer una regla de decisión sobre la cual se basaría nuestra decisión final. Por ejemplo: si la media obtenida está próxima a la afirmada por el ingeniero (10), entonces se podría afirmar dice la verdad. Si por el contrario la muestra nos proporciona una media muy distinta a 10, entonces se diría que el ingeniero miente sobre el algoritmo en cuestión. Esto supone un problema y es el hecho de cuándo considerar que la media es lo suficientemente distinta como para determinar que la afirmación del ingeniero es errónea. Por ejemplo si la media de la muestra es 8.5, ¿se podría desmentir la afirmación inicial? El contraste de hipótesis nos proporciona una forma de establecer este criterio y poder rechazar o aceptar la afirmación inicial.

1.2.1. Hipótesis nula y alternativa

En todo contraste de hipótesis siempre siempre se dan dos posibilidades o hipótesis, las cuales se representan con los siguientes símbolos:

$$H_0 : \text{Hipótesis nula}$$
$$H_1 : \text{Hipótesis alternativa}$$

A modo de ejemplo, supongamos que unos programadores están trabajando en la optimización de un algoritmo de aprendizaje. El objetivo es mejorar el algoritmo de forma que los resultados que proporcione sean menores de 100. Se toma una muestra de los resultados obtenidos por el nuevo algoritmo optimizado y se observa que la media de la muestra es de 92. Si no hubiera incertidumbre en la media muestral, entonces se podría concluir que la modificación reduciría los resultados a 92. Sin embargo, siempre existe incertidumbre en la media muestral. La media poblacional en realidad será poco mayor o menor a 92.

Los programadores están preocupados de que el nuevo algoritmo en realidad no mejore al anterior, es decir, que la media poblacional pudiera ser mayor o igual a 100. Quieren saber si esta preocupación está justificada. Se ha observado una muestra con media de 92 y existen dos posibles interpretaciones o, como se ha mencionado más arriba dos tipos de hipótesis que serán contrastadas más adelante mediante un determinado test estadístico:

1. La media poblacional es mayor o igual a 100 (la media muestral es, por tanto, menor debido sólo a la variación aleatoria de la media poblacional). El nuevo algoritmo no mejorará al anterior.

2. La media poblacional es menor que 100, y la media muestral lo refleja. El nuevo algoritmo sí mejorará al anterior.

La primera interpretación sería la hipótesis nula o H_0, que es la hipótesis que se supone cierta de partida, es decir, es la hipótesis que establece que lo que indica la muestra es solamente debido a la variación aleatoria entre la muestra y la población. La segunda interpretación o H_1, es la hipótesis alternativa y es la que reemplazará a la hipótesis nula si ésta es rechazada. H_1 establece que lo que indica la muestra es verdadero, ya que representa a toda la población.

En este caso, los programadores están preocupados de que la hipótesis nula sea cierta. Un test estadístico o prueba de hipótesis hallará, entre otras cosas una medida cuantitativa de la factibilidad de la hipótesis nula (denominado estadístico de contraste que para este ejemplo viene dado por la media obtenida en la muestra) y se podrá decir a los programadores (después de que el test tome la decisión) si su preocupación está o no justificada. Por tanto, a modo de resumen este ejemplo nos proporciona dos hipótesis:

$$H_0 : \mu \geq 100 \text{ vs. } H_1 : \mu < 100$$

La realización de un contraste de hipótesis no consiste en decidir cuál de las dos hipótesis (H_0, H_1) es más creíble, sino en decidir si la muestra proporciona o no suficiente evidencia para descartar H_0. Para realizar la prueba de hipótesis o test estadístico se pone la hipótesis nula en juicio, es decir se empieza suponiendo que H_0 es verdadera. Se podría poner como analogía el supuesto de *"En un juicio, el acusado siempre es inocente hasta que se demuestre lo contrario."* Es decir:

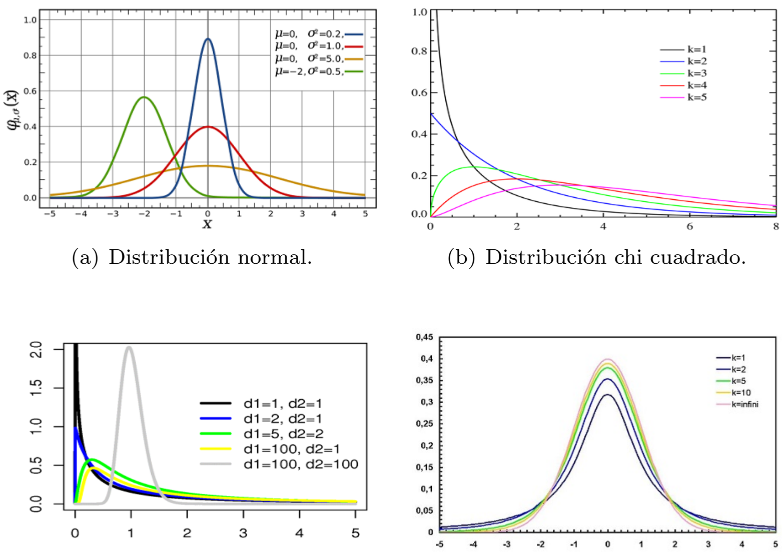
$$\begin{array}{l} H_0 : \text{El acusado es inocente} \\ H_1 : \text{El acusado es culpable} \end{array}$$

y, mientras no se tenga suficiente evidencia para aceptar H_1, hay que creer que lo que dice H_0 es cierto. La muestra aleatoria proporcionará la evidencia. Si el juicio (test o prueba de hipótesis) determina que el acusado es inocente, sólo se puede decir que no se tiene suficiente evidencia para asegurar que el acusado es culpable, mientras que si aceptamos la hipótesis alternativa, se estará bastante seguro de que el acusado sí es culpable.

1.2.2. Estadístico de contraste

Los tests estadísticos o pruebas de hipótesis, calculan internamente una medida cuantitativa que proporciona la factibilidad de la hipótesis nula. Esta medición se extrae de la muestra proporcionada. Por ejemplo, si queremos contrastar la hipótesis de que la media poblacional es 5, un estadístico a calcular puede ser la media de una muestra. En este caso, la muestra viene determinada por los resultados obtenidos por los algoritmos y cada uno de los tests tienen una forma particular de hallar este estadístico mediante una fórmula que lo caracteriza. Estos estadísticos siguen una determinada distribución de probabilidad. Por ejemplo en esta plataforma web, los tests implementados harán uso de estadísticos que siguen distribuciones como:

- Distribución normal (p. ej. test de Wilcoxon).
- Distribución chi cuadrado χ^2 (p. ej. test de Friedman).
- Distribución f de Fisher-Snedecor (p. ej. test de Iman-Davenport).
- Distribución t de Student (p. ej. t-test).



(a) Distribución normal. (b) Distribución chi cuadrado.

(c) Distribución f. (d) Distribución t.

Figura 1.1: Funciones de distribución de probabilidad [2].

La figura 1.1, nos muestra el aspecto que presentan las distintas distribuciones de probabilidad. Todas ellas son continuas, pues se puede tomar cualquier valor dentro de un intervalo, a diferencia de las distribuciones discretas.

Atendiendo a la distribución normal, μ y σ^2 son la media y la varianza respectivamente. La varianza, es una medida de dispersión que indica cómo se distribuye la población. Por ejemplo: en una distribución normal de media 0 y varianza 1, aproximadamente el 68 % de la población se encuentra en el intervalo $[-1, 1]$. Si un estadístico sigue una distribución normal con media μ y varianza σ^2, se expresa como:

$$\text{Estadístico} \sim N(\mu, \sigma^2)$$

En las otras tres distribuciones se habla del parámetro K o grados de libertad. Los grados de libertad son el número de valores que pueden ser asignados de forma arbitraria, antes de que el resto de las variables tomen un valor automáticamente, producto de establecerse las que son libres [3]. En nuestro caso, K viene determinado por el número de algoritmos o variables relacionadas que tiene la muestra de datos con la que se están aplicando los tests, y cada test que use este parámetro lo modifica de acuerdo a su fórmula característica para el estadístico. La media y la varianza de estas tres distribuciones vendrán determinadas por el parámetro k.

Las distribuciones de probabilidad que pueda seguir un estadístico nos dan un valor diferente de probabilidad para cada valor diferente del estadístico. Este valor de probabilidad indica lo probable que es obtener ese valor del estadístico siendo la hipótesis nula cierta. Por ej. si es cierta la hipótesis nula de que la media de una población es 5, es más probable que obtengamos una media de una muestra igual a 4.5 que a 3.

1.2.3. Decisiones y tipos de error

Cuando se lleva a cabo un contraste de hipótesis sólo se pueden tomar dos decisiones. Los datos de la muestra, evidenciarán qué decisión se debe tomar:

1. Aceptar la hipótesis nula (H_0) (Rechazar la hipótesis alternativa H_1)
2. Rechazar H_0 (Aceptar la hipótesis alternativa)

		Decisión	
		No se rechaza Ho	Se rechaza Ho
Realidad	Ho es verdadera	Decisión correcta	Error de tipo I
	Ho es falsa	Error de tipo II	Decisión correcta

Figura 1.2: Decisiones y tipos de error.

La muestra, en el caso de este proyecto, vendrá dada por los resultados obtenidos por los algoritmos. Sin embargo, cuando se toma la decisión se pueden cometer dos tipos de error. La figura 1.2, nos muestra las decisiones y los dos tipos de errores que se pueden cometer.

Se puede tomar una decisión correcta o errónea. La probabilidad de *"Error tipo I"* se denota por α y se denomina nivel de significación:

$$P(\text{Error tipo I}) = P(\text{Rechazar } H_0 | H_0 \text{ es cierta}) = \alpha.$$

La probabilidad de *"Error tipo II"* se denota por β:

$$P(\text{Error tipo II}) = P(\text{Aceptar } H_0 | H_0 \text{ es falsa}) = \beta.$$

Por otra parte, la *"Potencia"* es la probabilidad de detectar que una hipótesis es falsa. Los tests estadísticos o pruebas de hipótesis implementados en el presente proyecto serán mejores o peores dependiendo de su potencia:

$$P(\text{Potencia}) = P(\text{Rechazar } H_0 | H_0 \text{ es falsa}) = 1 - \beta.$$

El *"Error tipo I"*, es lo más peligroso que puede ocurrir, mientras que el *"Error tipo II"* no es tan importante. Volviendo al ejemplo de la analogía del juicio, es mucho peor condenar a un inocente que absolver a un culpable. Obviamente, lo ideal sería que tanto α como β fuesen nulos y que no se cometiese ningún error, o que ambos valores fuesen muy pequeños. Como no se pueden disminuir ambos errores a la vez, se controla el *"Error tipo I"*, que es el más importante. Este nivel de significación α es un parámetro que debe seleccionar la persona que quiere realizar un test estadístico en base a cómo de importante considere rechazar H_0 cuando es cierta.

1.2.4. Intervalos de confianza

El nivel de significación fijado divide en dos regiones el conjunto de posibles valores del estadístico de contraste: la región de aceptación y la región de rechazo o región crítica. Se denomina región de aceptación a la región que conduce a la aceptación de H_0 y región de rechazo a la región que conduce al rechazo de H_0 en favor de H_1.

La determinación de las regiones de aceptación o de rechazo dependen de cómo se establezca la hipótesis alternativa H_1. Por ejemplo, si hablamos de un contraste en el que se esté contrastando una determinada media (μ_0) se podría establecer como H_1 que la media en realidad sea menor, mayor o distinta a (μ_0):

Media menor (test unilateral con cola a la izquierda):

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu < \mu_0$$

Media mayor (test unilateral con cola a la derecha):

$$H_0 : \mu = \mu_0$$

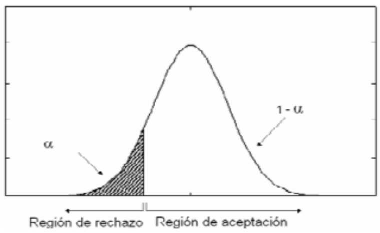
$$H_1 : \mu > \mu_0$$

Media distinta (test bilateral o de dos colas):

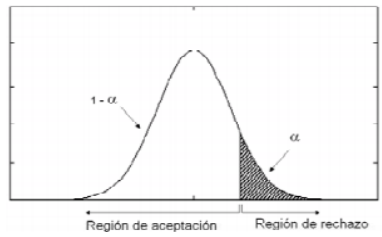
$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0$$

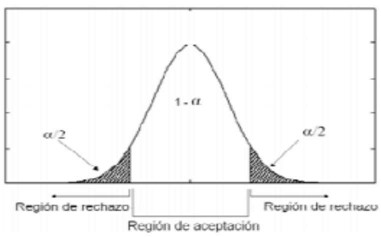
En la figura 1.3, podemos ver cómo quedarían establecidos los intervalos para el ejemplo:



(a) Test unilateral. Cola a la izquierda.



(b) Test unilateral. Cola a la derecha.



(c) Test bilateral o de dos colas.

Figura 1.3: Regiones de aceptación y rechazo.

1.2.5. Decisión final y Concepto de p-valor

Si el valor del estadístico cae en la región de aceptación, se acepta la hipótesis nula, ya que no existen razones suficientes para rechazar H_0 con el nivel de significación dado. Por tanto, en este caso se diría que el contraste es estadísticamente no significativo, es decir, no existe evidencia estadísticamente significativa en favor de H_1. La figura 1.4 nos muestra el punto crítico: si se supone por ejemplo que el estadístico obtenido por el test o prueba de hipótesis es 5 y el punto crítico fuese por ejemplo 4.5, ya no se podría aceptar H_0 ya que se caería en la región de rechazo.

La decisión de rechazar o aceptar la hipótesis nula, se puede determinar también mediante el p-valor, que es el parámetro utilizado para realizar los tests estadísticos en esta plataforma web. El p-valor proporciona un forma más eficiente de determinar si el contraste es o no estadísticamente significativo, ya que no sería necesario recalcular regiones de aceptación y rechazo cada vez que el usuario de los tests cambia de nivel de significación.

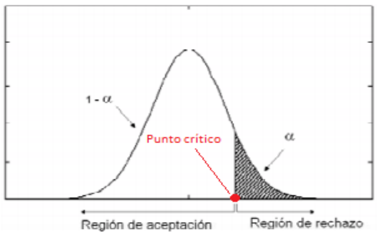


Figura 1.4: Punto crítico.

1.2.6. Tests paramétricos y no paramétricos

1.2.7. Etapas en la resolución de un contraste de hipótesis

En un contraste de hipótesis siempre se siguen una serie de pasos definidos. Como se ha ido viendo a lo largo del capítulo, los pasos para la realización de una prueba de hipótesis o test estadístico son las siguientes:

1. Especificación de la hipótesis nula H_0 y de la hipótesis alternativa H_1.

2. Suponer que H_0 es verdadera (el test sirve para que a partir de la muestra de datos podamos rechazar H_0 en beneficio de H_1).

3. Calcular un estadístico de prueba o estadístico de contraste. Este estadístico se usa para evaluar la fuerza de la evidencia en contra de H_0 (medir la discrepancia entre la hipótesis y la muestra).

4. Establecer un nivel de significación α en base a cómo de importante se considere rechazar H_0 cuando realmente es verdadera.

5. El nivel de significación fijado divide en dos regiones el conjunto de posibles valores del estadístico de contraste: la región de aceptación y la región de rechazo o región crítica.

6. Si el valor del estadístico cae en la región de rechazo, se rechaza la hipótesis nula, ya que esto evidencia que los datos obtenidos de la muestra no son compatibles con H_0. Por tanto, en este caso se diría que el contraste es estadísticamente significativo, es decir, existe evidencia estadísticamente significativa en favor de H_1.

7. Si el valor del estadístico cae en la región de aceptación, se acepta la hipótesis nula, ya que no existen razones suficientes para rechazar H_0 con el nivel de significación dado. Por tanto, en este caso se diría que el contraste es estadísticamente no significativo, es decir, no existe evidencia estadísticamente significativa en favor de H_1.

8. La decisión de rechazar o aceptar la hipótesis nula, se puede determinar también mediante el p-valor, que es el parámetro utilizado para realizar los tests estadísticos en esta plataforma web. El p-valor proporciona un forma más eficiente de determinar si el contraste es o no estadísticamente significativo, ya que no sería necesario recalcular regiones de aceptación y rechazo cada vez que el usuario de los tests cambia de nivel de significación.

1.3. Organización del documento

Bibliografía

[1] Tom M. Mitchell, *Machine Learning*, McGraw Hill, 1997.

[2] Categoría:Distribuciones continuas. Artículo de Wikipedia (*http://es.wikipedia.org/wiki/Categoria:Distribuciones_continuas*). Consultado el 26 de junio del 2014.

[3] Grado de libertad (estadística). Artículo de Wikipedia (*http://es.wikipedia.org/wiki/Grado_de_libertad_(estadistica)*). Consultado el 26 de junio del 2014.