Banjercito

Diplomado en Ciencia de Datos

Notas del Curso

Módulo III — Semana 4

Instructor: Alan Badillo Salas

Septiembre 2024

1 Introducción

En la semana 3 estudiamos el análisis de la varianza para un eje de datos segmentado por grupos. El principal objetivo fue crear una segmentación en el eje de datos que permitiera separar los datos en grupos. Si cada grupo se comportaba de manera similar a otro grupo, entonces la segmentación no reflejaría un cambio significativo entre los datos, y por lo mismo, la varianza de los grupos sería similar, implicando que los grupos no se contradicen respecto al eje.

De manera contraria, si un grupo o segmento tiene una media y varianza distinta a otro grupo, entonces, el grupo contradice el comportamiento del otro y esto significa que los grupos reflejan características diferentes, por lo que se aprueba una hipótesis de que los grupos de un mismo eje tienen diferente comportamiento bajo esa segmentación.

Este resultado se puede medir mediante la Prueba F de Fisher que mide la probabilidad de que dos segmentos de datos compartan una misma distribución en las colas, es decir, que compartan muestras que vuelvan ambiguo el determinar a qué segmento debería pertenecer el dato.

Esta semana estudiaremos la regresión simple y la regresión múltiple, las cuales nos permitirán encontrar la recta o el plano respectivamente que se ajusta mejor a nuestros datos, encontrando el valor de tendencia continuo sobre todo el eje.

1.1 Contenido de la Semana 4

Esta semana revisaremos los siguientes temas:

1. Predicción de datos
   (a) Ejes independientes
   (b) Eje dependiente
2. Regresión lineal simple
   (a) Correlación entre dos ejes
   (b) Línea de tendencia
   (c) Regresión lineal $x \to y$
   (d) Coeficientes $y = m \cdot x + b$
   (e) Caso de estudio sobre la planta iris
   (f) Caso de estudio sobre la supervivencia del titanic
3. Regresión lineal múltiple
   (a) Correlación entre múltiples ejes
   (b) Plano de tendencia
   (c) Regresión lineal $x_1, x_2, \ldots, x_n \to y$
   (d) Coeficientes $y = m_1 \cdot x_1 + m_2 \cdot x_2 + \ldots + m_n \cdot x_n + b$
   (e) Caso de estudio sobre la planta iris
   (f) Caso de estudio sobre la supervivencia del titanic

2 Análisis por Grupos

Un eje de datos puede ser la edad, la estatura, el peso o el salario de una persona, pero también la altura de un tornillo en una línea de producción, la concentración de gas en un refresco o el monto de crédito aprobado en un préstamo. Un eje por sí mismo no aporta mucha información al análisis global de un estudio, para ello, podemos basarnos en otros ejes para segmentar los datos y encontrar cambios significativos entre los datos.

Por ejemplo, vamos a pensar que analizamos el salario de las personas, entonces encontramos que su salario promedio es de $10,500.00 pesos, esto nos sirve para saber en promedio cuánto gana una persona. Pero ¿Qué pasa si segmentamos los datos en hombres y mujeres? Es decir, ¿Qué pasa si tomamos las muestras que corresponden a aquellas personas que son hombres y por otro lado separamos las muestras de personas que son mujeres?