Benemérita Universidad Autónoma de Puebla
Facultad de Ciencias de la Computación

Máquinas de aprendizaje

Reporte: Aplicaciones con bootstrap

Docente: Abraham Sánchez López

Alumno	Matrícula
Taisen Romero Bañuelos	202055209

Cambios generados por el uso de bootstrap

En las prácticas anteriores usamos modelos que daban por hecho algunas cosas, como la normalidad o independencia de variables. En esta práctica, como lo hemos estado haciendo últimamente, ponemos en tela de juicio algunas cosas que antes no se consideraban. Un ejemplo de eso es la estabilidad de los clusters generados con k-means. Antes hacíamos los clusters y los interpretamos sin validar su calidad (estabilidad en este caso).

Según entendí, descubrimos que el cluster 3 no es tan estable porque la alimentación de las ciudades que lo conforman no eran tan similares entre sí, aunque también podría verse como que se parecen un poco más a las ciudades de otros clústers, ya que la estabilidad checa qué tan frecuente es el cambio de un elemento en "x" cluster durante las iteraciones que programemos. Pienso yo que esto podría ser una señal para aumentar un poco el número de clústers que queremos generar. Digamos que bootstrap nos muestra los resultados de variar ligeramente los datos múltiples veces y eso nos ayuda a saber si hay que aumentar el número de clusters, aunque me pregunto cómo se verá cuando haya que disminuirlo.

Sobre los modelos mixtos, me confundí un poco porque luego de leer el PDF de clustering estaba mezclando las cosas en mi cabeza, también porque entre uno que otro término técnico me confundía. Pero se puede decir que usamos bootstrap para calcular los intervalos de confianza de ciertas métricas. Hicimos algunos remuestreos para ver qué tanto varían las métricas y según si varían mucho o poco vemos si tiene un buen intervalo de confianza o no. Y hablando de esos intervalos de confianza, yo pensaría que basarse en los datos reales generaría un intervalo de confianza (IC) más estrecho, pero al parecer no fue así. Es decir, aunque el bootstrap paramétrico y no paramétrico dieron resultados más o menos similares, el paramétrico (el que usa regresión lineal para simular datos nuevos) tiende a generar IC más estrechos, pero de todas formas, yo no me confiaría tanto como para sólo usar el modelo paramétrico porque si uno configura un poco mal el modelo bien puede dar resultados más deseables. Creo que sería interesante hacer una práctica más profunda dedicada a la comparación de ambos métodos, ya que parece que dan para hablar un poco más porque está el dilema de fiarnos de un método que sigue la tendencia natural de los datos (porque se basa en los datos que se tienen) versus un método que hace una simulación de datos similares que en principio parece funcionar.