



UNIVERSIDAD DE CHILE

Minería de Datos

Welcome to the Machine Learning class

Valentin Barriere

Universidad de Chile – DCC

CC5205, Otoño 2024

Sesgos y Causalidades en la Modelización

Outline : Generalization and Approximation errors

Generalization and Approximation
errors

Correlación vs Causalidad

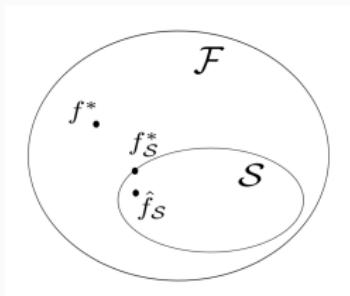
Sesgos

Trade-offs y Dilemas

En resumen

Complejidad y modelos: Objetivo Riesgo Minimum

- $\mathcal{F} = \{f : \text{funciones medibles } \mathcal{X} \rightarrow \mathcal{Y}\}$
- Mejor solución $f^* = \arg \min_{f \in \mathcal{F}} \mathcal{R}(f)$
- Clase de funciones $\mathcal{S} \subset \mathcal{F}$ utilizadas como modelos
- Objetivo ideal en \mathcal{S} : $f_S^* = \arg \min_{f \in \mathcal{S}} \mathcal{R}(f)$
- Estimación obtenida en \mathcal{S} : se obtiene \hat{f}_S tras un entrenamiento



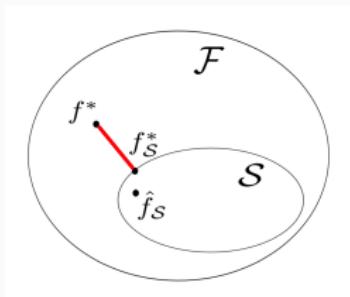
Error de aproximación y error de generalización

$$\mathcal{R}(\hat{f}_S) - \mathcal{R}(f^*) = \underbrace{\mathcal{R}(f_S^*) - \mathcal{R}(f^*)}_{\text{approximation error}} + \underbrace{\mathcal{R}(\hat{f}_S) - \mathcal{R}(f_S^*)}_{\text{estimation error}}$$

- El error de aproximación puede ser grande si el modelo \mathcal{S} no es adaptado
- El error de estimación puede ser grande si el modelo es complejo
- $\mathcal{R}(f^*)$ viene de la limitacion de los datos

Complejidad y modelos: Objetivo Riesgo Minimum

- $\mathcal{F} = \{f : \text{funciones medibles } \mathcal{X} \rightarrow \mathcal{Y}\}$
- Mejor solución $f^* = \arg \min_{f \in \mathcal{F}} \mathcal{R}(f)$
- Clase de funciones $\mathcal{S} \subset \mathcal{F}$ utilizadas como modelos
- Objetivo ideal en \mathcal{S} : $f_S^* = \arg \min_{f \in \mathcal{S}} \mathcal{R}(f)$
- Estimación obtenida en \mathcal{S} : se obtiene \hat{f}_S tras un entrenamiento



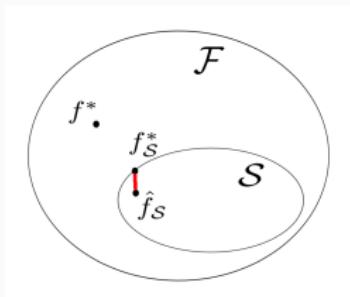
Error de aproximación y error de generalización

$$\mathcal{R}(\hat{f}_S) - \mathcal{R}(f^*) = \underbrace{\mathcal{R}(f_S^*) - \mathcal{R}(f^*)}_{\text{approximation error}} + \underbrace{\mathcal{R}(\hat{f}_S) - \mathcal{R}(f_S^*)}_{\text{estimation error}}$$

- El error de aproximación puede ser grande si el modelo \mathcal{S} no es adaptado
- El error de estimación puede ser grande si el modelo es complejo
- $\mathcal{R}(f^*)$ viene de la limitacion de los datos

Complejidad y modelos: Objetivo Riesgo Minimum

- $\mathcal{F} = \{f : \text{funciones medibles } \mathcal{X} \rightarrow \mathcal{Y}\}$
- Mejor solución $f^* = \arg \min_{f \in \mathcal{F}} \mathcal{R}(f)$
- Clase de funciones $\mathcal{S} \subset \mathcal{F}$ utilizadas como modelos
- Objetivo ideal en \mathcal{S} : $f_S^* = \arg \min_{f \in \mathcal{S}} \mathcal{R}(f)$
- Estimación obtenida en \mathcal{S} : se obtiene \hat{f}_S tras un entrenamiento



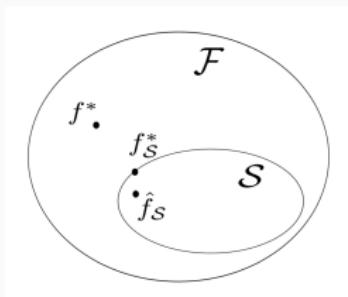
Error de aproximación y error de generalización

$$\mathcal{R}(\hat{f}_S) - \mathcal{R}(f^*) = \underbrace{\mathcal{R}(f_S^*) - \mathcal{R}(f^*)}_{\text{approximation error}} + \underbrace{\mathcal{R}(\hat{f}_S) - \mathcal{R}(f_S^*)}_{\text{estimation error}}$$

- El error de aproximación puede ser grande si el modelo \mathcal{S} no es adaptado
- El error de estimación puede ser grande si el modelo es complejo
- $\mathcal{R}(f^*)$ viene de la limitacion de los datos

Complejidad y modelos: Objetivo Riesgo Minimum

- $\mathcal{F} = \{f : \text{funciones medibles } \mathcal{X} \rightarrow \mathcal{Y}\}$
- Mejor solución $f^* = \arg \min_{f \in \mathcal{F}} \mathcal{R}(f)$
- Clase de funciones $\mathcal{S} \subset \mathcal{F}$ utilizadas como modelos
- Objetivo ideal en \mathcal{S} : $f_S^* = \arg \min_{f \in \mathcal{S}} \mathcal{R}(f)$
- Estimación obtenida en \mathcal{S} : se obtiene \hat{f}_S tras un entrenamiento

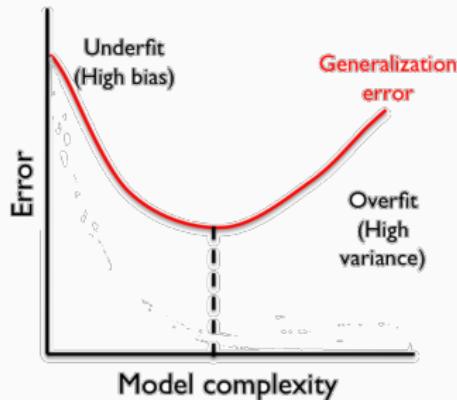


Error de aproximación y error de generalización

$$\mathcal{R}(\hat{f}_S) - \mathcal{R}(f^*) = \underbrace{\mathcal{R}(f_S^*) - \mathcal{R}(f^*)}_{\text{approximation error}} + \underbrace{\mathcal{R}(\hat{f}_S) - \mathcal{R}(f_S^*)}_{\text{estimation error}}$$

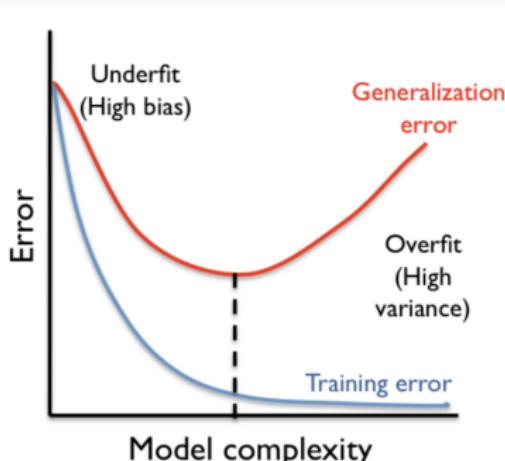
- El error de aproximación puede ser grande si el modelo \mathcal{S} no es adaptado
- El error de estimación puede ser grande si el modelo es complejo
- $\mathcal{R}(f^*)$ viene de la limitacion de los datos

Sobre-aprendizaje y sub-aprendizaje



- Según la complejidad del modelo (por ejemplo, tiempo de entrenamiento, cantidad de parámetros) se observa un comportamiento diferente
- Los modelos poco complejos son aprendidos fácilmente pero el error de aproximación puede ser grande (sub-aprendizaje)
- Los modelos muy complejos pueden tener el objetivo correcto pero un gran error de estimación (sobre-aprendizaje)

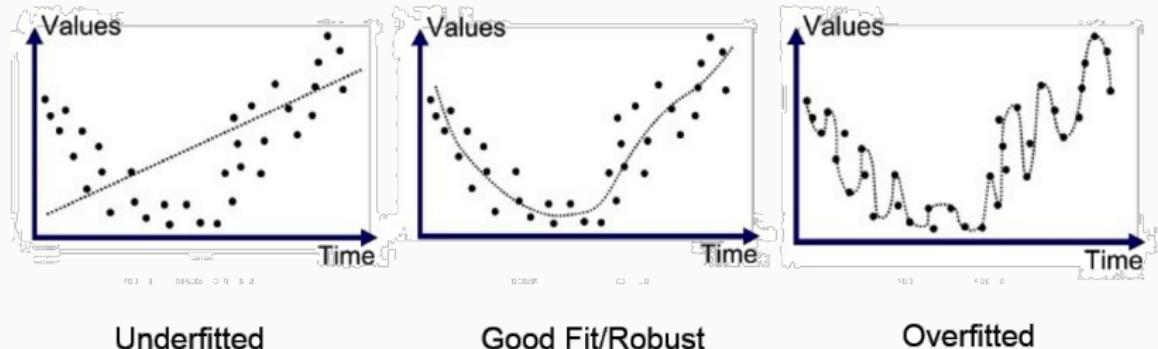
Sobre-aprendizaje: Problema



Error y riesgos

- El **riesgo empírico** (error en el conjunto de entrenamiento) disminuye con el aumento de la complejidad del modelo
- El **riesgo real** (error en observaciones de un nuevo conjunto) es muy diferente.
¡Tenemos un **problema de generalización!**
- Sobre-aprendizaje : los parámetros aprendidos son demasiado específicos para el conjunto de entrenamiento
- Se debe usar un criterio diferente al error en el conjunto de entrenamiento

Sobre-aprendizaje: Complejidad



Complejidad

- Si el modelo es demasiado simple, entonces ya no sigue los datos
- Si el modelo es demasiado complejo, el modelo aprende todas las irregularidades del conjunto de datos \mathcal{D}_n : el objetivo no es memorizar los datos de entrenamiento, sino generalizar los nuevos datos
- Ejemplo : si el modelo es el de la curva del medio más una componente de ruido no considerada en las variables, el modelo de la derecha aprende ese ruido

Caso simple

Supongamos que tenemos

- Un **polinomio de orden N** $X_N = \sum_i^N \alpha_i X^i$
- Quiero **predecir la variable Y** , compuesta por X_N y un ruido ε que no pudimos medir: $Y = X_N + \varepsilon$
- Mi **modelo es un polinomio de orden k** $\hat{Y} = X_k = \sum_i^k w_i X^i$

Con $k \leq N$, definiendo $Y_k = \sum_1^k \alpha_i X^i$ y $\varepsilon_{N-k} = \sum_{k+1}^N \alpha_i X^i$ obtenemos:

$$Y = Y_k + \varepsilon_{N-k} + \varepsilon$$

Es posible modelar Y usando \hat{Y} hasta cierto punto porque:

- ε_{N-k} no puede ser modelado debido a una **limitación del modelo**
- ε no puede ser modelado debido a una **limitación de los datos**

Caso simple

Supongamos que tenemos

- Un **polinomio de orden N** $X_N = \sum_i^N \alpha_i X^i$
- Quiero **predecir la variable Y** , compuesta por X_N y un ruido ε que no pudimos medir: $Y = X_N + \varepsilon$
- Mi **modelo es un polinomio de orden k** $\hat{Y} = X_k = \sum_i^k w_i X^i$

Con $k \geq N$, definiendo $\varepsilon_{N+} = \sum_{N+1}^k w_i X^i$ obtenemos:

$$\hat{Y} \sim X_N + \varepsilon_{N+}$$

Es posible modelar Y usando \hat{Y} pero aparecerán otros problemas:

- ε_{N+} se enfocará en intentar aprender ε , lo cual es imposible de predecir
- ε sigue sin poder modelarse debido a una **limitación de los datos**, creando el error inevitable $\mathcal{R}(f^*)$

Caso simple

Si tenemos un $k \ll N$

Nuestro modelo tendrá baja complejidad, y en este caso ε_{N-k} puede ser alto. El modelo está **subajustando** los datos y obtendremos un **error de aproximación**.

Si tenemos un $k = N$

Nuestro modelo tendrá la complejidad perfecta para capturar los datos, y en este caso ε_{N-k} puede ser cercano a cero. Aun así, el modelo no podrá modelar ε ya que su **información no está contenida en las observaciones**.

Si tenemos un $k \gg N$

Nuestro modelo será demasiado complejo para la tarea, también aprenderá a memorizar el ε que no puede ser predicho con las observaciones. Ha perdido su poder de generalización. El modelo está **sobreajustando** los datos y obtendremos un **error de generalización**.

¿Qué es este ε ? Información faltante en las observaciones

Veamos el ejemplo simple de un **modelo que predice el género de una persona usando el largo del cabello** (una variable booleana).

- El modelo usará las observaciones para alcanzar el objetivo
- Para minimizar el error, se apoyará en la **probabilidad a priori del largo del cabello respecto al género.**
- Las predicciones serán correctas para los grupos mayoritarios



ε no es ruido

ε tendrá un **valor distinto de cero para mujeres con cabello corto y hombres con cabello largo**, induciendo un **sesgo negativo** en el comportamiento del modelo respecto a estos dos grupos.

¿Qué pasa si $\varepsilon = 0$? Información faltante en las observaciones

En el caso perfecto donde toda la información del objetivo está contenida en las observaciones, $\varepsilon = 0$, sin embargo, surgen otros problemas:

- El modelo perfecto puede ser difícil de aprender: necesitaría más datos (por ejemplo, ejemplos difíciles) o más parámetros.
- Pero si el modelo tiene demasiadas variables de entrada y/o parámetros, se verá afectado por el ruido que esto introduce (cf. la maldición de la dimensionalidad)¹
- Especialmente cuando los datos están sesgados, el modelo se apoyará en heurísticas simples para predecir.

Durante el proceso de entrenamiento, los parámetros convergerán hacia lo que **más ayuda al modelo a predecir**, incluso si se basa en **correlaciones y no en causalidades**.

¹La maldición de la dimensionalidad ocurre al trabajar con datos de alta dimensión, generando mayor complejidad computacional, sobreajuste y correlaciones espurias

Supongamos que $\varepsilon = 0$

¡Aún podemos tener problemas!

Outline : Correlación vs Causalidad

Generalization and Approximation
errors

Correlación vs Causalidad

Sesgos

Trade-offs y Dilemas

En resumen

Poder Predictivo y Correlaciones Espurias

Optimizar una función de pérdida clásica nos dará un modelo con el mejor rendimiento posible, incluso si se basa en correlaciones y no en causalidades.

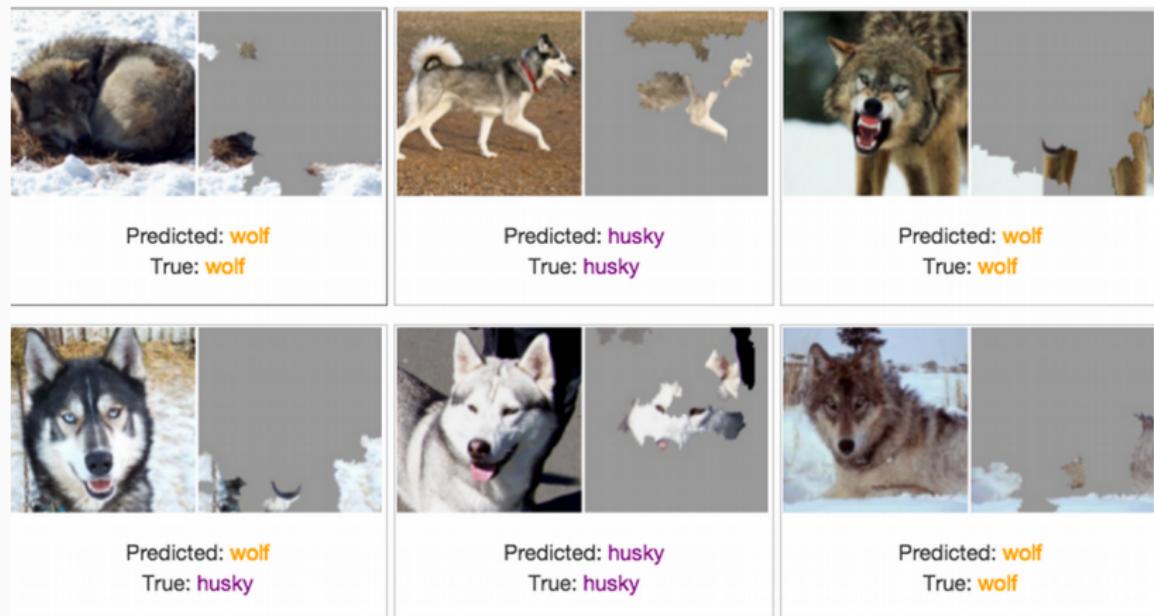
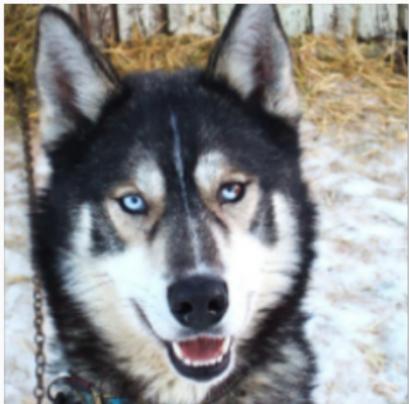
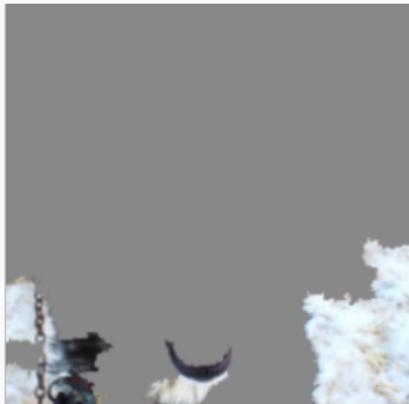


Figure 1: Explicación de las decisiones de un clasificador Husky/Lobo

Poder Predictivo y Correlaciones Espurias



(a) Husky classified as wolf



(b) Explanation

Figure 2: Explicación de las decisiones de un clasificador Husky/Lobo

¡El algoritmo comete errores que perjudican a los perros Husky que viven en lugares nevados! Pobrecitos :(

Correlaciones Espurias

The distance between Neptune and Earth

correlates with

Remaining Forest Cover in the Brazilian Amazon

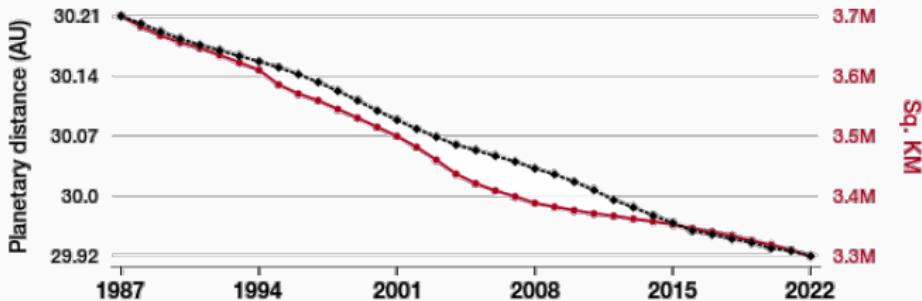


Figure 3: Un ejemplo de correlación espuria [2]

Correlaciones Espurias

The distance between Neptune and Earth
correlates with
Remaining Forest Cover in the Brazilian Amazon

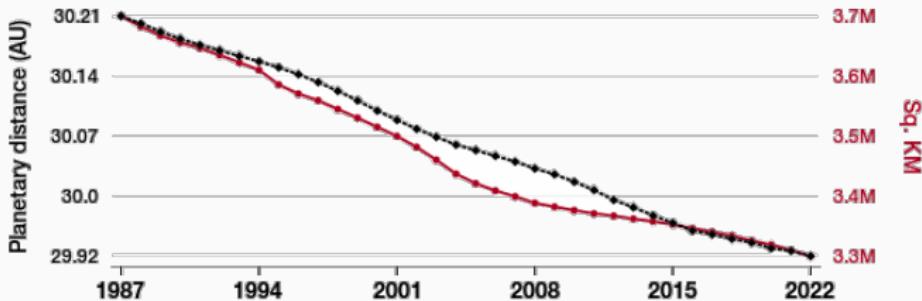


Figure 3: Un ejemplo de correlación espuria [2] e ilustraciones con IA



AI academic paper

(Because $p < 0.01$)

A Cosmic Dance: The Neptunian Distance and Amazonian Resilience

The Journal of Interplanetary Resilience Studies

Jan 2024

Reminder: This paper is AI-generated. Not real!

Las correlaciones espurias pueden ser fáciles de usar

Correlaciones Espurias

Relación matemática en la que dos o más eventos o **variables están asociadas pero no relacionadas causalmente**, ya sea por coincidencia o por la presencia de un tercer factor oculto (denominado "variable de confusión", "variable latente" o "variable de respuesta común").

- En ciertos contextos, **la información puede estar compartida** entre varias variables. Por ejemplo, en el conjunto de datos Husky/Lobo, la variable nieve comparte mucha información con la variable animal.
- Obviamente, esto es una **limitación de los datos**, ya que no es porque el animal esté junto a la nieve que sea un lobo.
- Tu modelo podría querer usar esta información para clasificar mejor porque **puede ser más fácil**.

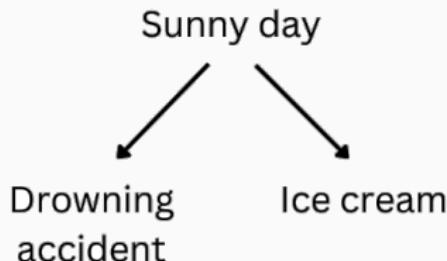
Las correlaciones espurias pueden ser fáciles de usar

Correlaciones Espurias

Relación matemática en la que dos o más eventos o **variables están asociadas pero no relacionadas causalmente**, ya sea por coincidencia o por la presencia de un tercer factor oculto (denominado "variable de confusión", "variable latente" o "variable de respuesta común").

- En ciertos contextos, **la información puede estar compartida** entre varias variables. Por ejemplo, en el conjunto de datos Husky/Lobo, la variable nieve comparte mucha información con la variable animal.
- Obviamente, esto es una **limitación de los datos**, ya que no es porque el animal esté junto a la nieve que sea un lobo.
- Tu modelo podría querer usar esta información para clasificar mejor porque **puede ser más fácil**.
- ¡Como cuando tu cerebro usa un estereotipo porque es más fácil que pensar de verdad! **Es en realidad un sesgo cognitivo humano.**

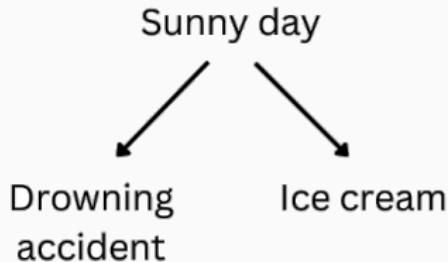
Correlaciones espurias o Variables de Confusión



Case 1: Spurious correlation

- Un día soleado impacta tanto en accidentes por ahogamiento como en ventas de helados
- Con suficientes variables y muestras, el modelo debería funcionar

Correlaciones espurias o Variables de Confusión



Case 1: Spurious correlation

- Un día soleado impacta tanto en accidentes por ahogamiento como en ventas de helados
- Con suficientes variables y muestras, el modelo debería funcionar
- ¿Qué pasa si no tenemos la variable clima?

Drowning Ice cream
accident

Case 1: Spurious correlation

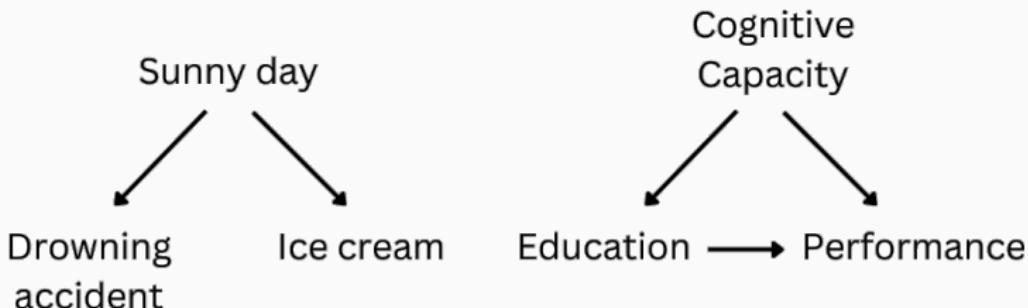
- Un día soleado impacta tanto en accidentes por ahogamiento como en ventas de helados
- Con suficientes variables y muestras, el modelo debería funcionar
- *¿Qué pasa si no tenemos la variable clima?*

Drowning → Ice cream
accident

Case 1: Spurious correlation

- Un día soleado impacta tanto en accidentes por ahogamiento como en ventas de helados
- Con suficientes variables y muestras, el modelo debería funcionar
- ¿Qué pasa si no tenemos la variable clima?

Correlaciones espurias o Variables de Confusión



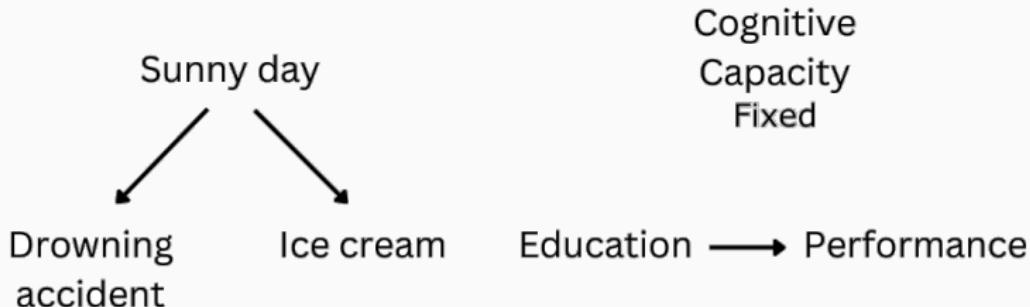
Case 1: Spurious correlation

- Un día soleado impacta tanto en accidentes por ahogamiento como en ventas de helados
- Con suficientes variables y muestras, el modelo debería funcionar
- ¿Qué pasa si no tenemos la variable clima?

Case 2: Confounder

- Los factores de confusión son causas comunes
- Es imposible medir el impacto de la educación sobre el desempeño laboral si una tercera variable influye en ambas

Correlaciones espurias o Variables de Confusión



Case 1: Spurious correlation

- Un día soleado impacta tanto en accidentes por ahogamiento como en ventas de helados
- Con suficientes variables y muestras, el modelo debería funcionar
- ¿Qué pasa si no tenemos la variable clima?

Case 2: Confounder

- Los factores de confusión son causas comunes
- Es imposible medir el impacto de la educación sobre el desempeño laboral si una tercera variable influye en ambas
- La solución es fijar una variable para estudiar las otras

Sesgos Cognitivos y Pensamiento Rápido/Lento [3]

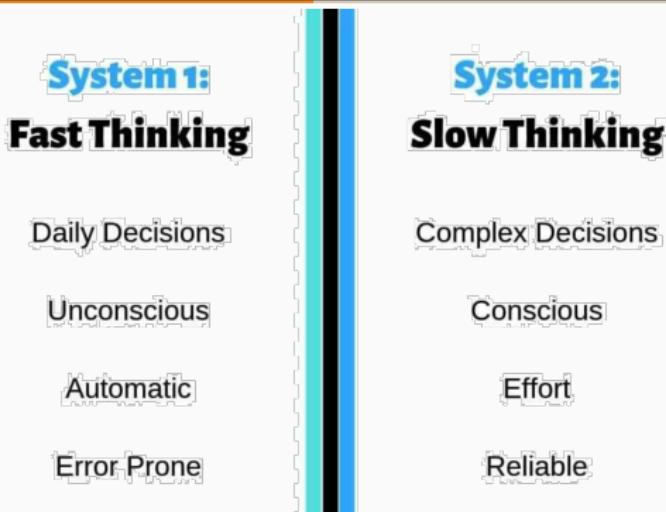


Figure 4: Libro ganador del Nobel "Pensar rápido, pensar despacio" de Daniel Kahneman

- **Sistema 1:** modo de pensamiento automático, rápido e intuitivo. Se basa en heurísticas (atajos mentales), juicios rápidos, estereotipos.
- **Sistema 2:** más lento, deliberativo y analítico. Se activa para resolver problemas complejos y razonar con evidencia.

Sesgos Cognitivos

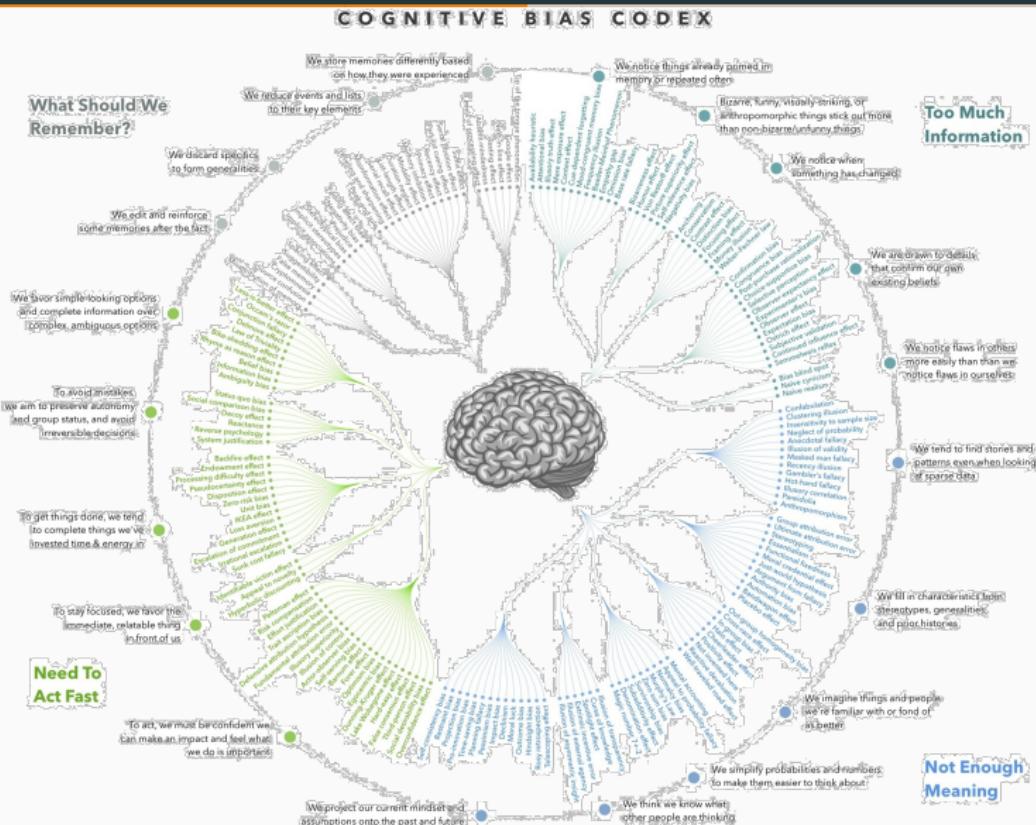


Figure 5: "Simplificar probabilidades y números para facilitar su comprensión" o "Encontrar patrones incluso en datos escasos"

Cerrando esta parte de analogías

- Así como los priors pueden estar sesgados en un conjunto de datos, los estereotipos surgen de información incompleta o generalizada
- Esto genera sesgos sistemáticos al procesar nueva información: "sesgo de confirmación" o prior sesgado
- **Los modelos de ML entrenados en datos sesgados pueden desarrollar priors sesgados** que llevan a predicciones injustas
- Igual que los humanos pueden desarrollar estereotipos sesgados

Cerrando esta parte de analogías

- Así como los priors pueden estar sesgados en un conjunto de datos, los estereotipos surgen de información incompleta o generalizada
- Esto genera sesgos sistemáticos al procesar nueva información: "sesgo de confirmación" o prior sesgado
- **Los modelos de ML entrenados en datos sesgados pueden desarrollar priors sesgados** que llevan a predicciones injustas
- Igual que los humanos pueden desarrollar estereotipos sesgados

En conclusión:

- Los datos pueden estar sesgados por correlaciones espurias debidas al azar o variables de confusión
- El modelo aprovechará ese sesgo
- Como lo haría un humano

Outline : Sesgos

Generalization and Approximation
errors

Correlación vs Causalidad

Sesgos

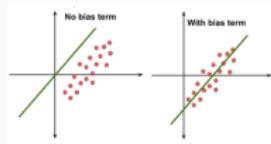
Trade-offs y Dilemas

En resumen

¿Qué pueden ser los sesgos?

Un sesgo puede ser una desviación respecto a la norma, al promedio o al valor cero:

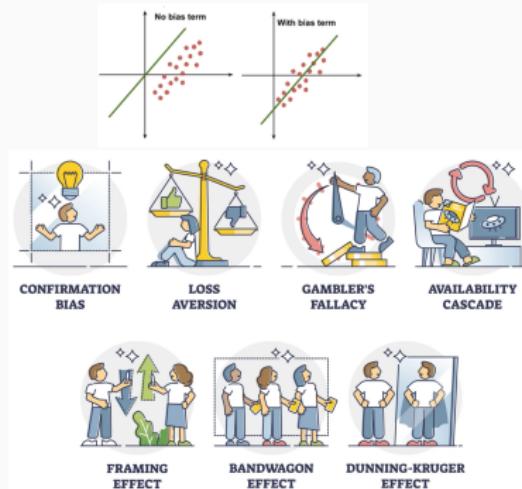
- Un sesgo en un modelo lineal para ajustar datos



¿Qué pueden ser los sesgos?

Un sesgo puede ser una desviación respecto a la norma, al promedio o al valor cero:

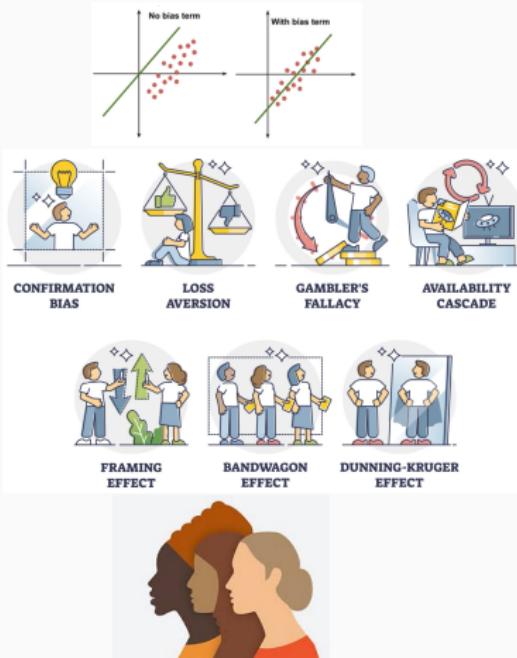
- Un sesgo en un modelo lineal para ajustar datos
- Sesgo cognitivo: sesgo de disponibilidad, sesgo de confirmación, efecto Dunning-Kruger, etc.



¿Qué pueden ser los sesgos?

Un sesgo puede ser una desviación respecto a la norma, al promedio o al valor cero:

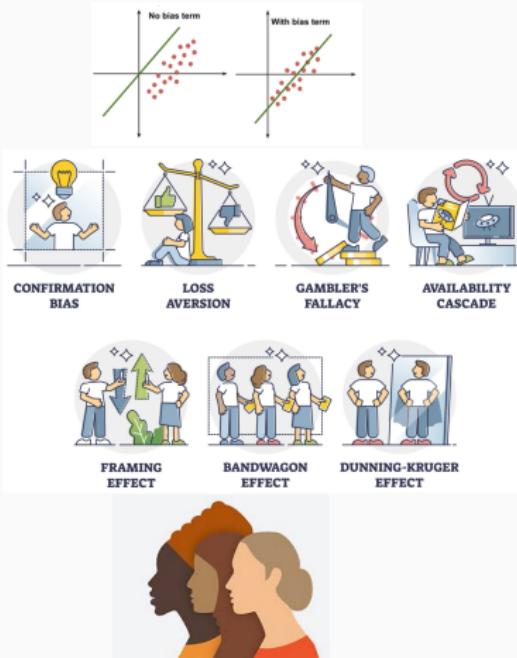
- Un sesgo en un modelo lineal para ajustar datos
- Sesgo cognitivo: sesgo de disponibilidad, sesgo de confirmación, efecto Dunning-Kruger, etc.
- Sesgo social, como sesgos culturales: las personas tienen diferentes normas sociales



¿Qué pueden ser los sesgos?

Un sesgo puede ser una desviación respecto a la norma, al promedio o al valor cero:

- Un sesgo en un modelo lineal para ajustar datos
- Sesgo cognitivo: sesgo de disponibilidad, sesgo de confirmación, efecto Dunning-Kruger, etc.
- Sesgo social, como sesgos culturales: las personas tienen diferentes normas sociales



En un proceso de toma de decisiones, un sesgo puede verse como un cambio en la acción causado por una variable no causal.

¿Qué pueden ser sesgos en modelos predictivos?

Definición

Una desviación de una norma definida o esperada. En IA, los modelos a menudo reflejan los sesgos presentes en los datos de entrenamiento, llevando a resultados sesgados.

Sesgos en aplicaciones de IA:

- **Aprobación de créditos:** modelos de IA han negado préstamos injustamente a mujeres o minorías por reflejar sesgos históricos en préstamos.
- **Sistemas de contratación:** algoritmos sesgados favorecen a hombres sobre mujeres igual de calificadas, perpetuando desigualdades de género.
- **Justicia penal:** herramientas como COMPAS han aumentado las tasas de encarcelamiento en minorías debido a evaluaciones sesgadas de riesgo.

¿Qué pueden ser sesgos en modelos predictivos?

Definición

Una desviación de una norma definida o esperada. En IA, los modelos a menudo reflejan los sesgos presentes en los datos de entrenamiento, llevando a resultados sesgados.

Sesgos en aplicaciones de IA:

- **Aprobación de créditos:** modelos de IA han negado préstamos injustamente a mujeres o minorías por reflejar sesgos históricos en préstamos.
- **Sistemas de contratación:** algoritmos sesgados favorecen a hombres sobre mujeres igual de calificadas, perpetuando desigualdades de género.
- **Justicia penal:** herramientas como COMPAS han aumentado las tasas de encarcelamiento en minorías debido a evaluaciones sesgadas de riesgo.

Los modelos de IA amplifican los sesgos existentes en la sociedad, reforzando desigualdades en lugar de mitigarlas.

Equidad

Generalmente, al hablar de modelos injustos, buscamos sesgos negativos hacia ciertos grupos objetivo.

Esto puede darse de distintas formas:

- Polarización hacia ciertos grupos: sentimiento más negativo hacia nombres árabes, predicciones de reincidencia más altas en personas negras, ...
- Desempeños heterogéneos entre grupos: sistemas de reconocimiento facial que fallan más en usuarios asiáticos, ASR que solo reconoce acentos específicos del español, ...

Sesgo del pasado para predecir el futuro

- Los modelos predictivos aprenden patrones del pasado para hacer predicciones
- Los datos reflejan el pasado, aun queremos predecir el futuro

Ejemplo

El modelo ha aprendido que las niñas tienen menos probabilidad de obtener empleos bien remunerados.



Sesgo del pasado para predecir el futuro

- Los modelos predictivos aprenden patrones del pasado para hacer predicciones
- Los datos reflejan el pasado, aun queremos predecir el futuro

Ejemplo

El modelo ha aprendido que las niñas tienen menos probabilidad de obtener empleos bien remunerados.



¿Queremos un modelo que se base en correlaciones espurias del pasado para predecir el futuro?

Spurious Correlations

- Predictive models are learning patterns over data to make predictions
- Predictive models are relying on whatever they can, and it can be spurious correlations at best, or even non-causal variables

Example

Let's take the extreme case of a 1 feature classifier that predicts whether or not someone will fail its exam based on the color of the skin.



Spurious Correlations

- Predictive models are learning patterns over data to make predictions
- Predictive models are relying on whatever they can, and it can be spurious correlations at best, or even non-causal variables

Example

Let's take the extreme case of a 1 feature classifier that predicts whether or not someone will fail its exam based on the color of the skin.



Should we use this classifier to filter out "probable best students" that should enter to the university in the future?

Outline : Trade-offs y Dilemas

Generalization and Approximation
errors

Correlación vs Causalidad

Sesgos

Trade-offs y Dilemas

En resumen

Correlaciones Espurias

- Modelos predictivos aprenden patrones de datos para hacer predicciones
- Pero esos modelos se apoyan en lo que puedan, incluso si son correlaciones espurias o variables no causales

Ejemplo

El caso extremo de un clasificador con una sola variable que predice si una persona reprobará un examen basándose en el color de piel.



Correlaciones Espurias

- Modelos predictivos aprenden patrones de datos para hacer predicciones
- Pero esos modelos se apoyan en lo que puedan, incluso si son correlaciones espurias o variables no causales

Ejemplo

El caso extremo de un clasificador con una sola variable que predice si una persona reprobará un examen basándose en el color de piel.



¿Deberíamos usar este clasificador para filtrar a los “mejores candidatos” que entrarán a la universidad?

Compromisos (Trade-offs)

Estas variables no causales:

- Pueden tener un gran poder predictivo, lo que las hace **muy útiles para el rendimiento**
- Sin embargo, estos modelos no son robustos frente a datos nuevos porque se apoyan en patrones espurios. Están sobreajustando datos pasados que contienen sesgos.

Por ejemplo:

Compromisos (Trade-offs)

Estas variables no causales:

- Pueden tener un gran poder predictivo, lo que las hace **muy útiles para el rendimiento**
- Sin embargo, estos modelos no son robustos frente a datos nuevos porque se apoyan en patrones espurios. Están sobreajustando datos pasados que contienen sesgos.

Por ejemplo:

- Si un modelo predice el salario de una persona, asignará menor salario a las mujeres. **El modelo es más preciso si se apoya en este sesgo.**
- Pero si la sociedad cambia y deja de ser misógina, las mujeres ganarán lo mismo. Entonces el modelo no funcionará bien. **No es robusto a cambios en los datos.**

Compromisos funcionales básicos en Cognición

- **Desempeño:** Capacidad de alcanzar un resultado deseado
- **Eficiencia:** Hacer más con menos recursos
- **Robustez:** Mantener desempeño frente a perturbaciones
- **Flexibilidad:** Adaptarse a distintas situaciones sin fallar

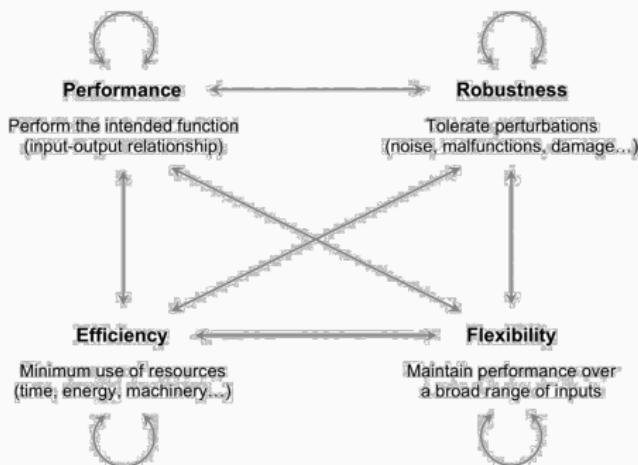


Figure 6: Compromisos funcionales [1]

Efectos de "Menos es Más"

Más rendimiento pero menos robustez

Un modelo con alto poder predictivo suele ser menos robusto y menos adaptable a nuevas situaciones donde la causalidad se expresa de forma distinta

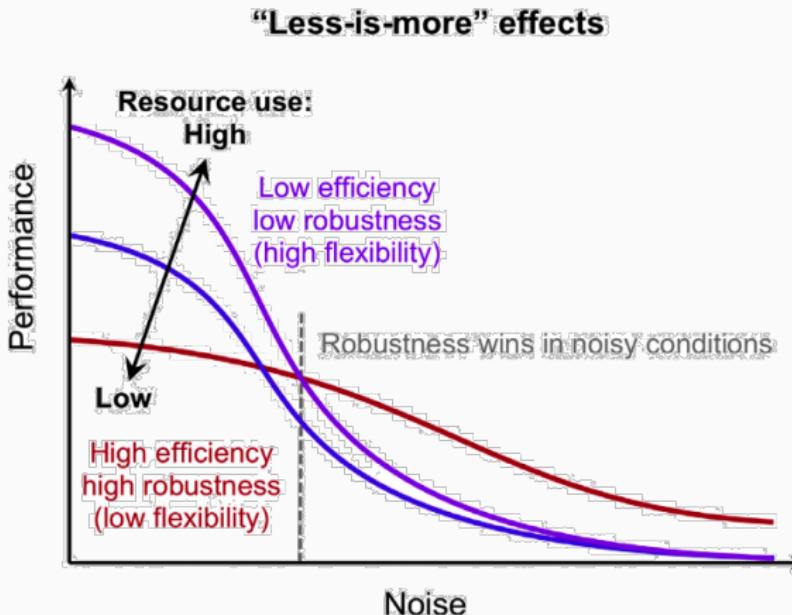


Figure 7: Los modelos con menos variables son más robustos [1]

Efectos de "Menos es Más"

Más rendimiento pero menos robustez

Un modelo con alto poder predictivo suele ser menos robusto y menos adaptable a nuevas situaciones donde la causalidad se expresa de forma distinta ⇒ **Si el futuro no refleja el pasado, el modelo falla**

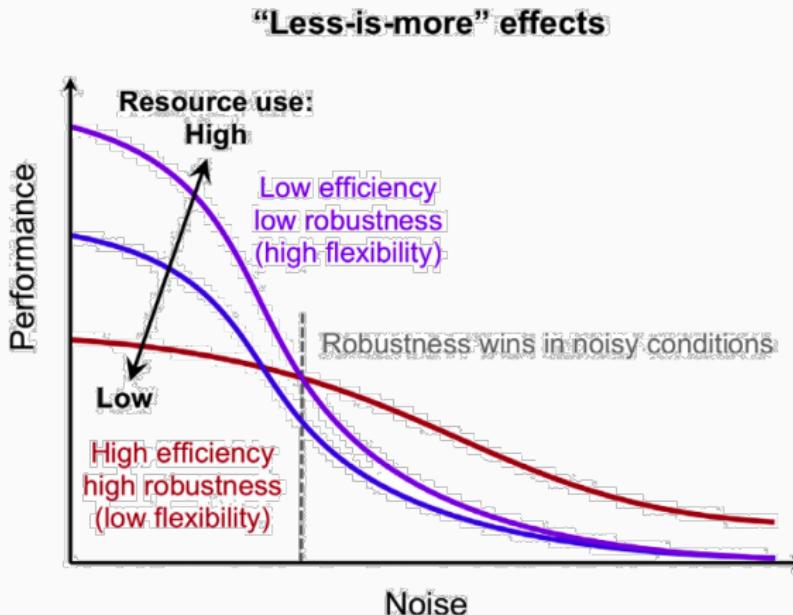


Figure 7: Los modelos con menos variables son más robustos [1]

Outline : En resumen

Generalization and Approximation
errors

Correlación vs Causalidad

Sesgos

Trade-offs y Dilemas

En resumen

Técnicas para tratar con datos sesgados

Si tienes un conjunto de datos con variables causales (CV), no causales (NCV) y atributos sensibles (SA), puedes:

- **Diferencia en pesos:** Entrenar modelos con CV y/o NCV y analizar las diferencias
- **Usar solo CV:** Intentar obtener el mejor modelo solo con variables causales
- **Diferencias de rendimiento:** Probar con CV y/o NCV
- **Fijar NCV:** Una por una y observar las diferencias en rendimiento y pesos
- **Evaluar rendimiento por grupo:** Separar el set de test por SA y analizar las diferencias

Dos visiones para una misma tarea

ML predictivo sobre la PSU (SAT)

Estoy creando un modelo para predecir el puntaje PSU de un estudiante. ¿Qué hago con ese modelo?

Caso I. Lo uso para etiquetar otros estudiantes fuera de los datos, suponiendo que la distribución es igual, y que los estudiantes más pobres reprobarán el examen este año, porque es la realidad. ¡Sé que el modelo se basa en razones erradas pero da buenos resultados!

Caso II. Lo uso para predecir el futuro, suponiendo que la próxima generación aún reproducirá esas desigualdades

Caso I. tiene un enfoque estadístico: Uso una muestra de la población para estimar el todo

Caso II. tiene un enfoque de ML: Trato de predecir lo que va a pasar

Ejemplo sobre un Candidato Laboral

Candidato

Supón que tienes que elegir entre dos candidatos para un puesto en tu empresa. Ambos tienen logros profesionales similares, pero uno no tiene título universitario. ¿A quién eliges?

- Un **modelo sesgado** sabrá que tener un título universitario tiene mayor probabilidad y probablemente elija a esa persona.
- Un **modelo no sesgado** elegirá a quien no tiene el título porque logró lo mismo con desventajas, lo que podría indicar mayor potencial.

En resumen

Correlaciones espurias vs. Modelos causales

- Correlaciones espurias: Mejoran precisión pero no son robustas ante cambios reales en los datos
- Modelos causales: Más interpretables y generalizables, aunque podrían reducir la precisión inicial

Consideraciones éticas

- Excluir variables sensibles puede reducir sesgos, pero disminuir precisión
- Incluirlas puede perpetuar desigualdades y dañar a grupos vulnerables (como los Huskies con nieve o las chicas con pelo corto)

Detección de sesgos

- Existen múltiples técnicas para detectar sesgos
- También se pueden usar enfoques de ML causal para modelos más confiables

Recursos

Contenido en línea:

- Un [curso interesante](#) sobre Machine Learning Causal
- Un buen [artículo de blog](#) sobre la diferencia entre correlación y causalidad

Clases:

- [Clase CC5117: Algoritmos, Redes y Equidad: Análisis de Sistemas Sociotécnicos](#)
- [Clase CS109 de Stanford: Probabilidad para Científicos de la Computación](#)
- [Clase CC6104: Pensamiento Estadístico](#)

Questions?

References i

-  M. Del Giudice and B. J. Crespi.
Basic functional trade-offs in cognition: An integrative framework.
Cognition, 179(June):56–70, 2018.
-  T. Given.
Spurious correlation #4,242, 2024.
-  D. Kahneman.
Thinking, Fast and Slow.
2011.