El Panorama del Machine Learning Parte 3

Principales Desafíos del Machine Learning

- Dos cosas que pueden salir mal son: "modelo malo" y "datos malos".

Cantidad Insuficiente de Datos de Entrenamiento

- Se necesita una gran cantidad de datos para que la mayoría de los algoritmos de machine learning funcionen correctamente.
- Incluso para problemas muy simples típicamente necesitas miles de ejemplos, y para problemas complejos como reconocimiento de imágenes o voz puedes necesitar millones de ejemplos.
- Pero: La investigación mostró que cuando hay suficientes datos disponibles, incluso algoritmos simples de machine learning pueden funcionar de manera similar en tareas complejas. Esto llevó a la idea de que los datos pueden importar más que el algoritmo en sí. Sin embargo, dado que los grandes conjuntos de datos no siempre están disponibles, el desarrollo de algoritmos sigue siendo importante.

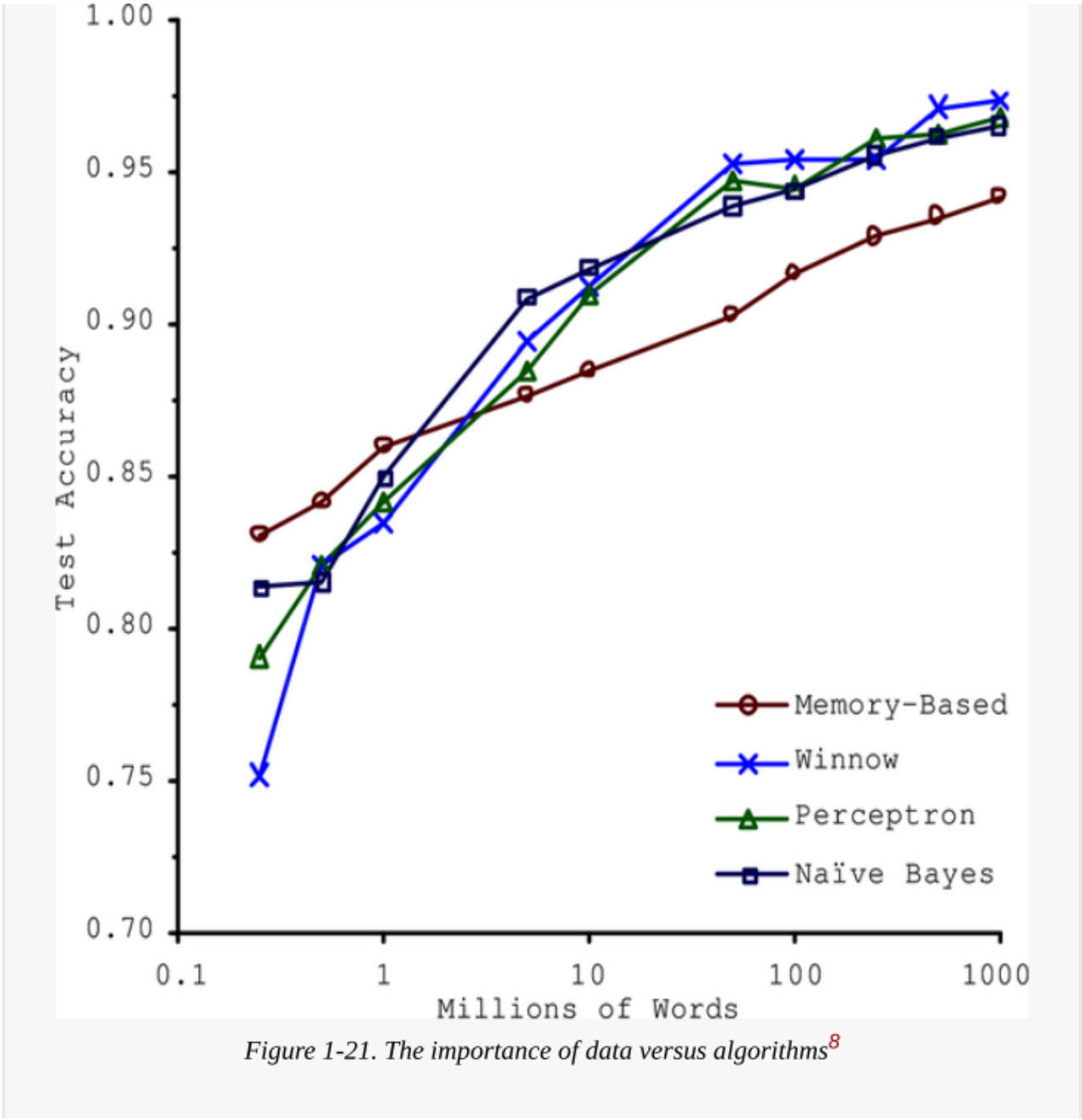

Figure 1-21. The importance of data versus algorithms[8]

Datos de Entrenamiento No Representativos

- Para entrenar un modelo adecuadamente, los datos deben representar los casos del mundo real donde el modelo será usado. Si el conjunto de datos es demasiado pequeño, puede contener errores o estar distorsionado por casualidad. Sin embargo, incluso un conjunto de datos grande puede ser pobre si fue recolectado de manera sesgada o incorrecta. Esto se llama sesgo de muestreo.
- No se trata solo de tener muchos datos — los datos deben reflejar con precisión la realidad.
- Ej:
  - En la elección estadounidense de 1936, el Literary Digest predijo al ganador equivocado porque encuestaron principalmente a personas adineradas (10 millones) y solo un pequeño porcentaje respondió (2.4 millones). Esto causó sesgo de muestreo y sesgo de no respuesta.
  - En el ejemplo de YouTube, recolectar videos de "música funk" a través de resultados de búsqueda solo capturaría contenido popular o sesgado por región, no todos los videos de funk. Esto también crea sesgo de muestreo.

Datos de Mala Calidad

- A menudo vale la pena el esfuerzo dedicar tiempo a limpiar tus datos de entrenamiento.
- Ej:
  - Si algunas instancias son claramente valores atípicos, puede ayudar simplemente descartarlos o intentar corregir los errores manualmente.
  - Si algunas instancias faltan algunas características (ej., 5% de tus clientes no especificaron su edad), debes decidir si quieres ignorar este atributo por completo, ignorar estas instancias, rellenar los valores faltantes (ej., con la edad mediana), o entrenar un modelo con la característica y un modelo sin ella.

Características Irrelevantes

- Tu sistema solo será capaz de aprender si los datos de entrenamiento contienen suficientes características relevantes y no demasiadas irrelevantes.
- Ingeniería de características: Una parte crítica de un proyecto de machine learning que implica crear y seleccionar características efectivas para entrenar el modelo. Involucra los siguientes pasos:
  - Selección de características: seleccionar las características más útiles para entrenar entre las características existentes.
  - Extracción de características: combinar características existentes para producir una más útil—como vimos antes.
  - Crear nuevas características recopilando nuevos datos

Sobreajuste de los Datos de Entrenamiento

- Sobre-generalizar es algo que los humanos hacemos con demasiada frecuencia, y desafortunadamente las máquinas pueden caer en la misma trampa.
- Sobreajuste (Overfitting): significa que el modelo funciona bien en los datos de entrenamiento, pero no generaliza bien.
- Ejemplo de sobreajuste de los datos de entrenamiento

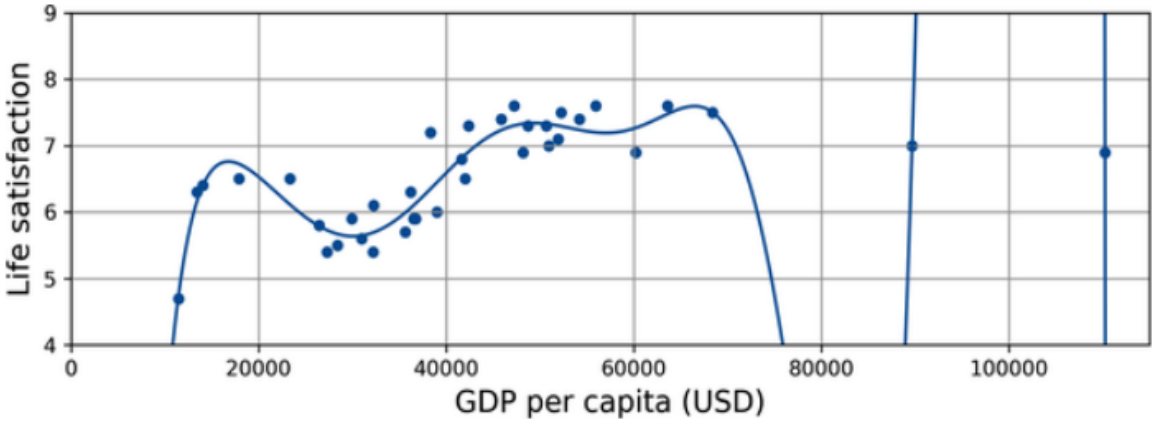

Figure 1-23. Overfitting the training data

  - Modelo de satisfacción con la vida de polinomio de alto grado que sobreajusta fuertemente los datos de entrenamiento.
- Los modelos complejos como las redes neuronales profundas pueden captar patrones muy sutiles, pero si el conjunto de datos es pequeño o ruidoso, pueden aprender patrones sin sentido que ocurrieron por casualidad. Estos patrones no generalizarán a nuevos datos.
- Regularización: Restringir un modelo para hacerlo más simple y reducir el riesgo de sobreajuste.

- Al limitar cuánto pueden cambiar los parámetros del modelo, se reduce la flexibilidad del modelo. El objetivo es encontrar un equilibrio: ni demasiado simple (subajuste) ni demasiado complejo (sobreajuste).

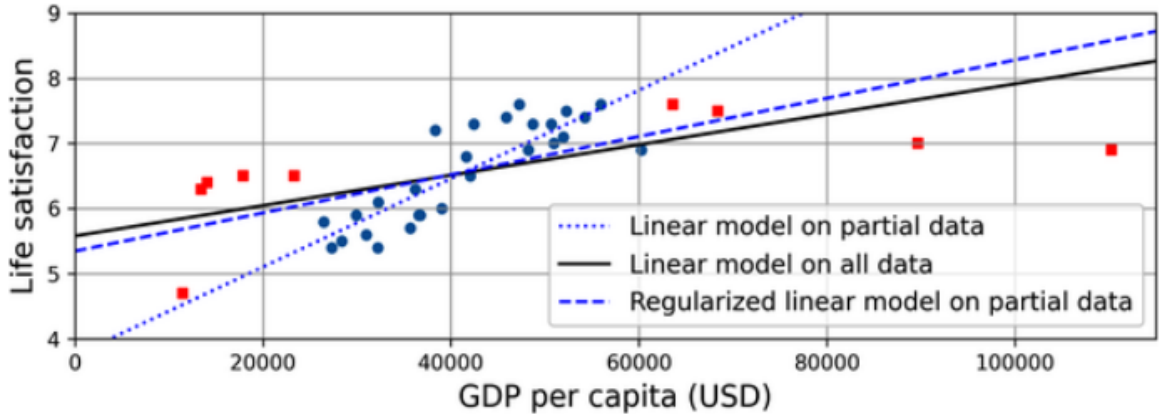

Figure 1-24. Regularization reduces the risk of overfitting

- El modelo regularizado (línea discontinua) se ajusta menos perfectamente a los datos de entrenamiento que el primer modelo, pero generaliza mejor a nuevos puntos no vistos.
- Un hiperparámetro de regularización controla cuánto se simplifica el modelo. Si el valor es demasiado alto, el modelo se vuelve demasiado simple y puede subajustar.

Subajuste de los Datos de Entrenamiento

- El subajuste (Underfitting) es lo opuesto al sobreajuste: ocurre cuando tu modelo es demasiado simple para aprender la estructura subyacente de los datos.
- Opciones para solucionar este problema:
  - Seleccionar un modelo más poderoso.
  - Alimentar mejores características al algoritmo de aprendizaje (ingeniería de características).
  - Reducir las restricciones en el modelo

Conclusión

- El sistema no funcionará bien si tu conjunto de entrenamiento es demasiado pequeño, o si los datos no son representativos, son ruidosos, o están contaminados con características irrelevantes (basura entra, basura sale). Por último, tu modelo no debe ser ni demasiado simple (en cuyo caso subajustará) ni demasiado complejo (en cuyo caso sobreajustará).