

Machine Learning Aplicado

Sesgo y Varianza

Ingeniería Civil Informática
Escuela de Ingeniería Informática
Universidad de Valparaíso

Profesor: Aaron Ponce Sandoval
correo: aaron.ponce@uv.cl

¿Cómo sabemos si un modelo está funcionando bien?

El rendimiento de un modelo de aprendizaje automático se considera bueno según su **predicción** y qué tan bien se **generaliza** en un conjunto de datos de prueba independiente. Basándonos en el rendimiento de los diferentes modelos, elegimos el modelo que tiene el rendimiento más alto.

Ejemplo: Problema

Queremos predecir quién ganará en las elecciones presidenciales en EE.UU

¿Los republicanos o los demócratas?



**Cuatro candidatos
entraron en la contienda**

**Se espera que
Biden se postule**

Republicanos



Trump



Haley



Ramaswamy



Biden

Demócratas



Williamson

Ejemplo: Adquisición de datos

Vamos a un vecindario y comenzamos a preguntarle a la gente si votaría por un demócrata o un republicano. **entrevistamos a 100 personas, 44 dicen que votarán por los demócratas, 40 dicen que votarán por los republicanos y 16 están indecisos.** Con base en estos datos, podemos hacer una predicción de que las posibilidades de que los demócratas ganen son más altas que las de los republicanos.

Ejemplo: Analisis

¿Podemos aplicar esta predicción a todo el condado, estado y luego a nivel nacional?

¿Podemos aplicar esta predicción a todo el condado, estado y luego a nivel nacional?

- ❑ No, porque la predicción podría cambiar si vamos a un vecindario, condado o estado diferente. Observaremos inconsistencias en la predicción.
- ❑ Esto significa que nuestro modelo no funciona bien, ya que no se puede utilizar de forma fiable para hacer predicciones.

El tamaño de la muestra es muy pequeña y no hay suficiente **variación** de datos

Objetivo de un Algoritmo de ML

- ❑ Cuando tenemos una entrada x y aplicamos una función f sobre la entrada x para predecir una salida y . La diferencia entre la salida real y la salida esperada es el error. El objetivo con el algoritmo de aprendizaje automático es generar un modelo que minimice el error del conjunto de datos de prueba.



Bias and Variance for regression

- ❑ Sea $\mathbf{F}(\mathbf{x})$ una función verdadera y desconocida con valor continuo con ruido.
- ❑ Se busca estimar en base a n muestras aleatorias de un conjunto \mathbf{D} generado por $\mathbf{F}(\mathbf{x})$
- ❑ Sea $\mathbf{g}(\mathbf{x}; \mathbf{D})$ una función de regresión estimada

Estamos interesado en la dependencia de la aproximación en el conjunto de entrenamiento \mathbf{D} de tamaño n

$$Err(x) = (g(x; D) - F(x))^2$$

Fórmula:

$$Y = g(x; D) + e$$

$$Err(x) = E[(g(x; D) - F(x))^2]$$

$$Err(x) = (E[g(x; D) - F(x)])^2 + E[(g(x; D) - E[g(x; D)])^2]$$

$$Err(x) = \underbrace{(E[g(x; D) - F(x)])^2}_{Bias^2} + \underbrace{E[(g(x; D) - E[g(x; D)])^2]}_{Varianza} + \underbrace{e}_{Error Irreducible}$$
$$\underbrace{Err(x) = Bias^2 + Varianza + Error Irreducible}_{Error Reductible}$$

El error irreducible es la parte del error de un modelo de aprendizaje automático que no puede ser reducida mediante el ajuste del modelo. Este componente del error se debe a factores que están fuera del alcance del modelo, como el ruido en los datos, la variabilidad natural en los datos o la falta de información relevante.

- ❑ El sesgo se refiere a la capacidad de un modelo de ajustarse a los datos de entrenamiento.
- ❑ Un modelo con alto sesgo no es lo suficientemente complejo para capturar la complejidad de los datos y subestima la relación entre las características de los datos de entrada y la variable de salida.

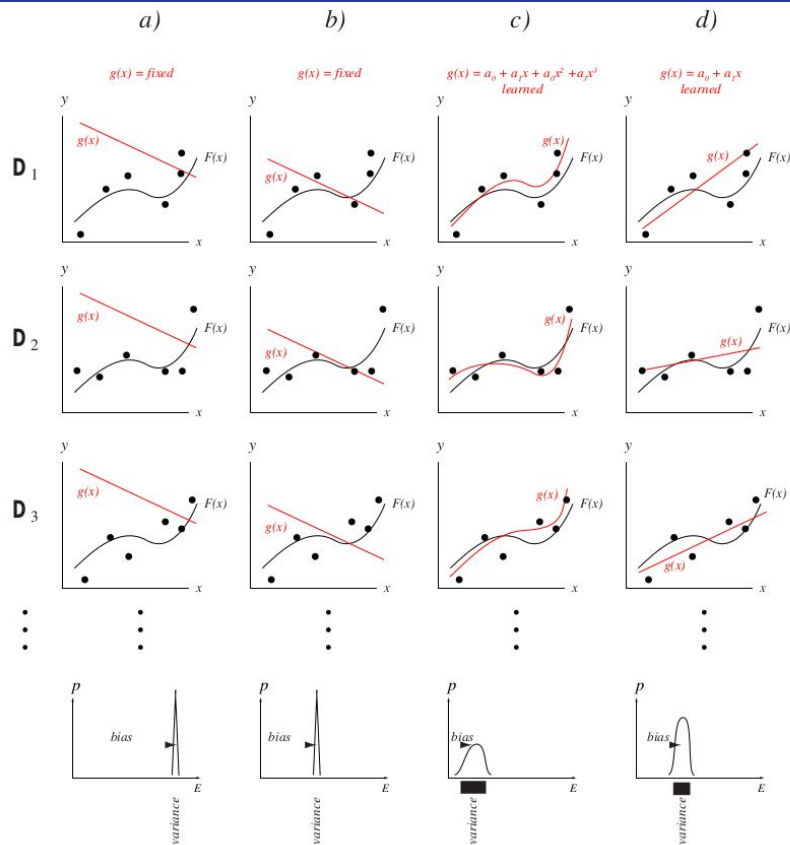
Esto se traduce en una alta tasa de error en los datos de entrenamiento y en una posiblemente alta tasa de error en los nuevos datos.

- ❑ La varianza se refiere a la capacidad de un modelo de generalizar a nuevos datos.
- ❑ Un modelo con alta varianza es demasiado complejo y se ajusta demasiado a los datos de entrenamiento.
- ❑ El modelo con alta varianza presenta una baja capacidad de generalización a nuevos datos.

.

Esto se traduce en una baja tasa de error en los datos de entrenamiento, pero una alta tasa de error en los nuevos datos

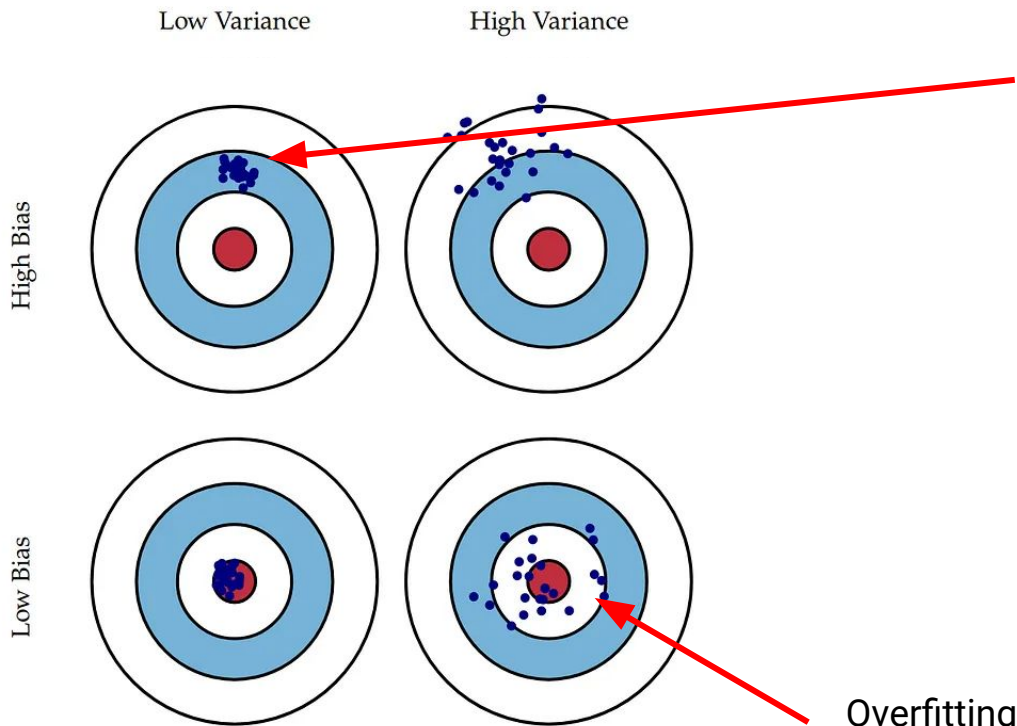
El dilema sesgo-varianza



Cada columna representa un modelo diferente, cada fila un conjunto diferente de $n = 6$ puntos de entrenamiento, D_i , muestreados aleatoriamente de la verdadera función $F(x)$ con ruido

- ❑ La columna a) muestra un modelo muy pobre: un $g(x)$ lineal cuyos parámetros se mantienen fijos, independientemente de los datos de entrenamiento. Este modelo tiene un alto sesgo y cero varianza.
- ❑ La columna b) muestra un modelo algo mejor, aunque también se mantiene fijo, independientemente de los datos de entrenamiento. Tiene un sesgo menor que en a) y la misma varianza cero
- ❑ La columna c) muestra un modelo cúbico, donde los parámetros se entrenan para ajustarse mejor a las muestras de entrenamiento en un sentido de error cuadrático medio. Este modelo tiene un sesgo bajo y una varianza moderada.
- ❑ La columna d) muestra un modelo lineal que se ajusta para adaptarse a cada conjunto de entrenamiento; este modelo tiene un sesgo y una varianza intermedios.

Underfitting and Overfitting



Underfitting

insuficiente ocurre cuando un modelo no puede capturar el patrón subyacente de los datos. Estos modelos suelen tener un alto sesgo y una baja varianza. Ocurre cuando tenemos muy poca cantidad de datos para construir un modelo preciso o cuando intentamos construir un modelo lineal con datos no lineales. Además, este tipo de modelos son muy simples para capturar los patrones complejos en datos como la regresión lineal y logística.

Overfitting

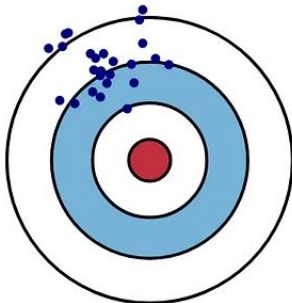
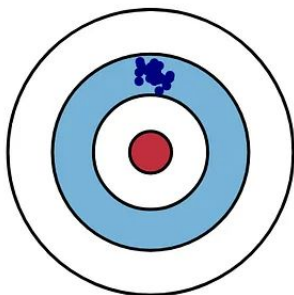
En el aprendizaje supervisado, el sobreajuste ocurre cuando nuestro modelo captura el ruido junto con el patrón subyacente en los datos. Ocurre cuando entrenamos mucho nuestro modelo sobre un conjunto de datos ruidoso. Estos modelos tienen un sesgo bajo y una varianza alta. Estos modelos son muy complejos, como los árboles de decisión, que tienden a sobreajustarse.

Underfitting and Overfitting

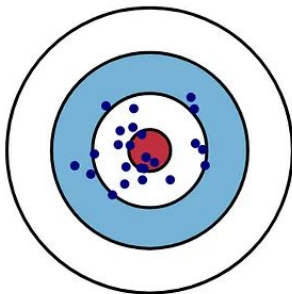
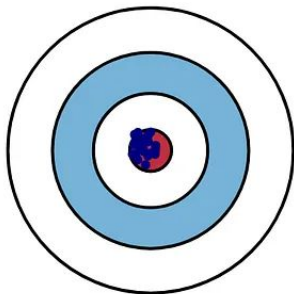
Low Variance

High Variance

High Bias

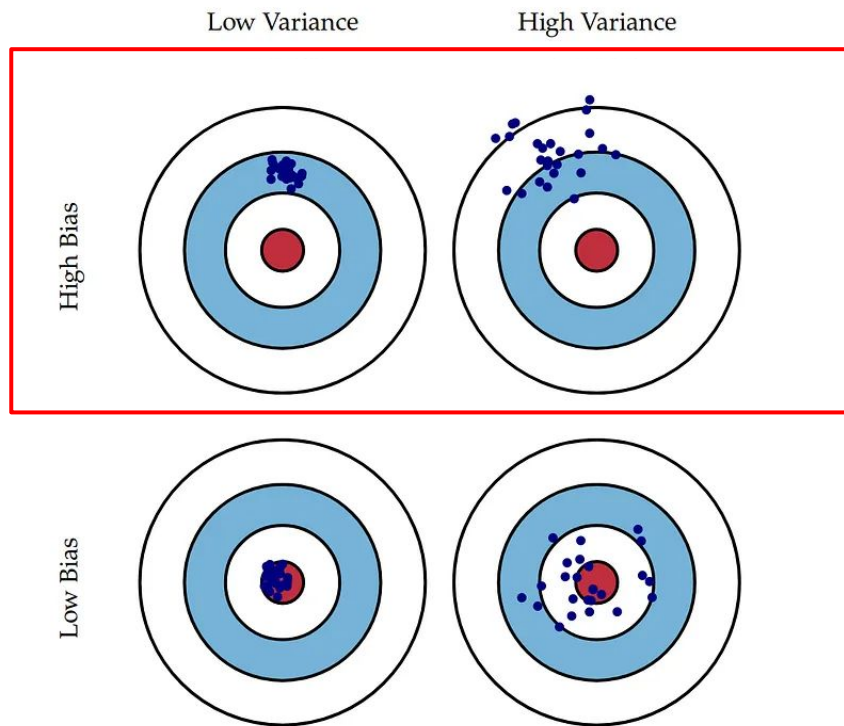


Low Bias



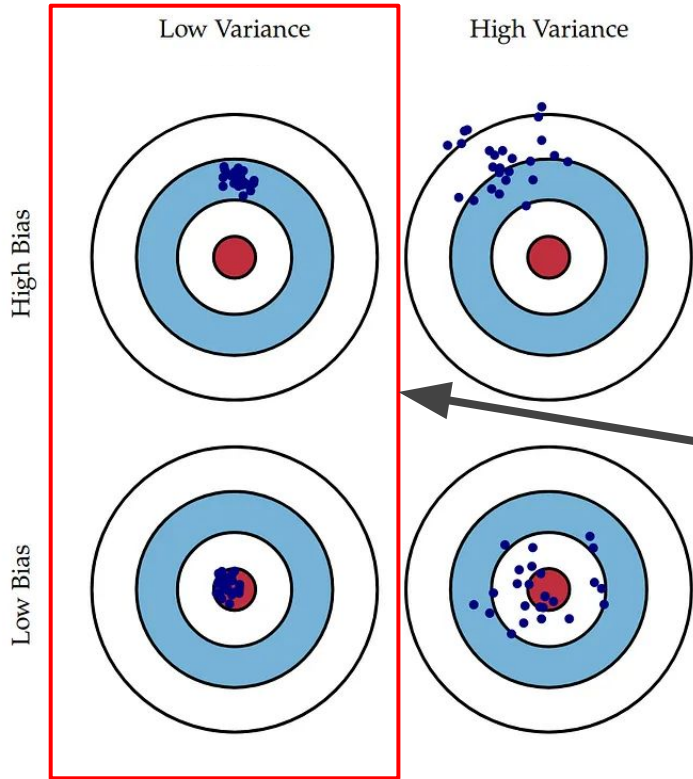
Predicciones “Correctas” que no están desviadas por un alto sesgo

Underfitting and Overfitting



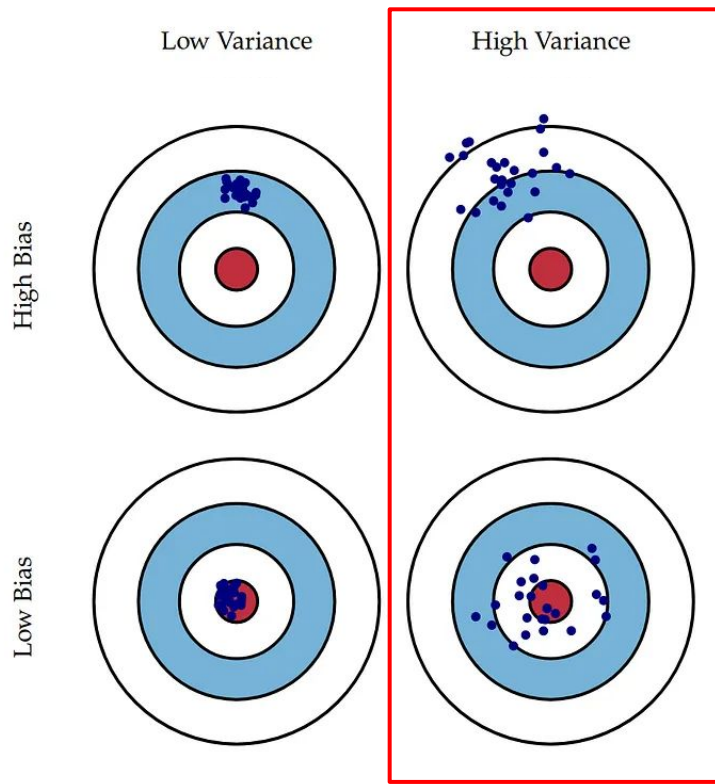
Predicciones “Incorrectas” fuera de lo esperado

Underfitting and Overfitting



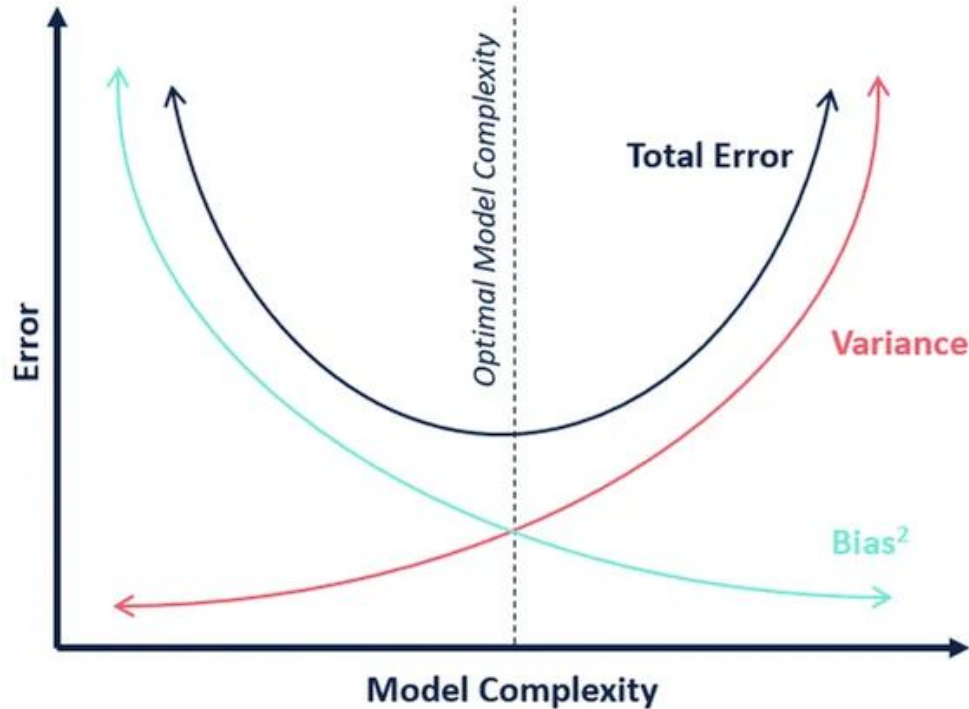
Las predicciones están en la misma vecindad ya sea para valores obtenidos correctos e incorrectos

Underfitting and Overfitting



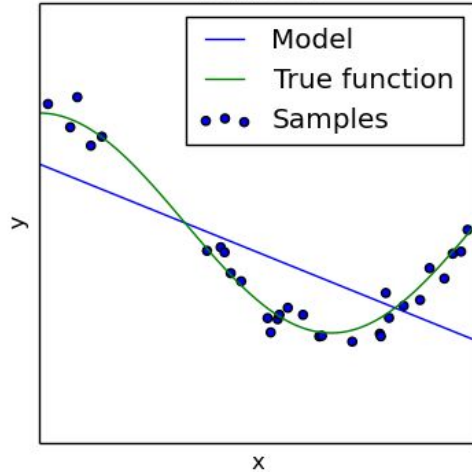
Las predicciones poco consistentes,
las predicciones son variadas

Equilibrio entre Sesgo y Varianza

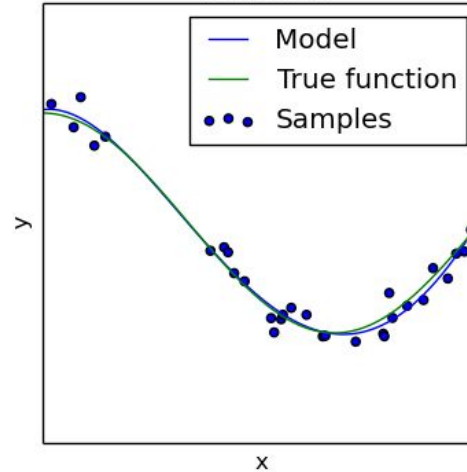


Como la complejidad del modelo tiene efectos inversos sobre el sesgo y la varianza, el problema se reduce nuevamente a buscar un equilibrio entre ambos

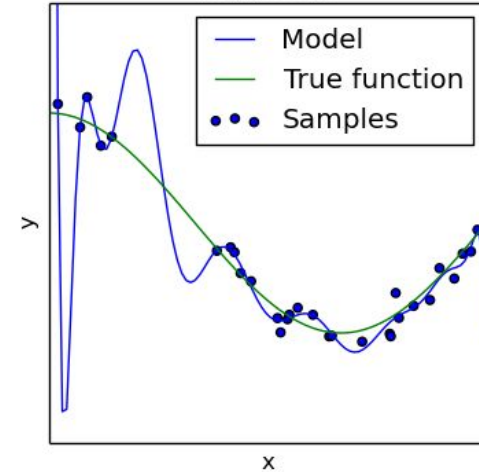
Ejemplo: Regresión



**Modelo con alto Sesgo:
Bajo ajuste de datos**

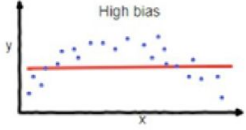
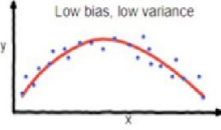
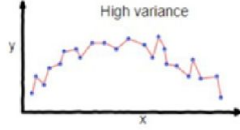
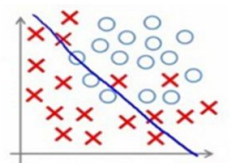
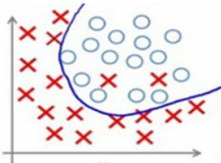
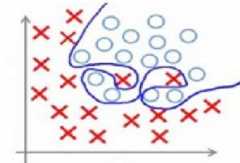

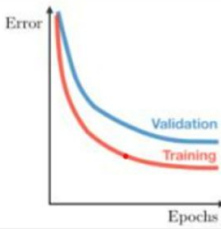
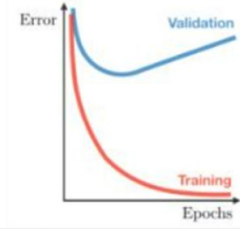


Modelo óptimo

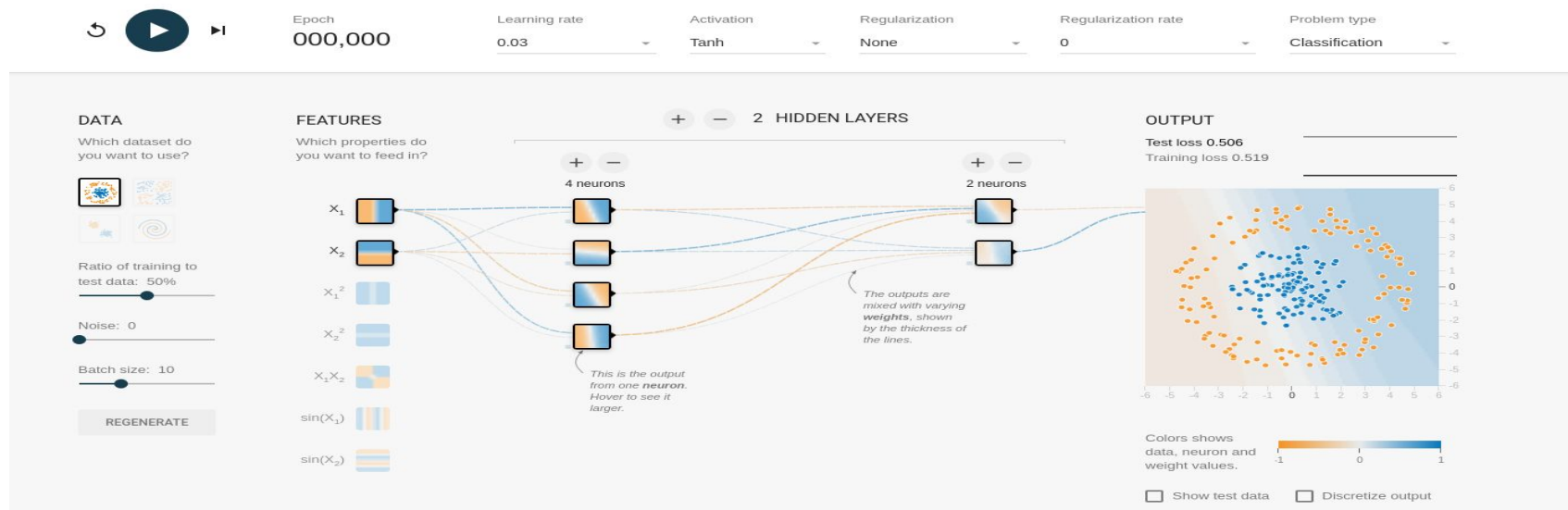


**Modelo con Alta varianza:
Sobreajuste de datos**

En resumen

	Underfitting	Optimal-fitting	Overfitting
Symptoms	<ul style="list-style-type: none"> • High Bias • High training error • Training error close to Testing error • Unable to capture the relationship between training data and labels • Inaccurate (not valid) 	<ul style="list-style-type: none"> • Low bias, low variance • Optimize training error • Training error slightly lower than test error • Capture well the relationship between training data and labels • Accurate (valid) and Consistence (reliable) 	<ul style="list-style-type: none"> • High Variance • Very low training error • Training error much lower than test error • Customize too much the relationship between training data and labels • Inconsistence (not reliable)
Regression			
Classification			
Deep Learning			

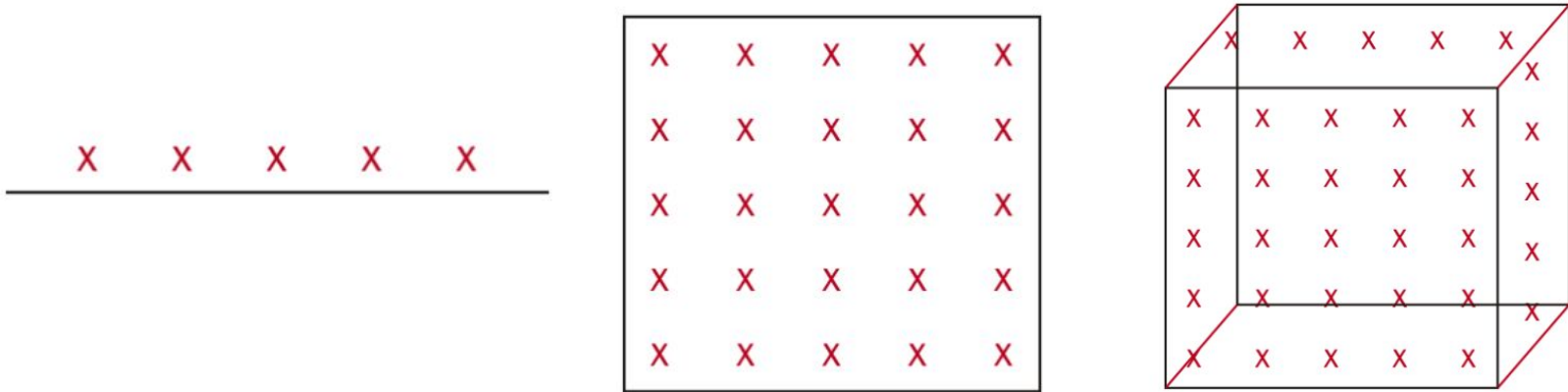
Tinker With a **Neural Network** Right Here in Your Browser.
Don't Worry, You Can't Break It. We Promise.



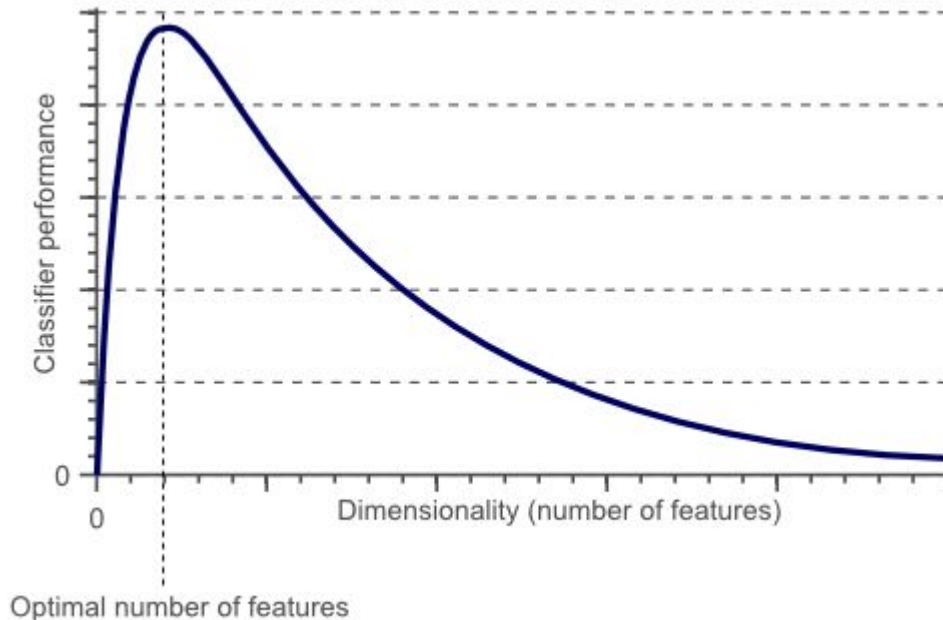
<https://playground.tensorflow.org/>

La maldición de la dimensionalidad

La maldición de la dimensionalidad es un fenómeno que se produce cuando el número de características o atributos de un conjunto de datos aumenta significativamente. Esto puede ser un problema, ya que la complejidad del modelo aumenta con el número de características.

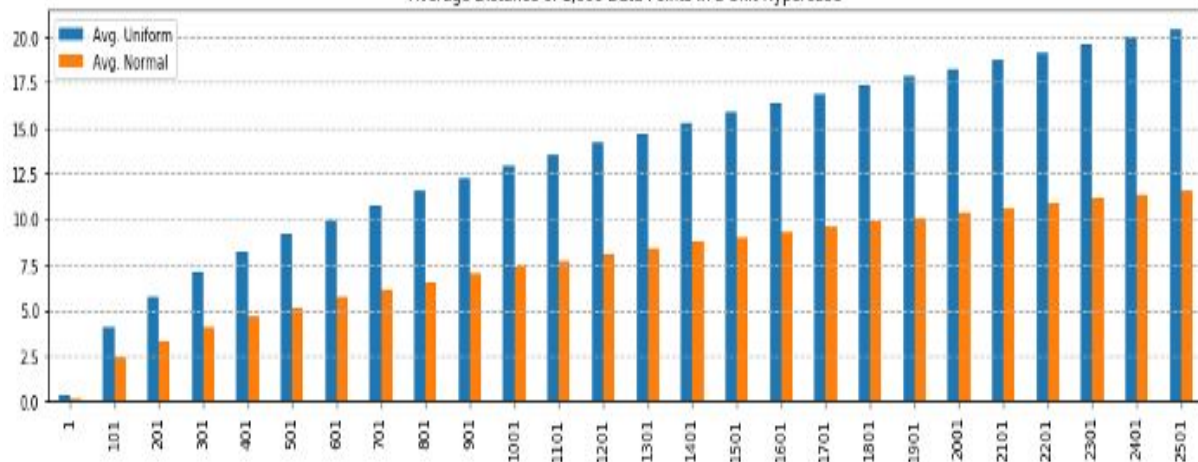


Hughes Phenomenon



El fenómeno de Hughes muestra que a medida que aumenta el número de características, el rendimiento del clasificador también aumenta hasta que alcanzamos el número óptimo de características. Agregar más funciones basadas en el mismo tamaño que el conjunto de entrenamiento degradará el rendimiento del clasificador.

Average Distance of 1,000 Data Points in a Unit Hypercube



$$d(p, q) = \sqrt{\sum_{i=1}^n (p_i - q_i)^2}$$

Es reducir la dimensionalidad de los datos mediante técnicas de selección de características o extracción de características. Esto implica identificar las características más importantes y relevantes para el modelo y descartar las demás.

- ❑ Análisis de Componentes Principales (PCA)
- ❑ Singular Value Decomposition (SVD):
- ❑ Representación de características con redes prr-entrenadas

El análisis de datos es el proceso de recopilar, organizar e interpretar datos para descubrir ideas y sacar conclusiones. Es la base de la ciencia de datos se puede utilizar para descubrir patrones, tendencias y correlaciones en los datos.

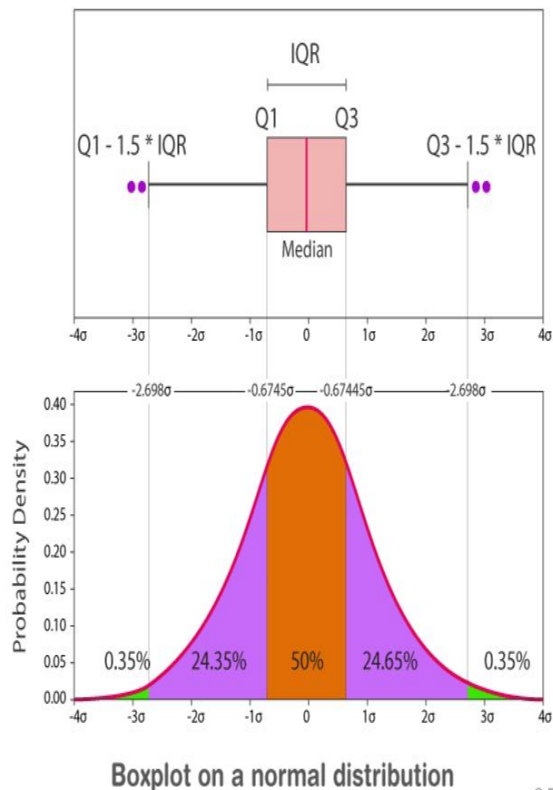
El análisis de datos permite:

- ☐ Comprender el comportamiento del cliente
- ☐ Mejorar la eficiencia operativa
- ☐ Toma de decisiones más inteligente

Tipos de Análisis de Datos

- ❑ Analítica Descriptivo
- ❑ Análisis Exploratorio
- ❑ Analítica Predictivo
- ❑ Analítica Prescriptiva
- ❑ Análisis Causal





El análisis descriptivo se utiliza para resumir datos y describir patrones y tendencias. Se utiliza para obtener una mejor comprensión de los datos e identificar oportunidades de mejora.

En este caso, se puede calcular la media, mediana, desviación estándar y otros estadísticos para obtener una mejor comprensión de las características de los datos.

- ❑ La edad media de los estudiantes en una universidad.
- ❑ La distribución de los ingresos de los hogares en una ciudad.
- ❑ El número de votos emitidos por cada partido político en una elección.

Análisis Exploratorio se utiliza para explorar las relaciones entre las variables. Puede incluir técnicas como gráficos de dispersión, diagramas de caja, tablas de contingencia y correlación

- ❑ La relación entre la altura y el peso de un grupo de personas.
- ❑ La asociación entre el nivel de educación y el salario de los trabajadores en diferentes sectores.
- ❑ La frecuencia de ciertas enfermedades en diferentes regiones geográficas.



El análisis predictivo se utiliza para pronosticar resultados futuros en función de datos pasados. Se utiliza para identificar riesgos y oportunidades potenciales y tomar decisiones en consecuencia.

- ❑ Predecir la demanda de productos en función del comportamiento de los clientes y las tendencias del mercado.
- ❑ Predecir el rendimiento de los estudiantes en un examen en función de sus calificaciones y su asistencia a clase.
- ❑ Predecir el tiempo de vida útil de una máquina en función de su historial de mantenimiento y su uso.

El análisis prescriptivo se utiliza para sugerir soluciones óptimas a un problema dado. Se utiliza para identificar el mejor curso de acción y tomar decisiones basadas en las predicciones.

El análisis prescriptivo se plantea como un problema de optimización :

- ❑ Optimización de recursos

El análisis causal se utiliza para identificar las causas fundamentales de un problema. Se utiliza para identificar oportunidades de mejora y tomar decisiones basadas en datos.

- ☐ Estudio de los efectos de un medicamento
- ☐ Análisis de la efectividad de una campaña publicitaria
- ☐ Evaluación de la efectividad de un programa de capacitación



Modelo de Analítica Ascendente de Gardner

