La maldición de la dimensionalidad

La maldición de la dimensionalidad (*curse of dimensionality*) se refiere al fenómeno que ocurre cuando la dimensionalidad del espacio de característica es muy grande para un conjunto de datos. En principio, el número de instancias necesarias para estimar un modelo con un determinado nivel de precisión crece exponencialmente con respecto al número de variables de entrada. ¿Qué pasa si se incrementa la dimensionalidad y no así la cantidad de datos? A medida que aumenta el número de atributos, la distribución de los datos será cada vez más "dispersa" en este espacio. Intuitivamente, se podría pensar que incluso los datos más cercanos estarán demasiado lejos en un espacio de alta dimensión para dar una buena estimación. De hecho, la variabilidad de las distancias entre los datos disminuye exponencialmente: cuando hay un gran número de atributos, los datos están todos "casi" a la misma distancia entre sí (es decir, los valores de la distancias entre puntos son muy similares).

Muchos algoritmos de aprendizaje funcionarán muy bien en espacios de baja dimensionalidad, pero tendrán dificultades cuando esta aumenta. ¿Qué podemos hacer? En teoría, una solución a la maldición de la dimensionalidad podría ser aumentar el tamaño del conjunto de datos con el fin de alcanzar una densidad suficiente de instancias para el aprendizaje. Sin embargo, en la práctica, el número de instancias de entrenamiento necesarias para alcanzar una densidad dada crece exponencialmente con el número de dimensiones.

El otro camino es **reducir la dimensionalidad de los datos**. La mayoría de los problemas del mundo real conllevan datos que nunca se distribuyen uniformemente en el espacio de entrada (esta observación se conoce como la "bendición" de la no uniformidad). Una distribución de datos no uniforme sugiere que las dimensiones de los datos no son todas independientes, sino que pueden estar altamente correlacionadas o contienen información redundante. Además, en muchos conjuntos de datos la mayor parte de la variabilidad se concentra en unas pocas dimensiones o direcciones. Estas dimensiones capturan la esencia de la información en los datos, mientras que las dimensiones restantes pueden contener ruido o información menos relevante. Sobre estas dos premisas actúan las técnicas de reducción de la dimensionalidad al descartar atributos no

informativos o derivar un nuevo conjunto de características que recoge la información esencial de los datos.

Bibliografía

"The curse of dimensionality". En: Geron, A. (2022). Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow. Capítulo 8. 3rd Edition. O'Reilly.

Universidad de los Andes | Vigilada Mineducación Reconocimiento como
Universidad: Decreto 1297 del 30 de mayo de 1964.
Reconocimiento personería jurídica: Resolución 28 del 23 de febrero de 1949 Minjusticia.