

Estadística

Modelando nuestra realidad

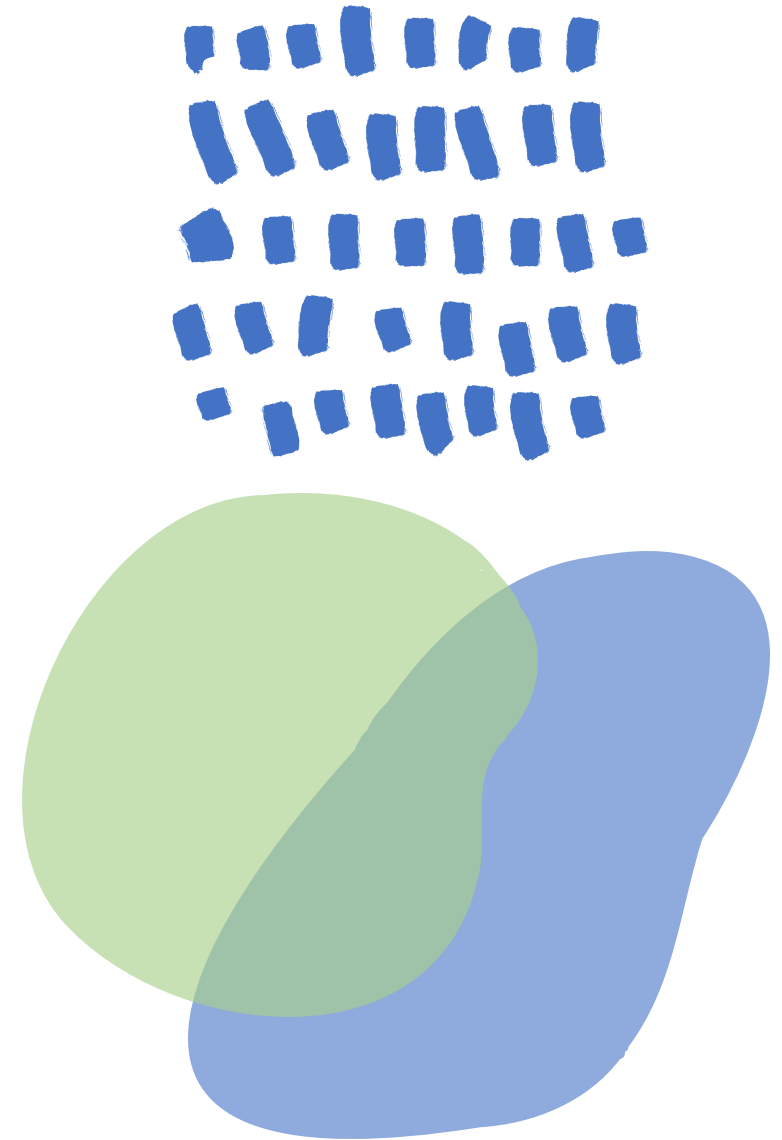
Erick I. Navarro-Delgado

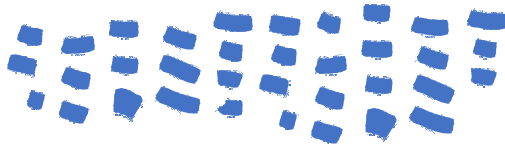
Candidato a Doctor en Bioinformática - UBC

Guadalupe Ayala Rodríguez

Estudiante de Maestría en Ciencias - UAEM

Clubes de Ciencia 2025





Temario

- Introducción: filosofía y motivación
- Revisión de conceptos clave: variables, variables aleatorias, distribuciones, teorema del límite central, sesgo-varianza trade-off, covariables
- Regresión lineal

Objetivos

- Ser familiares con la terminología
- Entender de forma clara los conceptos

Qué es la estadística?

- Ciencia que estudia la **colecta**, el **análisis**, **modelado**, e **interpretación** de datos, así como la **comunicación** de la incertidumbre de los resultados.
- Para utilizarla correctamente necesitamos:
- Entendimiento de las metodologías y del contexto en el cual se está usando
 - Usar la estadística como “recetas de cocina” puede resultar en procedimientos no robustos, malinterpretaciones y conclusiones erróneas
- Esencial en la forma en que la mayoría de las ramas científicas en la actualidad interpretan los datos y entienden al mundo

¿Cómo surgió la estadística (en el mundo occidental)?

Antes del siglo XIX: números y estadísticas son habitualmente subestimadas por la población

- “El poder de un estado no puede entenderse solamente con conocer su tamaño, población, producto interno bruto, o el número de animales que posee”
- La única forma de entender el mundo es a través de descripciones cuidadosas y conocimiento de la historia de ese fenómeno (cualitativo).

Siglo XIX: explosión de recolección de datos por el estado

- Inicialmente una nueva tecnología de los estados ante la competencia industrial, comercial y marcial. Ligado a afianzar el poder político, militar, colonial e industrial.
- Herramientas matemáticas comienzan a desarrollarse para entender estos datos
- Choque cultural: análisis empíricos desafían la forma de ver y entender al mundo
- Física social: científicos utilizan estadística para estudiar las características humanas

¿Cómo surgió la estadística (en el mundo occidental)?

Inicios del siglo XX: Se desarrollan la mayoría de las bases para la estadística moderna (frecuentista) para “entender la naturaleza humana” (biometría)

- Trasfondo racista y eugénico: ranking de personas y razas con el objetivo de “mejorar” la especie humana - purificar características deseadas
 - Galton desarrolla la regresión lineal y asocia caracteres biológicos con razas
 - Pearson desarrolla la correlación y utiliza resultados para impulsar políticas anti-inmigrantes de “razas inferiores” en Reino Unido
 - Correlación se interpretaba como causalidad
 - Spearman desarrolla la correlación y propone inteligencia diferencial en distintas razas
- Estadística juegan un papel crucial en la institucionalización del racismo (e.g. estadísticas criminales en UK)
- Historia de la estadística y sus prácticas se entrelazan con la **historia de intentar formalizar jerarquías de raza, sexo y clase.**

¿Cómo surgió la estadística (en el mundo occidental)?

Mediados del siglo XX: La estadística al servicio de la guerra (2ª guerra mundial)

- National Security Agency institucionaliza la colección de datos, algoritmos y formas de análisis – espías son pioneros de el almacenamiento de datos a gran escala
 - Industria les sigue el paso
- “Sobre-matematización” de la estadística: enfoque se mueve al desarrollo de modelos abstractos
- Primeras ideas de “máquinas con inteligencia” (Turing) e “inteligencia artificial” (McCarthy): objetivo de lograr inteligencia humana

Finales del Siglo XX: Surgimiento del análisis de datos

- Desarrollo de machine learning y minería de datos para analizar el creciente número de datos
- Se busca destilar los datos y encontrar patrones informativos que resuelvan preguntas
- Cantidad de datos es aún relativamente pequeña

¿Cómo surgió la estadística (en el mundo occidental)?

Siglo XXI: De minería de datos al “Big Data”

- Avance en ciencias de la computación permiten el almacenamiento masivo de datos
- Surge la disciplina de “ciencia de datos”
- Estadísticos aprenden ciencias de la computación para retomar conceptos de estadística y Machine Learning, y aplicarlos en datos masivos

Muy recientemente:

- Grandes avances en modelos de predicción - ahora se busca entender los modelos
- Explosión de la “arquitectura persuasiva” en mercadotecnia (ads personalizados)
 - El nuevo oro del siglo XXI: datos
- Surgimiento de comités de ética en equipos de Inteligencia Artificial
- Algunas herramientas de inteligencia artificial disponibles para la población

Conceptos básicos

Variable: un elemento, característica o factor que se encuentra sujeto a cambio o variabilidad

- La mayoría de preguntas en estadística pueden formularse como “cuál es la relación entre dos o más variables?”

Variable aleatoria: variable resultante de la medición de una cantidad sujeta a variación

- Respuesta/variable dependiente: variable en la cuál estamos interesados en un experimento (la expresión de un gen X, tirar un dado, etc.)
- Predictor/variable independiente : variable que (en nuestra hipótesis), puede explicar/predecir la respuesta

Una variable aleatoria tiene una **distribución**.

Conceptos básicos

- **Distribución de probabilidad:** función matemática que mapea eventos/estados de una variable aleatoria con su probabilidad.
- Probabilidad: Un número asignado a un evento/estado que describe qué tan probable es que ocurra
 - Entre 0 y 1
 - Representa la frecuencia del evento (multiplicando la probabilidad por el número de muestras)
- *Ejemplo en R

Tipos de variables aleatorias

- **Discreta:** estados posibles son números enteros
- **Continua:** cualquier estado es posible dentro de un intervalo

¿La variable que usamos en el ejemplo es variable discreta o continua?

¿Para qué usamos la estadística?

Para **inferir**: entender la relación entre dos o más variables

- Estamos interesados en cuál es la relación entre un predictor y una respuesta

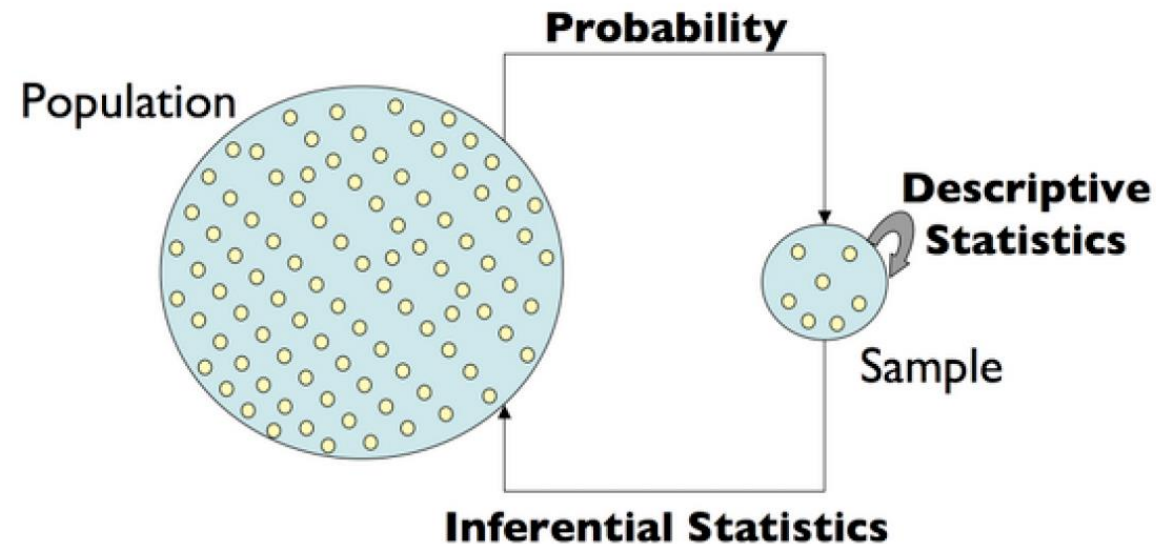
Para **predecir**: el valor de un resultado no observado con base en variables observadas

- La relación entre los predictores y la respuesta es irrelevante (no nos interesa saber $f(x)$).
- Buscamos modelos que generen buenas predicciones.
- La mayoría de modelos de IA se mueven en este paradigma

Una misma herramienta estadística puede servir para ambas cosas.

Inferencia estadística

- Marco teórico para obtener conclusiones acerca de una población a partir de una muestra de datos



- La probabilidad nos permite conocer la incertidumbre y realizar predicciones
- La estadística inferencial nos permite sacar conclusiones de los datos (pruebas de hipótesis, etc.)

Modelo matemático/estadístico

Objetivo: inferir una función (i.e. ecuación matemática) que prediga la variable de respuesta

- Se necesita **estimar parámetros** (i.e. ajustar el modelo a los datos)

Un modelo es una representación que (idealmente) describe los datos y (principalmente) la población de la que tomamos la muestra

Podemos usar un modelo para:

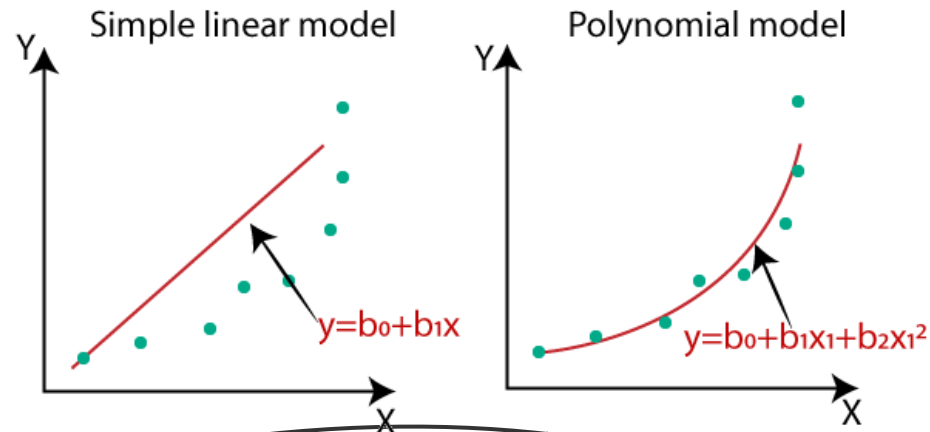
- Pruebas de hipótesis
- Predicción
- Simulación

Modelo lineal

- Supone una relación lineal entre el predictor y la respuesta
- Intuición: encontrar la línea que mejor explica los datos

Representación matemática: $y = \beta_0 + \beta_1 x + e$

*No todos los modelos lineales son “una línea” – la “relación lineal” se refiere a los coeficientes



Tipos de modelos lineales

- Regresión lineal simple: modelos que utilizan un solo predictor
- Regresión lineal múltiple: modelos que utilizan múltiples predictores
- Regresión lineal multivariada: modelos para múltiples variables de respuesta

Algunos conceptos:

- Los **valores ajustados** (los valores predichos): valores de Y que se generan si introducimos nuestros valores X en nuestro modelo.
- Los **residuos** = diferencias no explicadas por nuestro modelo .

Supuestos

Hay cuatro supuestos asociados con un modelo de regresión lineal:

1. **Linearidad:** La relación entre X y la media de Y es lineal.
2. **Homosedasticidad:** La varianza del residuo es la misma para cualquier valor de X .
3. **Independencia:** Las observaciones son independientes unas de otras.
4. **Normalidad:** Para cualquier valor fijo de X , Y se distribuye normalmente.

¿Cómo se ajusta un modelo lineal?

Web interactiva de modelo lineal OLS

- <https://setosa.io/ev/ordinary-least-squares-regression/>

Interpretación de modelos lineales múltiples

- En un modelo de regresión lineal de la forma:

$$y = \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$$

- El coeficiente β_k expresa el impacto de un cambio de una unidad en la variable predictora X_k , sobre la media de la respuesta $E(y)$, siempre que todas las demás variables se mantengan constantes.
- El signo del coeficiente da la dirección del efecto.

¡Ejemplos! ¿Qué significan esos coeficientes?

- $E(y) = 1.8 - 2.35X_1 + X_2$,
- $E(y) = 1.1 + 2.1X_1 + 1.5X_2$

Problemas

- Non-linearity of the response-predictor relationship
 - Identifiable with residual plots
- Correlation of error terms (non-independence)
 - E.g. siblings
- Heteroskedasticity
 - Can be mitigate by transforming Y (e.g. $\log(Y)$)
- Outliers/high leverage points
- Collinearity
 - Causes standard error of B's to grow -> reducing t statistics and impacting power
 - Detected through correlation matrix or Variance Inflation Factor

El diseño utilizado para obtener la muestra es esencial para generar modelos con buena generalización

- Las observaciones en una variable deben de ser:
 - Independientes – la observación 1 no depende de la observación 2
 - Provenir de la misma distribución (población) – tienen la misma distribución de probabilidad subyacente
 - Obtenidas de forma aleatoria
- ¿Qué tipo de observaciones violarían estos supuestos?
- ¿Qué pasa si se violan estas suposiciones?

¿Cómo saber qué tan bueno es tu modelo?

Mean Absolute Error

$$MAE = \frac{1}{n} \sum \left| y - \hat{y} \right|$$

Diagram illustrating the Mean Absolute Error (MAE) formula:

- $\frac{1}{n}$: Divide by the total number of data points
- \sum : Sum of
- y : Actual output value
- \hat{y} : Predicted output value
- $|y - \hat{y}|$: The absolute value of the residual

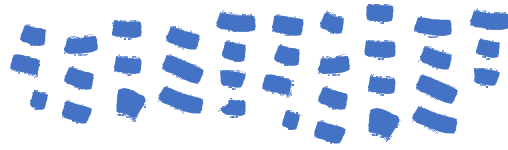
*Hay muchas más métricas complementarias

Referencias

- Chris Wiggins y Matthew L. Jones (2023). *How data happened*. W.W. Norton & Company. USA.
- Korthauer, K. (2022). STAT540: Review of Probability and Statistics (slides). The University of British Columbia.

Receso

10 minutos



Introducción a la inteligencia artificial y Machine Learning

Objetivos de aprendizaje

- Identificar similitudes y diferencias entre la IA, el ML, la estadística y la ciencia de datos
- Tipos de modelos de aprendizaje automático y sus usos
- Identificar algunas de las consideraciones éticas al usar ML

**¿ML en el mundo de la salud:
exageración o una nueva
realidad?**

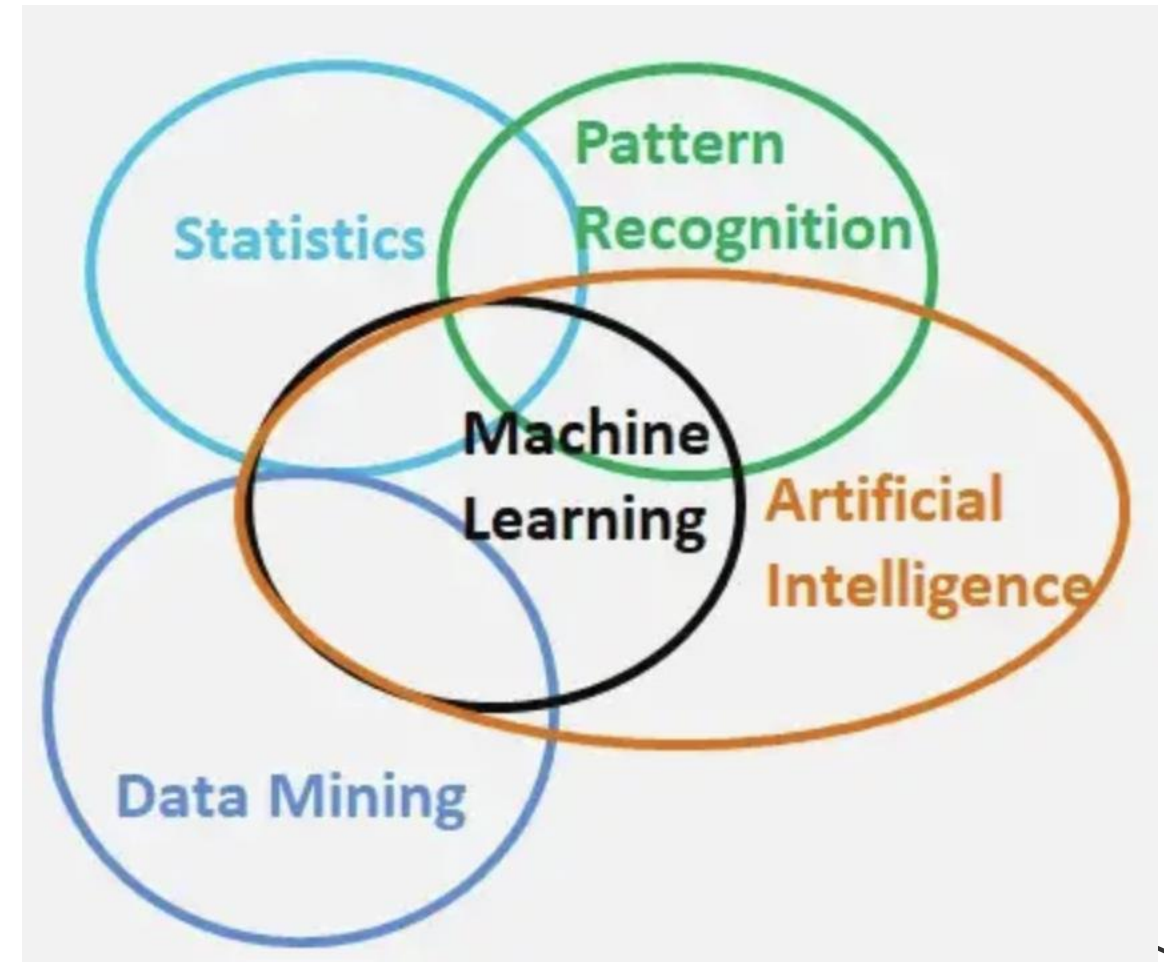
Estadística/Machine Learning/IA

Estadística:

- Enfoque: Inferencia de parámetros
- Los modelos deben cumplir con los supuestos
- Los modelos son a menudo interpretables

Machine learning:

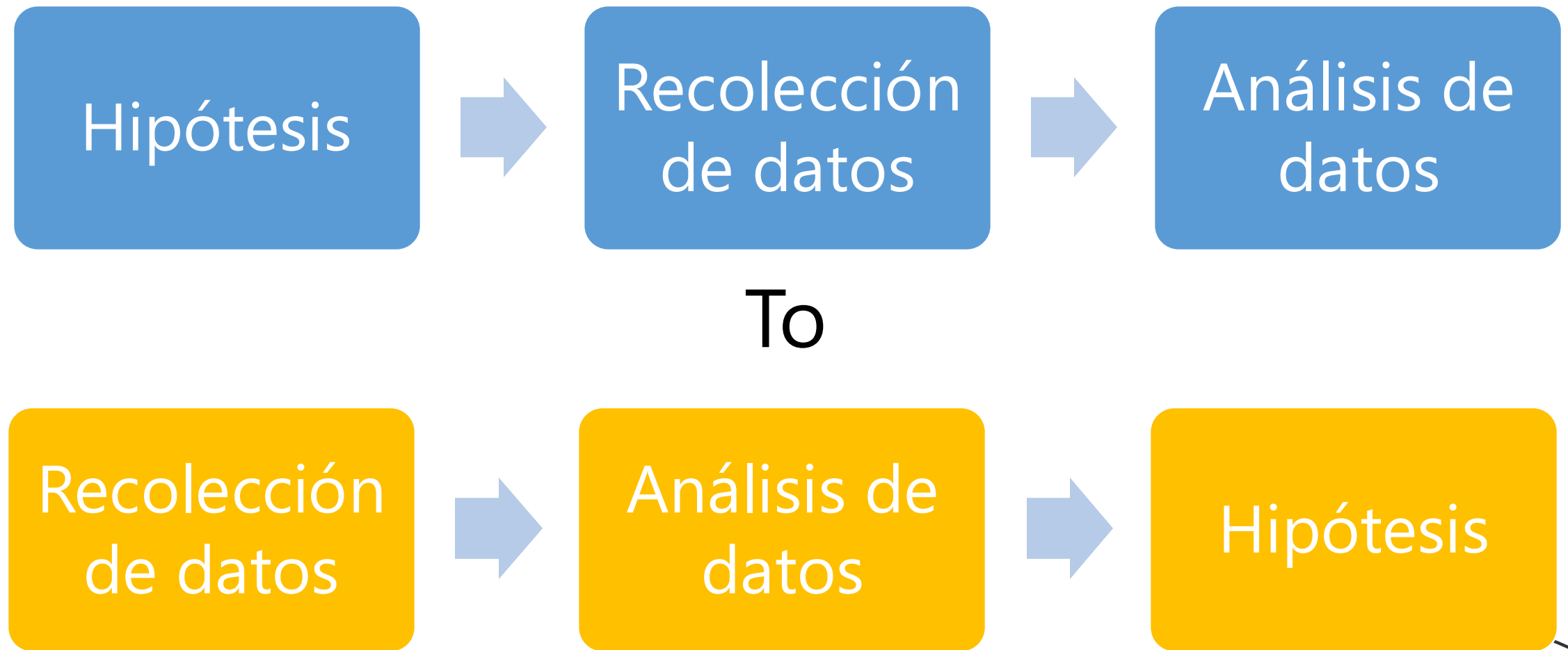
- Enfoque: predicción
- Hace pocas suposiciones
- Modelos más flexibles
- Adecuado para datos más grandes
- "Caja oscura"



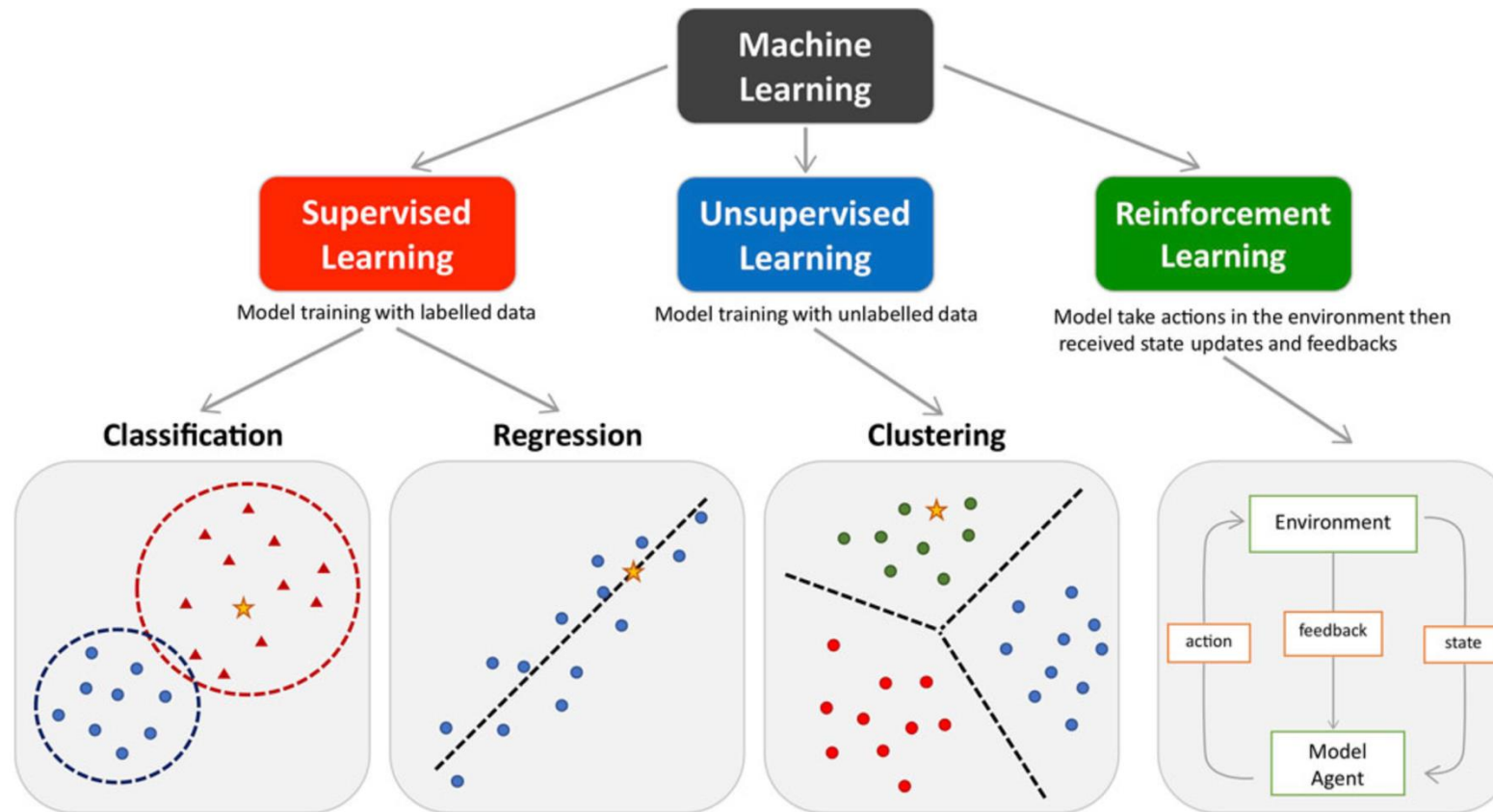
¿Qué es posible con ML?

- Descubrir subtipos de enfermedades o estratificar pacientes
- Identificar similitudes y diferencias entre los pacientes
- Predecir cuál es la terapia más eficaz
- Automatizar diagnósticos
- Predecir resultados clínico

Cambio de paradigma



Tipos de ML

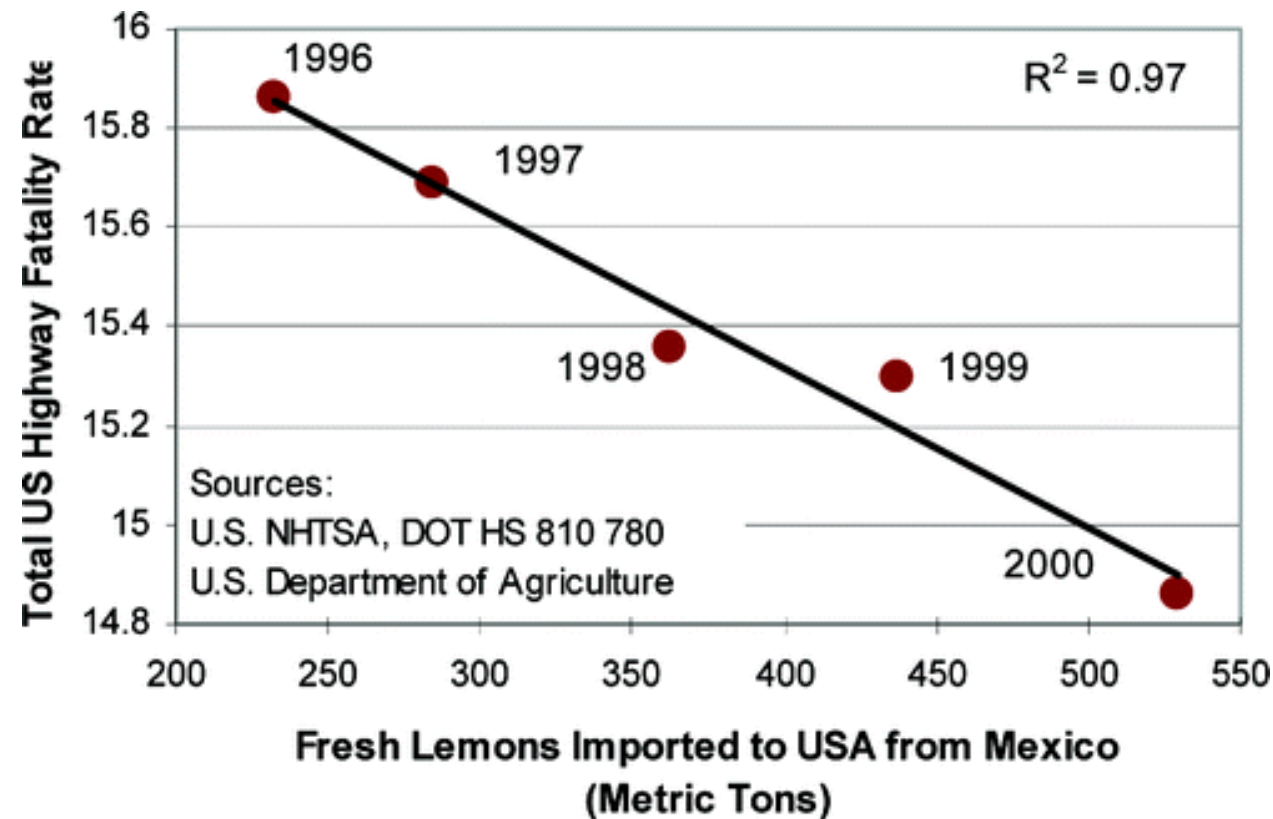


DOI: [10.3389/fphar.2021.720694](https://doi.org/10.3389/fphar.2021.720694)

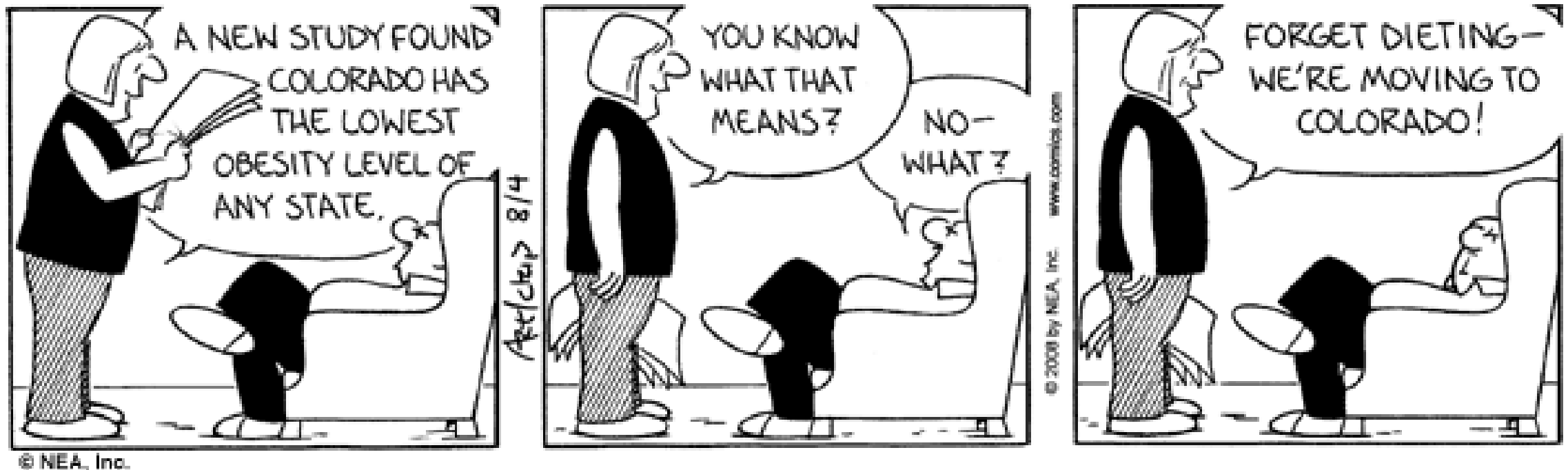
“La predicción permite identificar los mejores cursos de acción (por ejemplo, la elección del tratamiento) sin necesidad de comprender el mecanismo subyacente”

¿A favor o en contra?

¿Se pueden predecir las tasas de mortalidad en las carreteras a partir de la cantidad de limones importados?



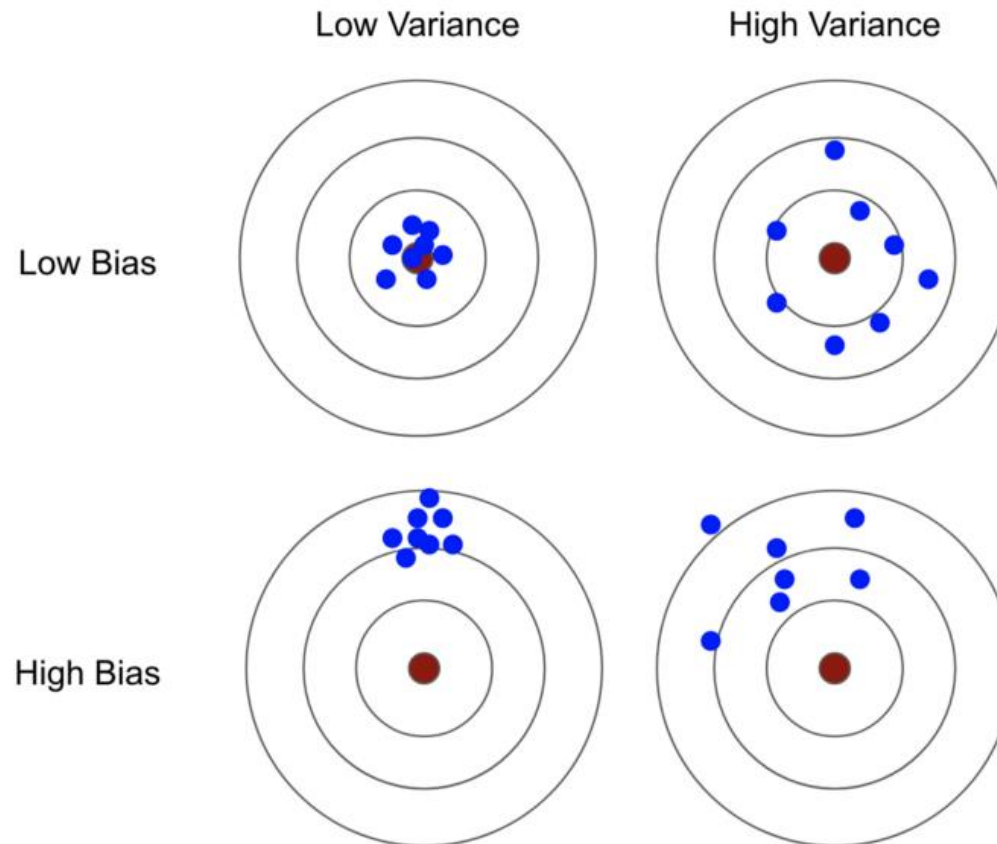
Correlación no es causalidad



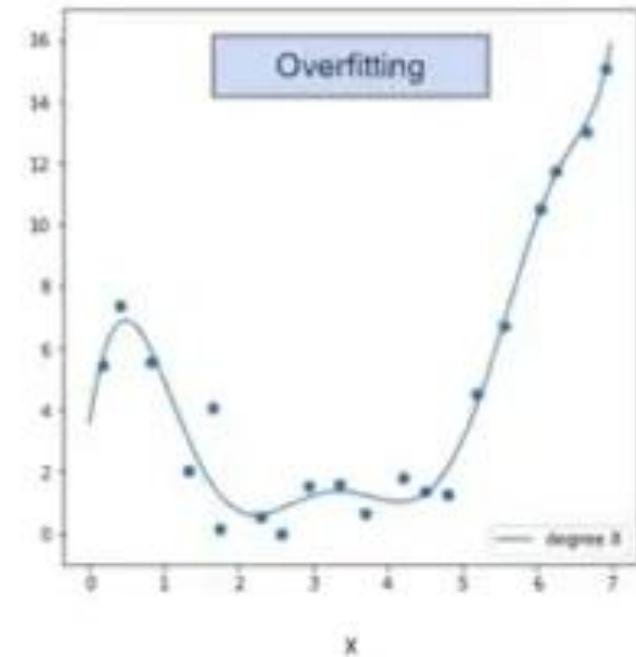
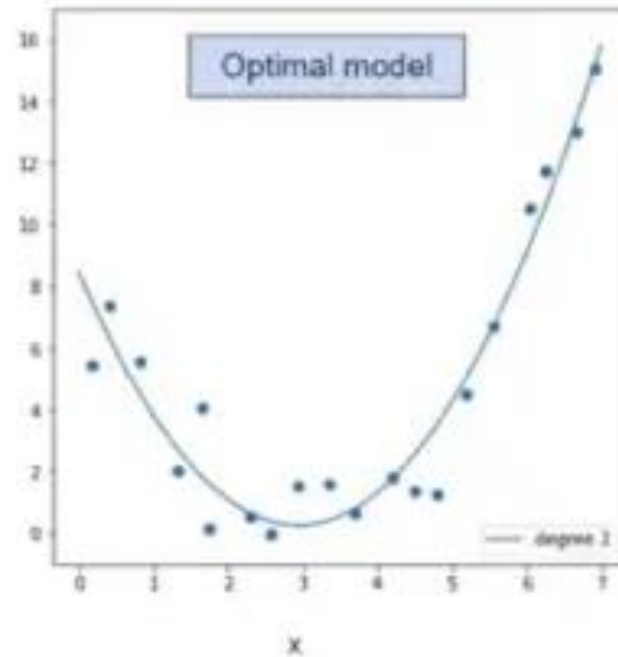
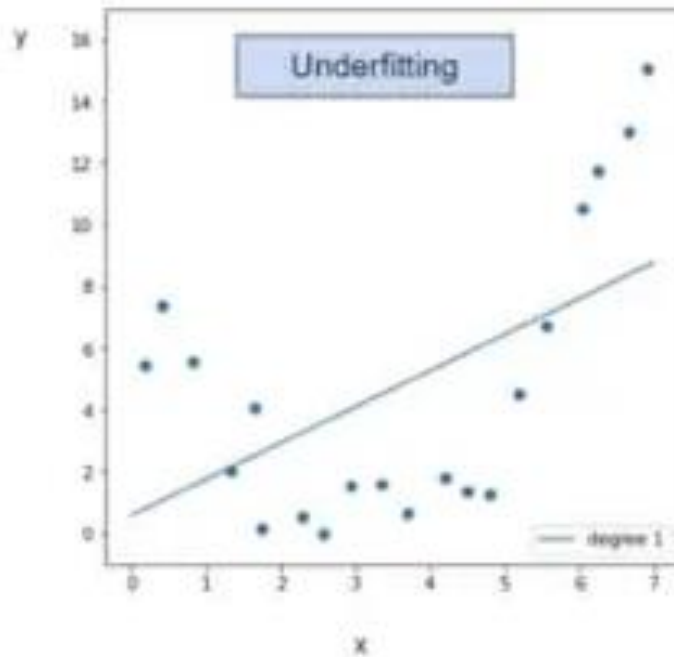
La IA/ML puede aprender patrones, pero no puede razonar: no puede responder a preguntas de "por qué", "qué pasaría si...", ni comprender conceptos físicos y biológicos rudimentarios

Una idea errónea

“El big data y el ML superan las barreras de las muestras pequeñas de la estadística”

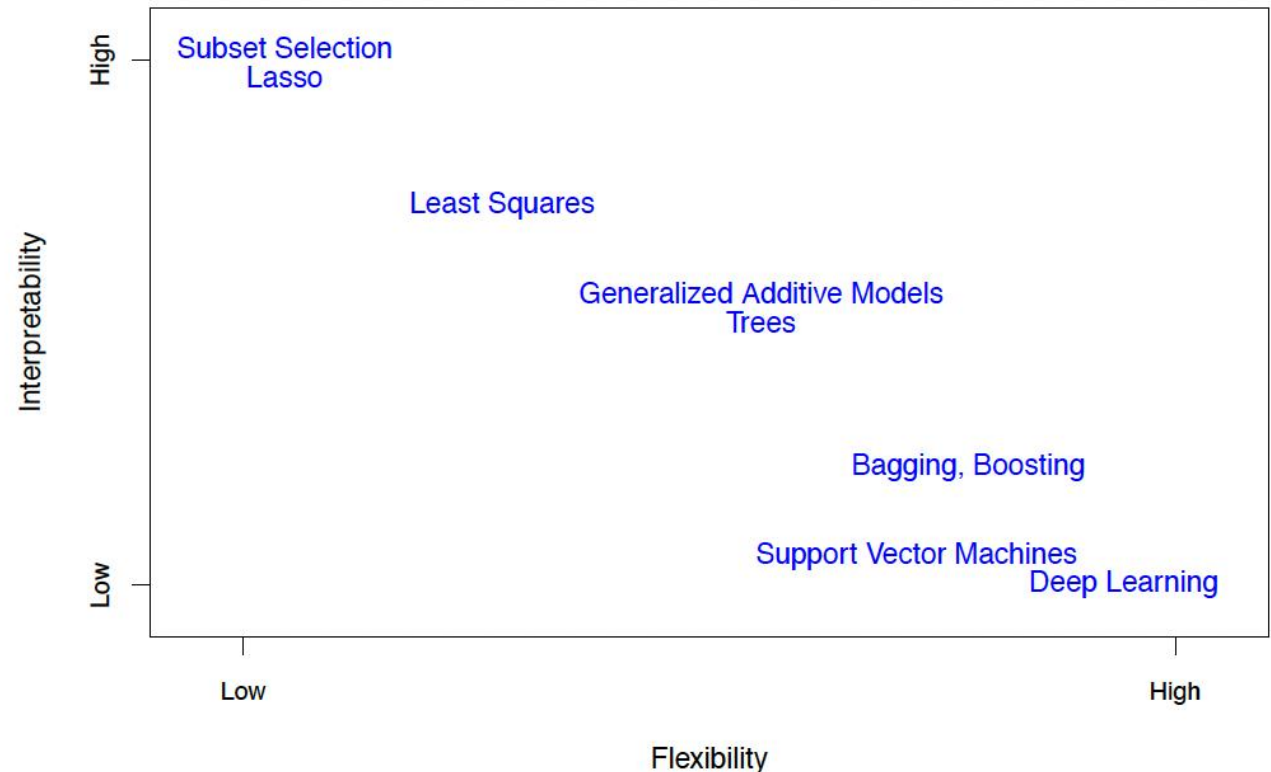


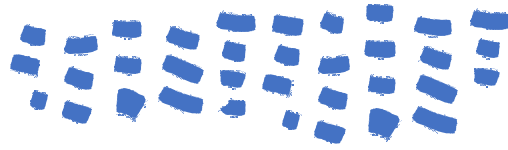
El balance underfitting/overfitting determina la generalización de nuestros modelos



La elección de un modelo depende de la situación y el objetivo

- Los métodos más flexibles tienen una varianza más alta pero un sesgo más bajo.
 - Varianza: qué tan sensible es a los cambios de datos / cuánto varía con diferentes conjuntos de datos
 - Sesgo: el error que se introduce al aproximar (forzar) un modelo con un método dado.
- El sesgo o varianza de un modelo es muy difícil de calcular, pero siempre debemos tener en cuenta esta relación. Es como un equilibrio entre overfitting y generalización.





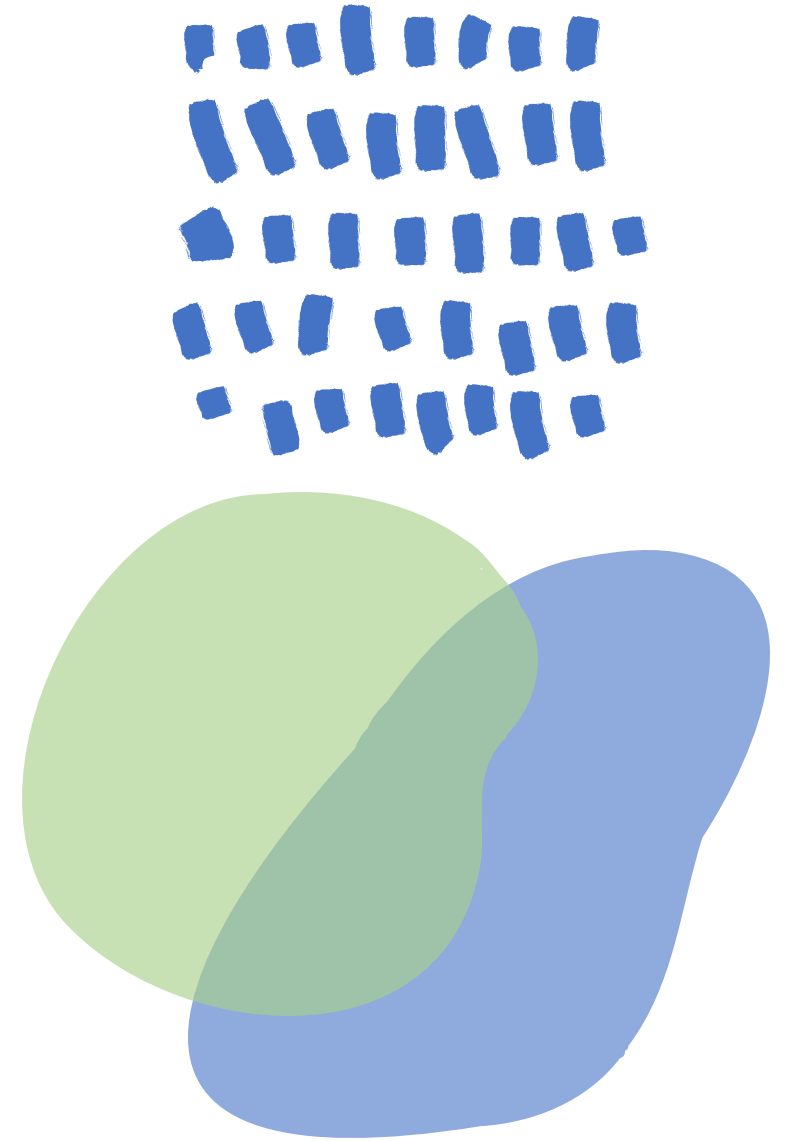
Receso

Comida (1 hora)



Problemas éticos de la inteligencia artificial

Discusión grupal



Inconvenientes y problemas éticos en la IA/ML

**¿Cuáles son los problemas
éticos de la IA?**

1. Issues arising from machine learning	
Privacy and data protection	Lack of privacy
	Misuse of personal data
	Security problems
Reliability	Lack of quality data
	Lack of accuracy of data
	Problems of integrity
Transparency	Lack of accountability and liability
	Lack of transparency
	Bias and discrimination
	Lack of accuracy of predictive recommendations
	Lack of accuracy of non-individual recommendations
Safety	Harm to physical integrity

2. Living in a digital world	
Economic issues	Disappearance of jobs
	Concentration of economic power
	Cost to innovation
Justice and fairness	Contested ownership of data
	Negative impact on justice system
	Lack of access to public services
	Violation of fundamental human rights of end users
	Violation of fundamental human rights in supply chain
	Negative impact on vulnerable groups
	Unfairness
Freedom	Lack of access to and freedom of information
	Loss of human decision-making
	Loss of freedom and individual autonomy
Broader societal issues	Unequal power relations
	Power asymmetries
	Negative impact on democracy
	Problems of control and use of data and systems
	Lack of informed consent
	Lack of trust
	Potential for military use
	Negative impact on health
	Reduction of human contact
	Negative impact on environment
Uncertainty issues	Unintended, unforeseeable adverse impacts
	Prioritisation of the “wrong” problems
	Potential for criminal and malicious use

3. Metaphysical issues	
	Machine consciousness
	“Awakening” of AI
	Autonomous moral agents
	Super-intelligence
	Singularity
	Changes to human nature

¿Cuáles son algunas de las consecuencias negativas?

Consecuencias

- Exacerbar la desigualdad
- Sesgado hacia aquellos que aportan la mayor cantidad de datos que pueden conducir a la discriminación
- Desafío al papel del juicio humano
- Pérdida de control personal
- Vigilancia

**¿Cuáles son los beneficios
éticos?**



**¿Somos capaces de superar los
problemas éticos?**



Receso

10 minutos



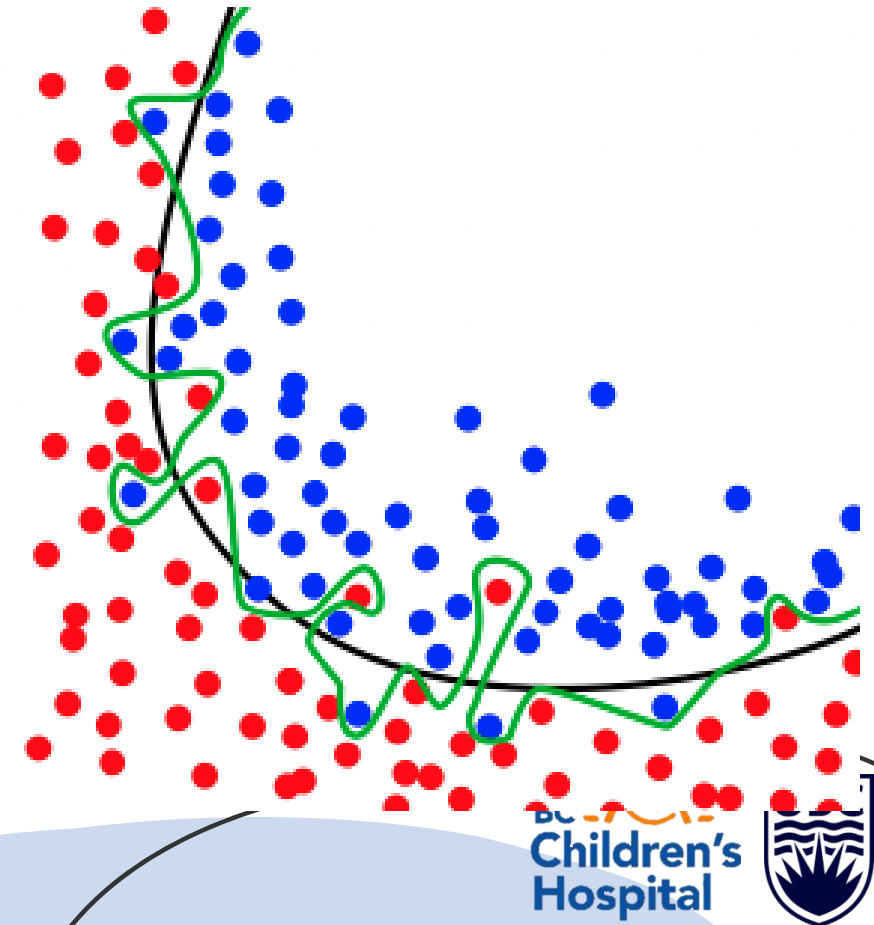
Introducción a modelos de regularización

¿Qué es el sobreajuste?

- El modelo predice muy bien con los datos con los que se entrenó, pero no funciona tan bien con los datos nuevos
- No logra generalizar

¿Por qué sucede?

- Muy pocas muestras para el training
- Datos ruidosos El modelo es demasiado complejo



¿Cómo detectar el sobreajuste?

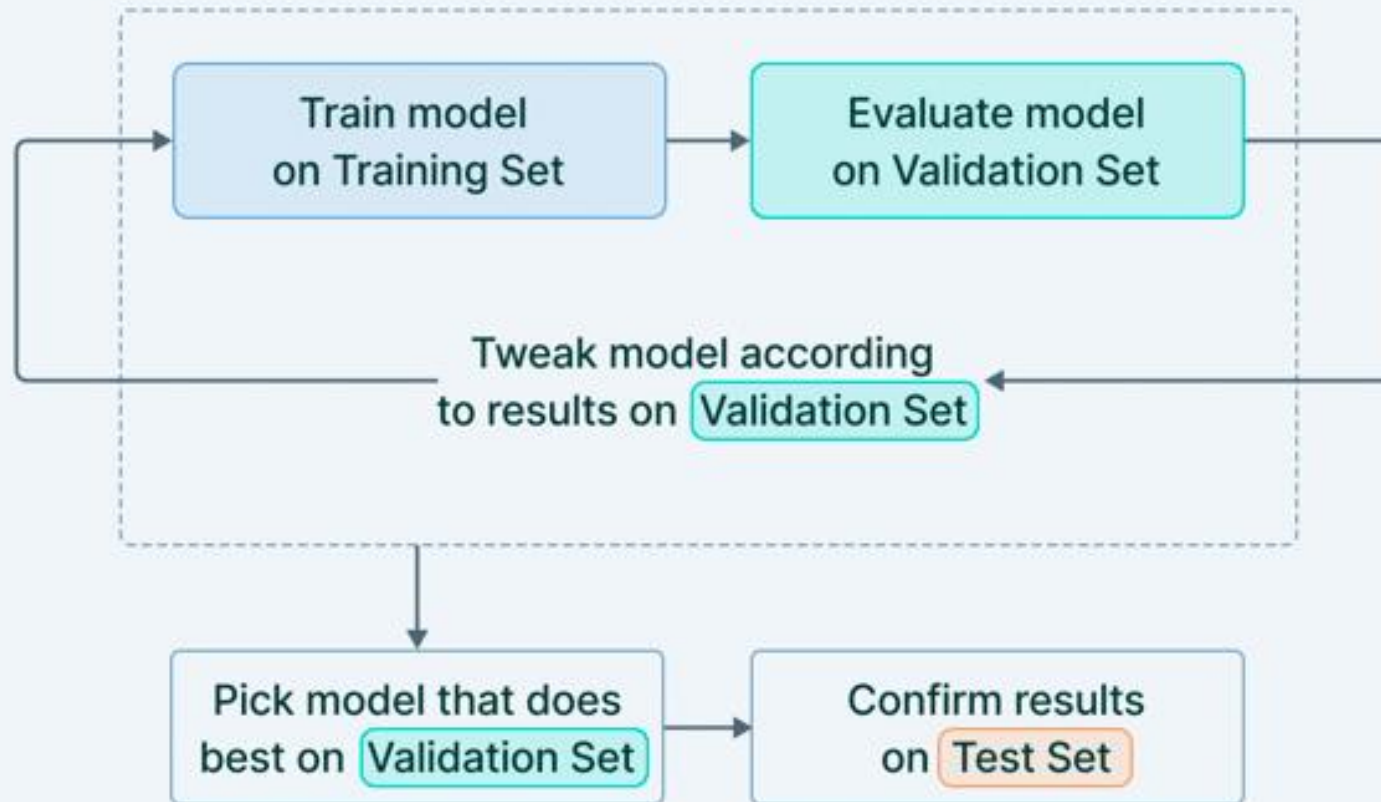
Evaluando modelos en conjuntos de datos que no se usaron para entrenarlos

- Train/Test split
- K-fold cross validation

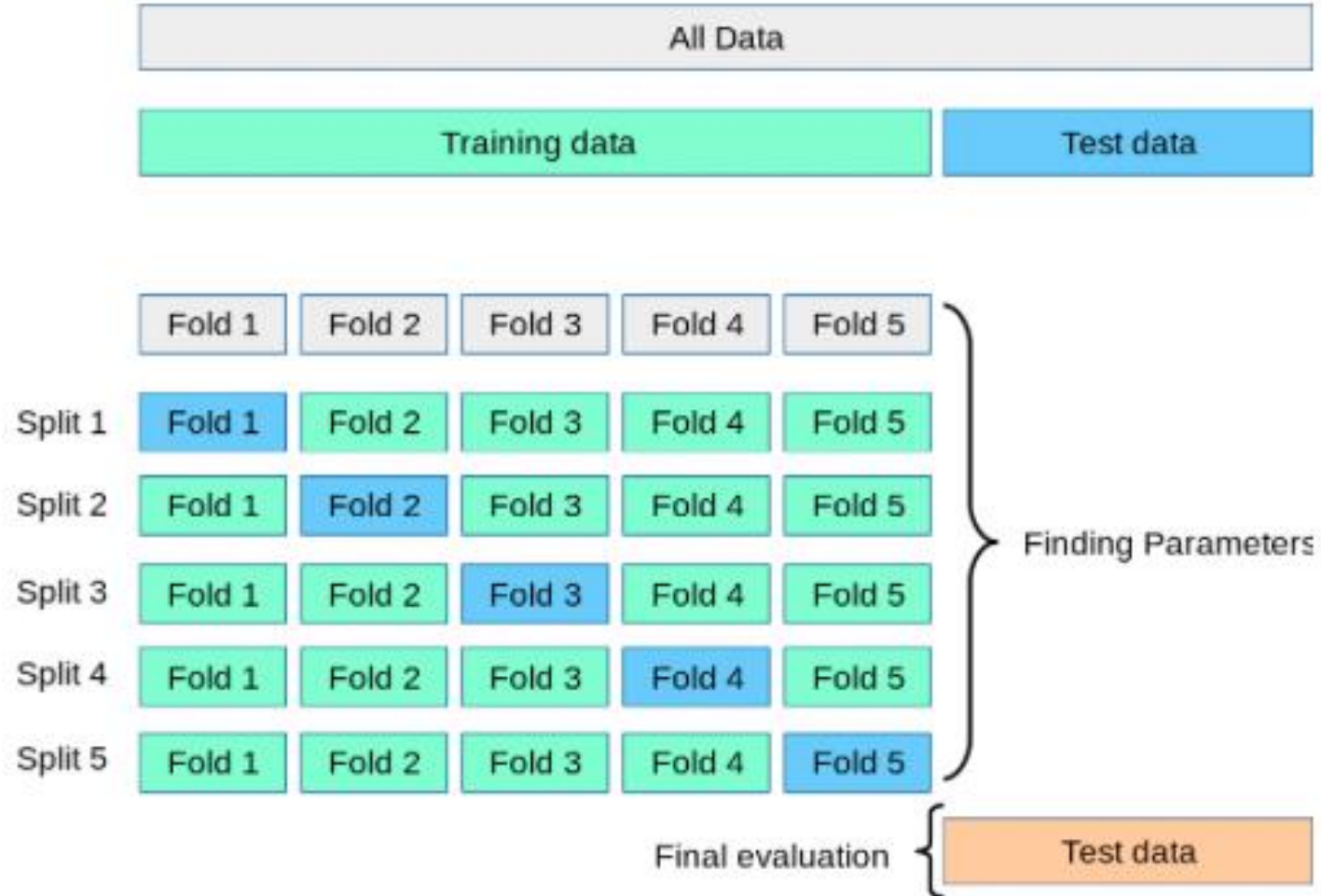
Train/Test/Validation

- Training set: datos utilizados para desarrollar el modelo. Debe ser representativo de la población que está modelando, incluida la elección de los hiperparámetros.
- Validation set: Un conjunto de datos independiente, que no se utiliza en el entrenamiento. Ayuda con la selección de modelos comparando modelos. El modelo aún se puede ajustar
- Test set: Un conjunto de validación final donde podemos comparar el rendimiento del modelo con el de la validación. El modelo está bloqueado.

Training data/validation/test



Cross validation: K-fold



Cómo evitar el sobreajuste

- **Pruning** Selección de característica/variable
- **Ensembling** Combinación de predicciones de varios algoritmos
- **Regularización** Aplicación de una penalización a la complejidad del modelo como parte del proceso de entrenamiento

¿Qué es el subajuste?

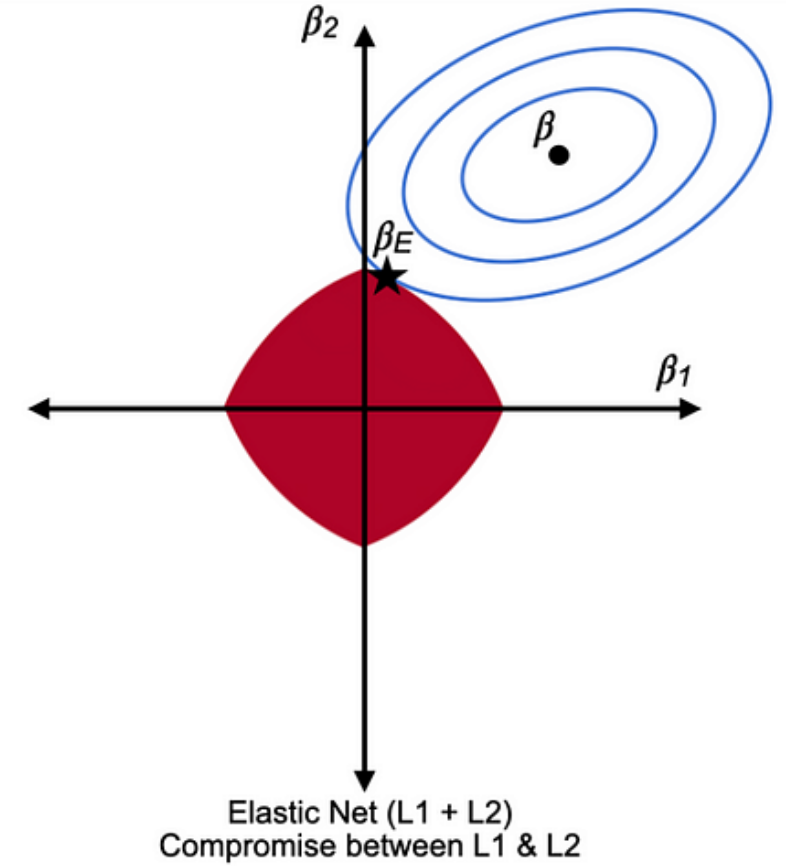
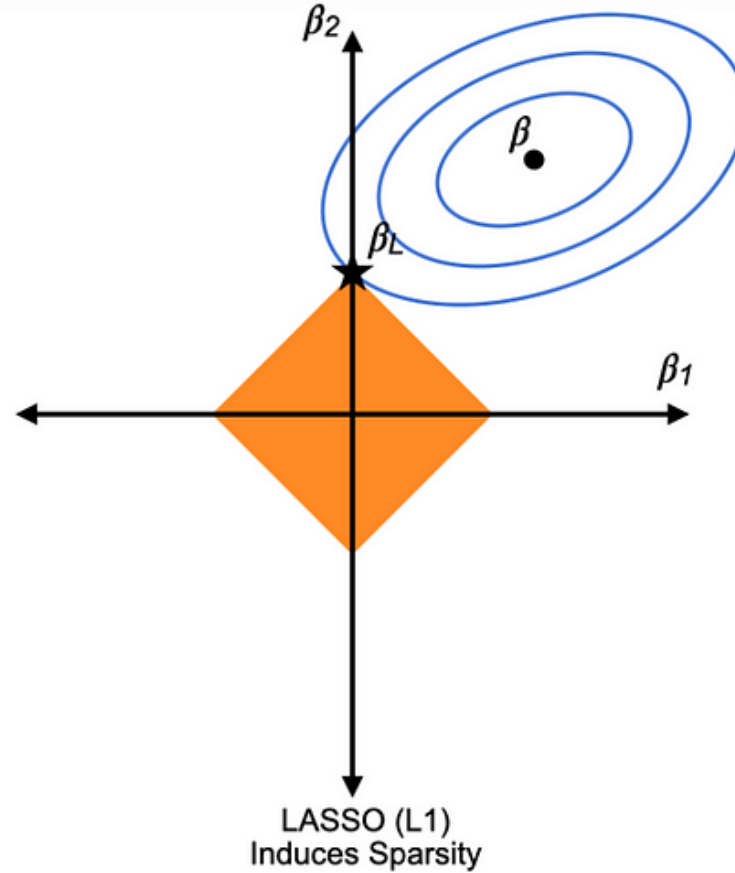
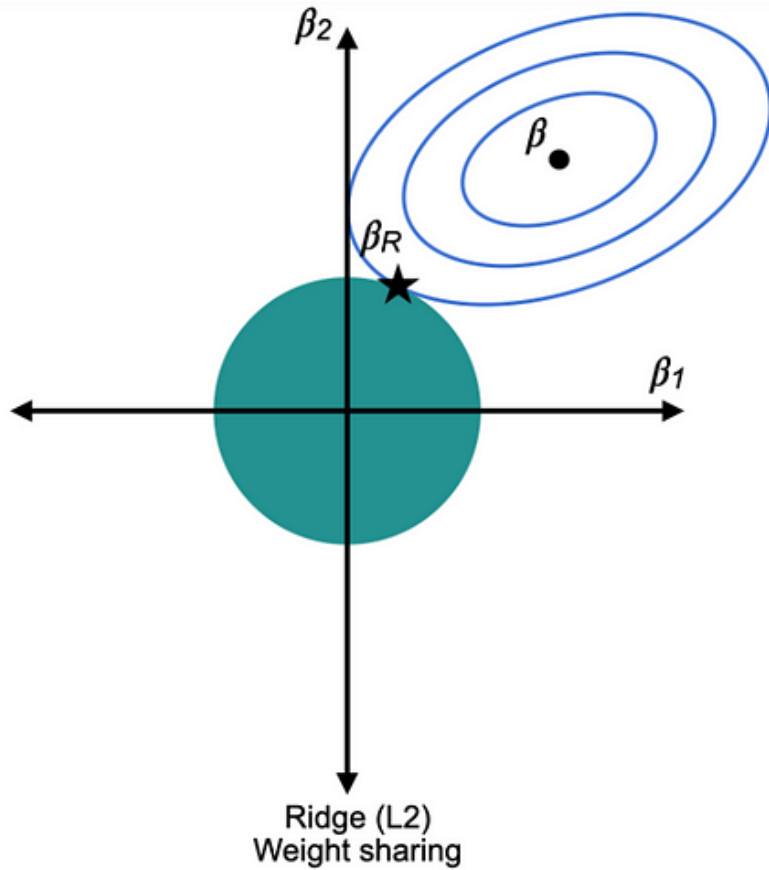
- El modelo no puede determinar una relación significativa entre X e Y
- No hay suficientes datos
- No hay señal en los datos
- Alto sesgo
- Alta varianza


Regularización

- Generalmente, se puede formular en forma de "función de pérdida no penalizada + función de penalización":

$$\hat{\beta} = \operatorname{argmin}_{\beta} \{L(D; \beta) + P(\lambda; \beta)\},$$

- $L(D; \beta)$ es una función de pérdida basada en los datos observados D y los coeficientes de regresión β para cuantificar la falta de ajuste.
- La función de penalización, $P(\lambda; \beta)$, mide la complejidad del modelo con el parámetro de ajuste λ . Como $\lambda \rightarrow +\infty$, se impone una mayor cantidad de penalización a β , y más componentes de β . Por otro lado, cuando $\lambda \rightarrow 0$, el modelo se vuelve más complejo (es decir, con más características).
- Un λ correctamente ajustado dará lugar a un número razonable de variables con una interpretabilidad satisfactoria y un rendimiento de predicción superior.




β RSS (Least Square) Coefficients  Contours of RSS


β_R Ridge Coefficients

β_L LASSO Coefficients

β_E Elastic Net Coefficients

 Ridge Constrained Region defining penalty term

 LASSO constrained region defining penalty term

 Elastic Net constrained region defining penalty term

LASSO - Least Absolute Shrinkage and Selection Operator

Agradecimientos

- Diapositivas de ML/IA inspiradas en la clase MEDI504B de la Dra. Aline Talhouk