Aprendizaje automático

Departamento de Ingeniería en Informática
ITBA

Trabajo Práctico 3

Objetivo: Aplicar diversos métodos estadísticos para aprender a hacer inferencia a partir de datos experimentales

Métodos de estadística paramétrica

1. Para cada especie de los “datos de los lirios Fisher” (ver ej. 5 tp2), suponiendo que las medidas tomadas sobre los lirios siguen una distribución gaussiana, hallar:
   (a) Los estimadores de máxima verosimilitud de la media y la varianza de los largos y anchos de los sépalos y de los pétalos.
   (b) Calcular los errores cuadráticos medios de los estimadores las medias de las 4 medidas tomadas sobre los lirios.
   (c) Hallar los intervalos de confianza de nivel 0.95 para las medias de las 4 medidas tomadas sobre los lirios.

2. En una muestra aleatoria de 80 componentes de cierto tipo se encontraron 12 defectuosos.
   (a) Obtener una estimación de la proporción de todos los componentes que no están defectuosos.
      i. ¿Es insesgado este estimador?
      ii. Calcular el error cuadrático medio.
   (b) Se va a construir un sistema seleccionando al azar dos de estos componentes y conectándolos en serie como se muestra en la figura:

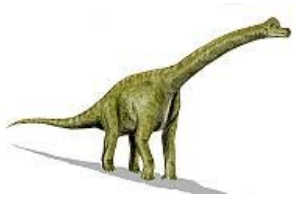
   La conexión en serie implica que el sistema funcionará sólo si ninguno de los dos componentes está defectuoso.
   Estimar la proporción de sistemas que funcionan correctamente.

3. Wiebe y Bortolotti (2002) examinaron los colores de las plumas de la cola de los carpinteros escapularios. Algunas de estas aves tienen unas plumas “raras” que tienen un color o un largo distinto del resto de las plumas de la cola, posiblemente porque han vuelto a crecer después de haberlas perdido.
   Ellos midieron la amarillez de estas plumas “raras” de 16 carpinteros y la compararon con la amarillez de una pluma “típica” de la misma ave.

	Grado de amarillez	
ave	**pluma típica**	**pluma rara**
1	-0.255	-0.324
2	-0.213	-0.185
3	-0.190	-0.299
4	-0.185	-0.144
5	-0.045	-0.027
6	-0.025	-0.039
7	-0.015	-0.264
8	0.003	-0.077
9	0.015	-0.017
10	0.020	-0.169
11	0.023	-0.096
12	0.040	-0.330
13	0.040	-0.346
14	0.050	-0.191
15	0.055	-0.128
16	0.058	-0.182

Los investigadores postulan que hay variación entre el color de las plumas típicas y las plumas raras. ¿Están en lo cierto?
Suponiendo la normalidad de los datos, testear la hipótesis de los investigadores.

4. Los datos en http://www.stat.ncsu.edu/working_groups/sas/sicl/data/brain.dat corresponden a los promedios de los pesos del cuerpo (en kg.) y de sus cerebros (en g.) para 28 especies distintas. Estos datos forman parte del conjunto de mediciones hechas por Allison y Cicchetti (1976).
   (a) Graficar 'peso del cuerpo' vs. 'peso de cerebro'.
      i. El dato 14 corresponde a los humanos, el 15 a los elefantes africanos, el 25 a los braquiosauros. Identificar estos datos.
      ii. ¿Existe una relación lineal entre las variables?



   (b) Graficar log(peso del cuerpo) vs. log(peso del cerebro) ¿Existe una relación lineal entre las variables?
   (c) Aplicar una transformación apropiada a las variables para obtener linealidad de los datos y ajustar la recta de regresión.
   (d) Obtener nuevamente la recta de regresión sin considerar los datos 14, 15 y 25. ¿Mejoró el ajuste?

5. En http://www.census.gov/population/www/censusdata/files/table-2.pdf se dan los datos correspondientes a la cantidad de habitantes obtenidos en los censos de población realizados en cada década en EE.UU. entre los años 1790 y 1990.
   (a) Graficar la cantidad de habitantes (en millones) correspondiente a cada década. ¿Existe una relación lineal entre las variables?
   (b) Ajustar un polinomio de grado 2 y estudiar los residuos.
   (c) Ajustar un polinomio de grado 3 y estudiar los residuos.
   (d) En base a los resultados de los ajustes anteriores, ¿qué modelo sería adecuado para ajustar los datos?
   (e) Utilizando los ajustes anteriores, predecir la población de EE.UU. en el año 2000.

Si en el censo realizado en dicho año se obtuvo que la población era de 281421906 habitantes, ¿qué modelo dio la mejor predicción?

6. Para los “datos de los lirios Fisher”, suponiendo que las medidas tomadas sobre los lirios siguen una distribución gaussiana, hallar los estimadores de máxima verosimilitud del vector de medias y de la matriz de covarianza de los largos y anchos de los sépalos y de los pétalos.

7. Hald (1952) estudió la evolución del calor en calorías por gramo de cemento (y) como función de la cantidad de los cuatros ingredientes en la mezcla: aluminato tricálcico (x_1), silicato tricálcico (x_2), ferrita de aluminio tricálcico (x_3) y silicato dicálcico (x_4). Los datos están disponibles en

   http://www.stat.ncsu.edu/working_groups/sas/sicl/data/setting.dat

   (a) Hallar la matriz de correlación muestral. ¿Qué deduce de los valores de esta matriz?
   (b) Aplicar el método de selección hacia adelante para encontrar un modelo adecuado para ajustar los datos.

8. Para los datos del cemento de Hald (ver ej. 7):
   (a) Hallar las componentes principales de los ingredientes y la varianza explicada por cada componente.
   (b) Hacer el biplot de las dos primeras componentes.