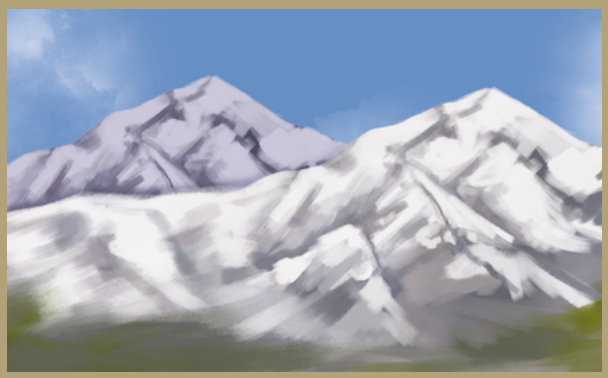
BOURBAKI
COLEGIO DE MATEMÁTICAS

Matemáticas y Datos... IV
La divergencia de Kullback–Leibler

En diversas etapas durante el pipeline de ciencia de datos podría ser importante calcular la divergencia que existe entre dos datasets pues esto nos dirá algo sobre la divergencia entre las dos poblaciones que conforman estos conjuntos de datos.

Es importante mencionar que para que este cálculo tenga sentido, forzosamente ambos datasets deben compartir las mismas características lo cual se traduce en que sus vectores aleatorios compartan un espacio de probabilidad.



Si dos vectores aleatorios X, X' comparten espacio de probabilidad discreto, definimos la divergencia de Kullback–Leibler de X' respecto a X mediante la siguiente fórmula:

$$KL(X', X) = x_1[log(x_1) - log(x'_1)] + \ldots + x_n[log(x_n) - log(x'_n)]$$

Mediante esta fórmula estamos suponiendo que la distribución X va a ponderar las diferencias que existen entre las poblaciones X y X', es decir las discrepancias respectos a los posibles miembros. Es importante notar que esta fórmula no es simétrica pues si X' pondera las diferencias obtenemos una nueva cantidad.

info@colegio-bourbaki.com
+5217351636845
Colegio de Matemáticas Bourbaki