Laboratorio 2
Estadística Computacional

Universidad Técnica Federico Santa María
Departamento de Informática

Sebastián Bórquez
<sborquez@alumnos.inf.utfsm.cl>

José García
<jigarcia@alumnos.inf.utfsm.cl>

Héctor Allende
<hallende@inf.utfsm.cl>

Rodrigo Naranjo
<rodrigo.naranjo@alumnos.usm.cl>

7 de mayo de 2019

Frecuentistas vs Bayesianos

1. Estadística Frecuentista

Contexto

La interpretación clásica, mayoritaria por lo menos hasta ahora, define la probabilidad en términos de experimentación. Si repites un experimento un número infinito de veces y compruebas que en 350 de cada 1.000 ocasiones se ha producido un determinado resultado, un frecuentista diría que la probabilidad de ese resultado es del 35 %. Basándose en esta definición, un frecuentista afirma que es posible asociar a cada evento una probabilidad de obtener un valor VERDADERO del mismo.

La aproximación clásica se basa por lo tanto en estudiar la probabilidad real"de las cosas, tratando de determinar hasta qué punto una medición realizada sobre un conjunto de experimentos se aproxima a la probabilidad real que subyace.

Ley de los grandes números

La ley de los grandes números es un teorema fundamental de la teoría de la probabilidad que indica que si repetimos muchas veces (tendiendo al infinito) un mismo experimento, la frecuencia de que suceda un cierto evento tiende a ser una constante.

Enunciado de la Ley de los grandes números (ley fuerte)

Sean $X_1, X_2, ..., X_n$ una muestra aleatoria iie, X_i una variable aleatoria con esperanza $E[X] = \mu$. Entonces para una muestra de tamaño infinito, el promedio de la muestra converge en probabilidad al valor esperado.

$$P(Lim_{n\to\infty}\overline{X}_n = \mu) = 1$$

Así, la ley de los grandes números señala que si se lleva a cabo repetidas veces un mismo experimento (por ejemplo lanzar una moneda, tirar una ruleta, etc.), la frecuencia con la que se repetirá un determinado suceso (que salga cara o sello, que salga el número 3 negro, etc.) se acercará a una constante. Dicha constante será a su vez la probabilidad de que ocurra este evento.

Actividad 1 - Demostración experimental (50 pts.)

En la primera parte de este laboratorio deberán corroborar y utilizar este teorema aprovechando la capacidad de realizar simulaciones utilizando software estadístico **R** o **Python**.

1.- Demuestre el teorema utilizando un caso particular, el lanzamiento de dos dados. Determine si la ley se cumple si se quiere calcular la probabilidad de obtener un par de 6 al lanzar dos dados. Para esto usted debe:

- Calcular la probabilidad teórica del experimento. **(5 pts.)**
- Realice 8 de simulaciones S_n donde se lanzan dos dados i_n veces, de cada simulación calcule la probabilidad experimental de obtener un par de 6. Las i_n veces que se lancen dados viene dado por la expresión $i_n = 10^{(n)}$. Luego calcule el error $|\epsilon_n|$, la diferencia en absoluto entre el valor real y el esperado para cada simulación. Finalmente reporte sus resultados mostrando a que valores tienden en dos gráficos:
  - Cantidad de lanzamientos vs Probabilidad experimental. **(12 pts.)**
  - Cantidad de lanzamientos vs Error. **(12 pts.)**

2.- ¿Qué importancia posee el supuesto de la infinita cantidad de muestras? ¿Qué sucede si en las simulaciones no apareció ningún par de 6? ¿Se cumple este supuesto con los datos disponibles en la realidad? **(8 pts.)**

3.- Estime la probabilidad teórica utilizando el método de simulaciones con un n lo suficientemente grande para los siguientes experimentos:

- La probabilidad de que al lanzar 5 dados, salgan dos pares, o un par y un trío, o una escala. **(6 pts.)**
- La probabilidad de obtener un número primo al lanzar 10 dados y sumar sus resultados. **(6 pts.)**
- La probabilidad de obtener un número cuadrado perfecto al lanzar 10 dados y sumar sus resultados. **6 pts.**

2. Estadística Bayesiana

Contexto

Por el contrario, la interpretación bayesiana se basa en un conocimiento limitado de las cosas. Afirma que sólo asocias una probabilidad a un evento porque hay incertidumbre sobre el mismo, es decir, porque no conoces todos los hechos. En realidad, un evento dado, o bien ocurrirá (probabilidad=100 %) o bien no ocurrirá (probabilidad=0 %). Cualquier otra cosa es una aproximación que hacemos del problema a partir de nuestro conocimiento incompleto del mismo. El enfoque bayesiano se basa por lo tanto en la idea de refinar predicciones a partir de nuevas evidencias. Un bayesiano definiría probabilidad como la expresión matemática que mide el nivel de conocimiento que tenemos para hacer una predicción.

El teorema de Bayes

El teorema de Bayes, es una proposición planteada por el matemático inglés Thomas Bayes que expresa la probabilidad condicional de un evento aleatorio A dado B:

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

Sin embargo, existe otra interpretación conocida como la **interpretación diacrónica**. Con esta interpretación obtenemos una expresión que nos indica como actualizar nuestra creencia de una hipótesis inicial **H** la evidencia entregada por los datos **D**.

$$P(H|D) = P(H)\frac{P(D|H)}{P(D)}$$

En esta interpretación los términos son conocidos como:

- $P(H)$: Priori, la probabilidad de la hipótesis antes de ver los datos.
- $P(H|D)$: Posteriori, la probabilidad de la hipótesis antes de ver los datos, lo que queremos computar.
- $P(D|H)$: Likehood, la probabilidad de la evidencia bajo una hipótesis.
- $P(D)$: Constante de normalización, la probabilidad de la evidencia bajo cualquier hipótesis.

Naive Bayes

Naive Bayes es un algoritmo de *aprendizaje supervisado* que aplica el teorema de Bayes con el supuesto *ingenuo* (naive) de la independencia condicional entre los pares de características dado la clase a que pertenece. El teorema de Bayes nos entrega la siguiente relación para la probabilidad de la clase y dado las características x_i.

$$P(y|x_0, ..., x_i, ..., x_n) = P(y)\frac{P(x_0, ..., x_i, ..., x_n|y)}{P(x_0, ..., x_i, ..., x_n)}$$

Dado el supuesto de independencia esto se simplifica a:

$$P(y|x_0, ..., x_i, ..., x_n) = P(y)\frac{\prod_{i=0}^{n} P(x_i|y)}{P(x_0, ..., x_i, ..., x_n)}$$

Actividad 2 - Heart Disease UCI (50 pts.)

Para la segunda parte de este laboratorio se les pide utilizar *naive bayes* para determinar la probabilidad de que una persona p, dado ciertas características x_i^p, sea propensa a sufrir un ataque cardiaco. Se utilizara una versión reducida del dataset Heart Disease UCI, este dataset consiste en diferentes casos de personas provenientes de diferentes hospitales, usted debe determinar las probabilidades de sufrir un ataque de un caso particular.

Los cálculos deben ser realizados utilizando software estadístico **R** o **Python**.

1.- Derive el teorema de Bayes a partir de la probabilidad conjunta y demuestre la simplificación del teorema dado la independencia de las x_i. **(8 pts.)**

2.- Se desea determinar cuál es la probabilidad de que un paciente p sufra un ataque. p es una mujer de 49 años, con un dolor de pecho (cp) del tipo *atypical angina* y un nivel de presión sanguínea menor a 120. Para esto usted debe:

- Defina su priori o creencia inicial, explique en que se basó para utilizar esta probabilidad. **(3 pts.)**
- Escriba la expresión simplificada de *naive bayes* para este caso particular (*hint*: ¿Para que valores de x_i calculamos su probabilidad?). **(2 pts.)**
- Utilizando **solo las personas del hospital 1**, calcule la probabilidad de cada característica x_i sea igual a la del paciente. (*hint*: $P(x_i = x_i^p)$ según los datos) **(5 pts.)**
- Utilizando **solo las personas del hospital 1**, calcule la probabilidad condicional de cada característica x_i sea igual a la del paciente dado que sufrió un ataque. (*hint*: $P(x_i = x_i^p|target = 1)$ según los datos) **(5 pts.)**
- Utilice sus resultados para evaluar la expresión de *naive bayes* y obtener su posteriori, ¿Cuál es la probabilidad de sufrir un ataque según la primera evidencia (hospital 1)? **(5 pts.)**

3.- Utilizando la probabilidad obtenida como nueva priori, iterativamente actualice su priori calculando su posteriori usando las personas de cada hospital. Muestre con un gráfico como cambia la probabilidad cada vez que utiliza nuevos datos. ¿Cuál es su probabilidad final? **(12 pts.)**

4.- Utilizando su priori inicial, calcule nuevamente su posteriori utilizando todo el dataset (todos los hospitales juntos). ¿Es diferente la probabilidad obtenida? ¿A que se debe su resultado? **(10 pts.)**

3. Sobre el desarrollo

Las sesiones y material usados serán hechas en R y Python. El desarrollo puede ser realizado con R o Python utilizando las herramientas presentadas en las sesiones. Las herramientas para el desarrollo son R

Markdown y Jupyter Notebooks, respectivamente. Para usar R se recomienda trabajar en RStudio, y para Python usar Jupyter Notebooks junto con Spyder, recomendado trabajar con Anaconda.

4. Sobre la Entrega

El informe puede realizarse en parejas o tríos. El informe **debe incluir el código** que usó en la ejecución, por lo que es necesario que use notebooks en el trabajo. Se aplicarán **descuentos** por código desordenado, ilegible o no modularizado. Se recomienda leer las siguientes convenciones de código: https://github.com/google/styleguide. La fecha de entrega es **el viernes 17 de Mayo**. El archivo a subir **debe ser el notebook** con el que trabajaron con los scripts ejecutados en formato HTML (o .ipynb en caso de usar Jupyter Notebooks) con nombre "Nombre1Apellido1-Nombre2Apellido2" a la sección de entregas de Moodle. En caso de atrasos, si el atraso es de 1 día, la nota máxima será 80. 2 o más días tendrán nota 0.