Exploratory Data Analysis

Machine Learning

Profesor: M.C.E. Orlando Uc

Contar una historia...

El trabajo de un científico de datos consiste en la resolución de problemas de negocio a partir de inferencias obtenidas al analizar diferentes bases de datos.

Pareciera que el trabajo de un *data scientist* se reduce a aplicar diferentes técnicas estadísticas, pero la verdad es que el trabajo de un científico de datos consiste en *contar una historia*.
Es decir, en poder traducir las conclusiones obtenidas de los datos en a lo más unas cuentas líneas y un par de gráficas que tengan armonía.

Contar una historia...

Cualquiera puede contar una historia, pero para que una historia sea interesante debe enganchar al espectador, debe tener un orden cronológico, debe estar bien escrita y debe rematar con un final lógico y bien fundamentado.

Todo lo anterior lleva a una *buena historia*.

Contar bien una historia...

Un científico de datos presta sus servicios a *tomadores de decisiones*, personas que tienen la responsabilidad de crear una estrategia de negocio, alcanzar los Indicadores Clave de Rendimiento (KPI, por sus siglas en inglés) y presentar soluciones innovadoras a problemas complejos.

Sin embargo, no todos los tomadores de decisiones tienen el conocimiento técnico suficiente para interpretar de manera adecuada las métricas de los resultados.

Entonces, se puede decir que en realidad **el trabajo de un científico de datos consiste en contar bien una buena historia**.

Exploratory Data Analysis (EDA)

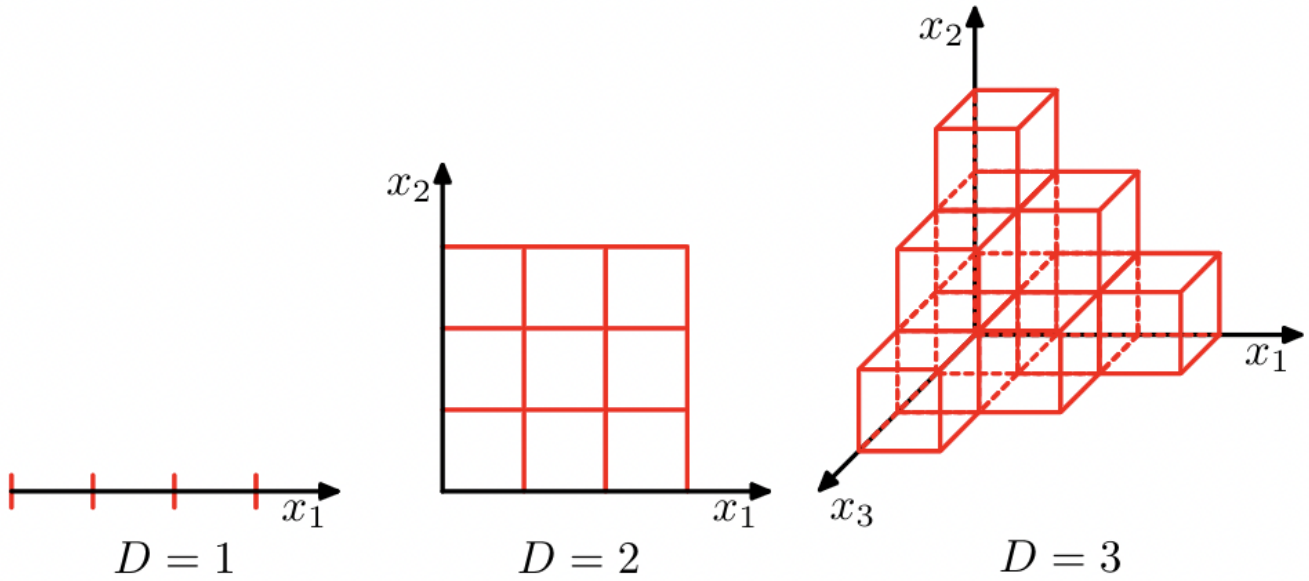
Se conoce como Análisis Exploratorio de Datos al conjunto de técnicas que permiten encontrar inferencias interesantes a través de los datos.

Por lo general, el EDA se acompaña de gráficos y diferentes técnicas de visualización.

La maldición de la dimensionalidad

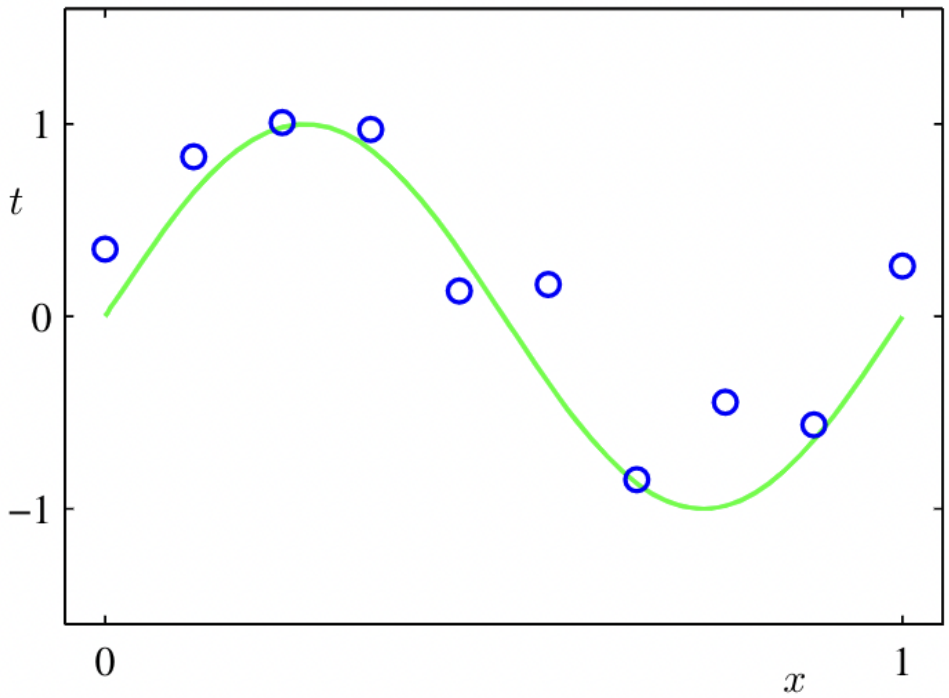
Para contar bien la historia que dicen los datos se deben seleccionar las variables, las estadísticas y los gráficos más adecuados.

Observa en la siguiente figura como, conforme aumenta el número de dimensiones en los que se representan los datos, aumenta también la complejidad del gráfico, y por tanto la dificultad de transmitir el *mensaje correcto* a la audiencia, a esto se le conoce como *la maldición de la dimensionalidad*.



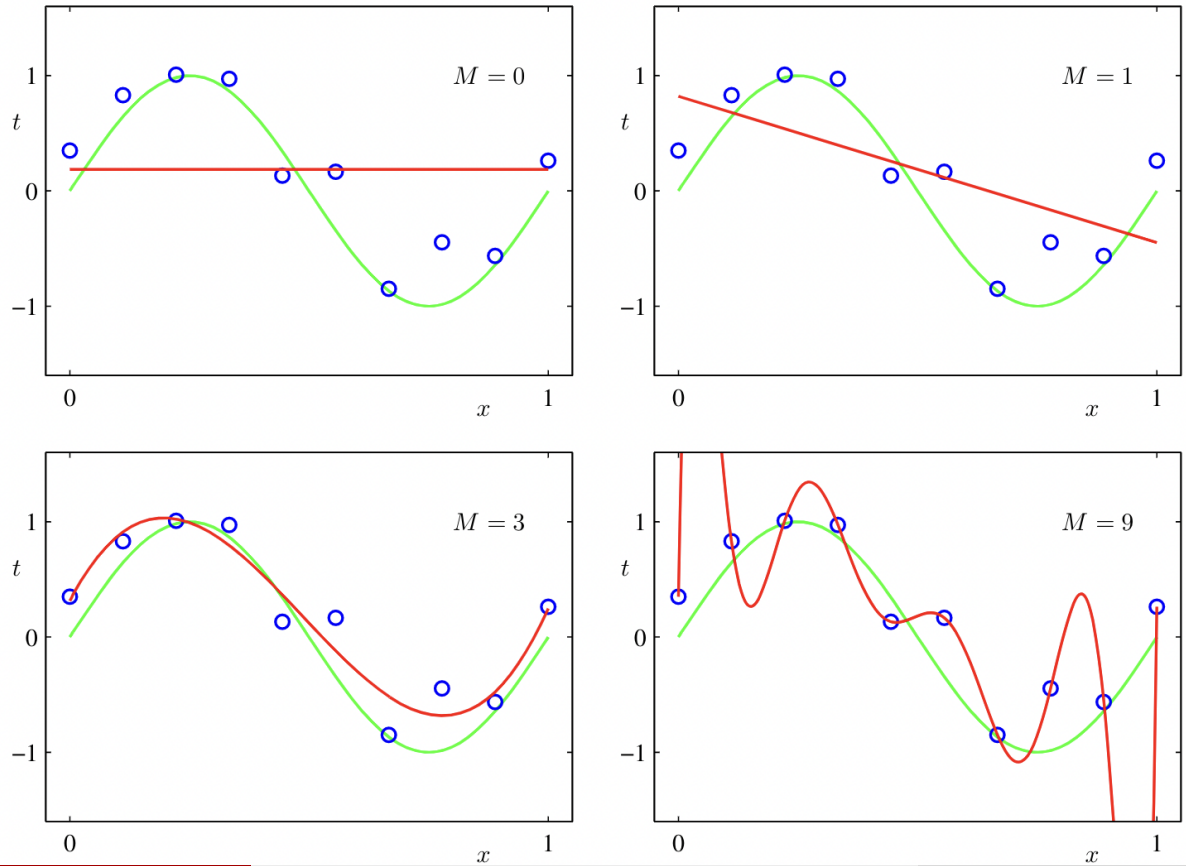
La maldición de la dimensionalidad

En la siguiente figura se pueden observar $N = 10$ puntos en color azul, y en color verde el gráfico de la función $\sin(2\pi x)$, ¿consideras que el gráfico es representativo?

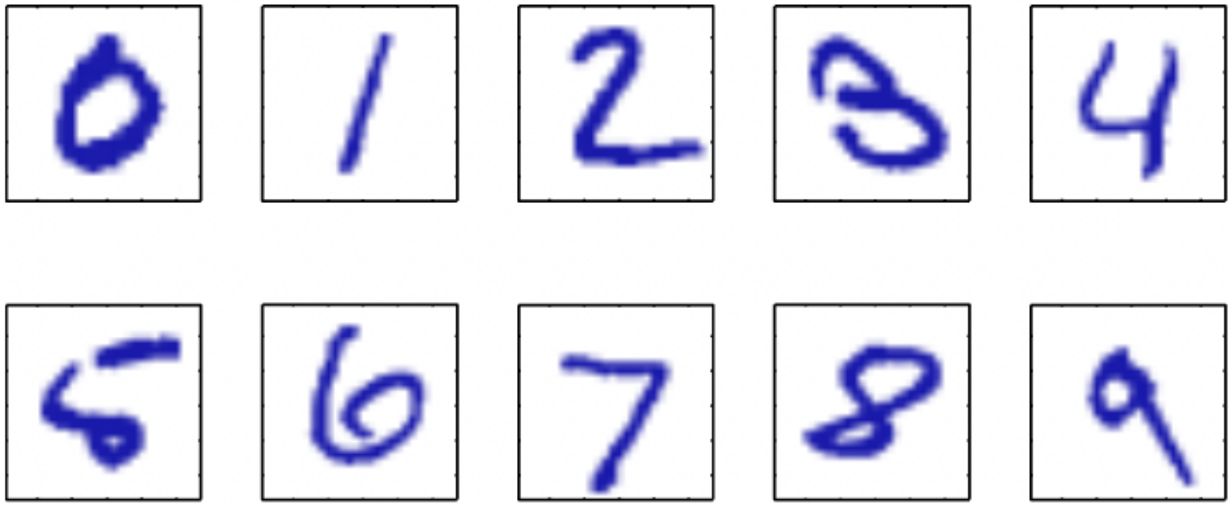


La maldición de la dimensionalidad

En la siguiente figura se encuentran gráficas de polinomios de diferente orden M. ¿Qué puedes comentar el respecto? ¿Consideras que un valor diferente de M cuenta una historia diferente? ¿Cómo elegirías un valor adecuado para M? ¿Existe una única solución?



Ahora observa los dígitos capturados a mano de la siguiente imagen, ¿cuál sería una buena manera de representarlos en bajas dimensiones? ¿Cómo traducirías los dígitos que observas en una base de datos? ¿Qué variables considerarías? ¿Cuántas variables tendrías?



Puedes consultar los libros:

- *Beautiful visualization: Looking at data through the eyes of experts* de Steele, J., & Iliinsky, N. (2010)
- *Pattern recognition and machine learning* de Bishop, C. M., & Nasrabadi, N. M. (2006).

¡Muchas gracias!