Laboratorio 1
Estadística Computacional

Universidad Técnica Federico Santa María
Departamento de Informática

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

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

Héctor Allende
<hallende@inf.utfsm.cl>

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

29 de julio de 2019

Distribuciones de probabilidad

El objetivo de esta experiencia será comprobar distintas afirmaciones teóricas sobre distribuciones probabilísticas mediantes resultados empíricos, y aplicar estas herramientas en un problema de carácter práctico. Para esta experiencia, trabajaremos con los paquetes de distribuciones incluidos en R, en el caso de Python, estos se encuentran repartidos en distintas librerías, mayormente *statmodels* y *numpy*. Para cada pregunta en donde se pida un gráfico, este debe ser claro y entendible (título, unidades, legible, etc). Recuerde también escribir las interpretaciones de cada gráfico. Puede usar cualquier paquete disponible para graficar sus resultados.

1. Contexto

La ventaja de trabajar las distribuciones a través de experimentos computacionales, además de poder trabajar una gran cantidad de datos, está en poder obtener muestras de cada distribución, permitiendo generar datos sintéticos de origen pseudo-aleatorio de acuerdo a la distribución respectiva. Es en base a esta capacidad de generar muestras que intentaremos corroborar ciertas propuestas teóricas y, además, usar las distribuciones en casos prácticos.

2. Distribuciones Discretas (30)

La distribución binomial es interesante para modelar, pero presenta ciertos problemas cuando existen muchos elementos involucrados en el problema. Es por esto que existen aproximaciones para esta distribución, las cuales son descritas de la siguiente forma:

Número de ensayos n	Probabilidad de éxito p	Sea $B(n,p)$
$n \geq 30$	$p < 0{,}1$	aproximar a la Poisson con $\lambda = np$
$n \geq 30$	$0{,}1 \leq p \leq 0{,}9$	aproximar a la Normal $N(np, np(1-p))$
$n \geq 30$	$0{,}9 < p$	aproximar recíproco a la Poisson con $\lambda = n(1-p)$
$n < 30$	cualquier p	no aproximar, calcular con la variable original

Responda las siguientes preguntas

1.- Explique porqué es posible aproximar una distribución discreta con una continua.

2.- Demuestre como estas aproximaciones se cumplen, generando muestras de la distribución binomial con parámetros fijos, y comparando los histogramas de las muestras con las respectivas funciones teóricas.

- Fije los parámetros n y p.
- Obtenga como mínimo 3 muestras de tamaños distintos, incrementando en orden de magnitud.
- Grafique el histograma de frecuencias de la muestra.
- Añada al gráfico la curva teórica de la distribución usada para aproximar.
- Incluya gráficos obtenidos como respuesta.

3.- ¿Tiene relación el parámetro n con el error de aproximación? Muestre con un gráfico como varía el error de aproximación, para p fijo, incrementando el valor de n.

3. Distribuciones Continuas (30)

En su trabajo como ingeniero de datos le solicitan obtener datos sintéticos para complementar el entrenamiento de un algoritmo de aprendizaje de maquinas. Las restricciones que deben poseer los datos son:

- Encontrarse en un dominio continuo, en el intervalo $[0, \infty[$.
- Ajustarse a una media y varianza específica.

1.- Explique porqué la distribución Gamma serviría en este caso.

2.- Genere datos provenientes de la distribución, para una media igual a 12, y varianza igual a 36. Aumente iterativamente la cantidad de datos generados hasta que el error de la media real vs muestral sea menor que 10^{-3}. Grafique como varía el error vs el tamaño de la muestra generada.

3.- Diseñe y ejecute un procedimiento iterativo, el cual le permita encontrar el valor del parámetro *shape* bajo el cual la probabilidad de obtener un dato menor a 12 sea 0.4, con un error menor a 10^{-3}. Mantenga el otro parámetro de la pregunta anterior. Grafique como varia el error a medida que se llega al valor del parámetro.

4. Distribuciones Multivariadas (30)

Dentro de los algoritmos de clasificación en el área de aprendizaje de máquinas, uno de los más simples es el método **Análisis discriminante lineal**, el cual, en simples palabras, construye una separación matemática que permite diferenciar un grupo de datos de otro. Probaremos el funcionamiento de esta técnica aprovechando el uso de la distribución normal multivariada.

1.- Obtenga 3 muestras normales bivariadas de tamaño 10000, con los siguientes parámetros:

$$X_1 : \mu_1 = (12, 10) \rightarrow \Sigma_1 = \begin{bmatrix} 3 & 0{,}9 \\ 0{,}9 & 5 \end{bmatrix}$$

$$X_2 : \mu_2 = (8, 6) \rightarrow \Sigma_2 = \begin{bmatrix} 5 & -0{,}7 \\ -0{,}7 & 6 \end{bmatrix}$$

$$X_3 : \mu_3 = (15, 4) \rightarrow \Sigma_3 = \begin{bmatrix} 10 & 0{,}2 \\ 0{,}2 & 7 \end{bmatrix}$$

y grafíquelas.

2.- Agregue al gráfico las siguientes zonas:

- Para cada muestra X_i, una elipse con centro μ_i y radios $\sigma_{i,1}$ y $\sigma_{i,2}$, y ángulo proporcional a la covarianza. *Hint:* $-1 \leq cov \leq 1 \rightarrow -45° \leq \theta \leq 45°$
- Para cada muestra X_i, una elipse con centro μ_i y radios $2\sigma_{i,1}$ y $2\sigma_{i,2}$ y ángulo proporcional a la covarianza.

Usando los valores obtenidos en la pregunta anterior, calcule el porcentaje de datos de cada muestra que se encuentre al interior de las elipses. ¿Cómo puede interpretar esta proporción de área?

3.- Suponga que LDA nos entrega como fronteras la siguientes líneas:

- Entre X_1 y X_2:
$$x_2 = \frac{-3}{2}x_1 + 23$$
- Entre X_2 y X_3:
$$x_2 = \frac{4}{3}x_1 - \frac{28}{3}$$
- Entre X_1 y X_3:
$$x_2 = \frac{1}{5}x_1 + 5$$

¿Qué tan buena es la clasificación de las muestras con estas fronteras? Obtenga la matriz de confusión para cada frontera con los datos muestreados.

4.- ¿Cuál es la probabilidad de que un dato de una muestra sea clasificado en otra muestra? Obtenga este valor de forma empírica. *Hint: Esto sería, dato de X_1 clasificado como X_2 o X_3, junto con X_2 clasificado como X_1 o X_3, y X_3 clasificado como X_1 o X_2.*

5. Conclusiones (10)

Mencione las conclusiones más relevantes e interesantes que ha encontrado en el análisis. La conclusión también lleva puntaje, tome su tiempo para encontrar información útil.

6. 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.

7. 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 **TBA**. 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.