Aprendizaje Estadístico Supervisado

Natalia da Silva

2024

Notación

- n número de observaciones
- p número de variables
- Una matriz de datos:

$$\mathbf{X}_{nxp} = (\mathbf{x}_1, \mathbf{x}_2 \ \dots \mathbf{x}_p) = \begin{pmatrix} x_{11} & x_{12} & \dots x_{1p} \\ x_{21} & x_{22} & \dots x_{2p} \\ \vdots & \vdots & \ddots \quad \vdots \\ x_{n1} & x_{n2} & \dots x_{np} \end{pmatrix}$$

Notación

- las filas de $\mathbf{X}$, las escribiremos como $x_1, x_2, \ldots x_n$
- donde x_i es un vector de longitud p, con p variables medidas para la i-ésima observación.

  la i-ésima observación se denota como:

$$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}$$

Notación

- las columnas de $\mathbf{X}$, las escribiremos como $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_p$

$$\mathbf{X} = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix}$$

- donde cada $\mathbf{x}_j$ es un vector de longitud n

  La j-ésima variable se denota como:

$$\mathbf{x}_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix}$$

Notación

y_i es la i-ésima observación de la variable que queremos predecir..

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

También utilizaremos Y para representar la variable de respuesta.

Los datos observados $D = \{(x_i, y_i)\}^n = \{(x_1, y_1), (x_2, y_2), \dots (x_n, y_n)\}$ con x_i un vector de longitud p.

Aprendizaje Estadístico

Variable de interés Y y un conjunto de p predictores diferentes $X = (x_1, x_2, \dots x_p)$.
Sea $X \in R^p$, Y puede ser numérica o categórica dependiendo el problema y $Pr(Y, X)$ la distribución de probabilidad conjunta.
Nos interesa estudiar la relación entre Y y $\mathbf{X} = (x_1, x_2, \dots x_p)$,
$Y = f(\mathbf{X}) + \epsilon$

- Componente Determinístico $(Y, f(X))$: Describe comportamiento medio
- Componente aleatorio (ε): Describe desviaciones del comportamiento medio

Aprendizaje Estadístico

El aprendizaje estadístico refiere a un conjunto de aproximaciones para estimar $f(X)$

$$Y = f(\mathbf{X}) + \epsilon$$

Tipos de Aprendizaje Estadístico

El aprendizaje estadístico da un marco para construir modelos a partir de datos.

.

Tipos de Aprendizaje Estadístico

Nos interesa estudiar la relación entre Y y $\mathbf{X}$,
$\mathrm{Y} = \mathrm{f}(\mathrm{X}) + \epsilon$
Tipos de problemas:

- Supervisado, la variable de respuesta y_i disponible para todas los x_i Problemas de regresión (y_i es numérica) o clasificación (y_i es categórica)
- No supervisado, y_i no está disponible para ningún x_i
- Semi supervisado, y_i disponible para algunas x_i

Importante identificar el tipo de problema de aprendizaje nos enfrentamos para identificar posibles métodos.

Ejemplo supervisado - regresión

Valuación de bosques:

- Respuesta: Valor económico de bosques
- Predictores: variables goegráficas y demográficas.
- Datos: Meta-análisis e imágenes satelitales

(Siikamäki et al. 2022)

Ejemplo supervisado - clasificación

(Tancredi 2024)

Uso de Plataformas y desempeño académico

- Respuesta: Alcanza nivel de Inglés
- Predictores: Uso de Little Bridge
- Datos: Escolares de 6to, en 2021.

Tipos de Aprendizaje Estadístico

Otros tipos de aprendizaje estadístico…

- Por refuerzo: se quiere aprender como actuar o comportarse basado en premios y castigos pero solo aprende con las recompensas.
- Profundo: basados en el aprendizaje de múltiples niveles de características o representaciones de datos, ejemplo redes neuronales con muchas capas.
- Transferencia: aprendo con un conjunto de datos y uso el modelo para predecir en otro contexto.

Este curso: aprendizaje supervisado, algunos métodos.

¿Porqué estimamos f?

Predicción

- En muchas situaciones $\mathbf{X}$ puede estar disponible, pero Y puede ser difícil de recolectar.
- Entonces nos gustaría usar $\mathbf{X}$ para predecir un nuevo valor de Y.
- No estamos muy preocupados si f es difícil de entender solo que haga un buen trabajo para predecir nuevos valores de Y.
  $\hat{Y} = \hat{f}(\mathbf{X})$
- $\hat{Y}$ predicción para Y.
- $\hat{f}(X)$ estimador de f.

  En este contexto usualmente no importa la forma de f (black box), sino la precisión de las predicciones del modelo.

¿Porqué estimamos f?

Inferencia

Muchas veces estamos interesados en entender la asociación entre Y y X, en este contexto el objetivo es estimar f pero no necesariamente obtener predicciones de Y.

En este caso $\hat{f}$ no puede ser una caja negra ya que queremos responder preguntas como:

- ¿Qué predictores están asociados con la variable de respuesta?
- ¿Cuál es la relación entre la variable de respuesta y cada predictor?
- ¿La relación de la variable de respuesta con cada predictor puede ser adecuadamente resumida con una relación lineal o la relación es más complicada?

Las dos culturas ...

- Breiman, L., (2001) Statistical modeling: The two cultures. *Statistical science.*
- Efron B., (2020) Prediction, Estimation, and Attribution. *Journal of the American Statistical Association*
- Shmueli, G., (2010). To explain or to predict? *Statistical science.*

Estadística *clásica*

¿Cómo desarrollamos métodos en estadística *clásica*?

- Modelo parámetrico, los parámetros se conectan con el problema científico.
- Propiedades estadísticas de estimadores
- Estimadores insesgados, y buscar la mínima varianza
- Teoría asintótica

Partimos de preguntas de interés, las traducimos en términos de modelos estadísticos, buscamos respuesta estadística con datos observados, y volvemos a responder la pregunta inicial.

Aprendizaje supervisado

- Se propone un *algoritmo* o modelo no-paramétrico.
- Selección *sistemática* de modelo guiada por datos.
- Error de predicción en datos *nuevos*.
- Minimizar pérdida esperada.
- Performance en simulaciones, datos reales simples/conocidos, y datos reales complejos.
- Disponibilidad de la implementación.

Explicar o predecir ...

Muchas veces se plantea que:

- La inferencia estadística, solo buscan *explicar/ entender*.
- Los algoritmos de Machine Learning solo buscan *predecir*.

Explicar o predecir es la cuestión. Hay que hacer todo!

Explicar o predecir ...

En el ejemplo de riqueza en bosques de los países:

- Entender determinantes del valor de bosques
- Predecir valor de bosques en todo el mundo
- Comparar modelos de distinto tipo de manera sistemática

Explicar o predecir ...

En el ejemplo de Ceibal en Inglés

- Entender factores que incrementen la probabilidad de alcanzar el nivel de Inglés esperado
- Predecir si un estudiante alcanza el nivel de inglés con datos hasta julio

Explicar o predecir ...

Si solo miramos performance predictiva:

- No consideramos posibles sesgos en los datos **NPL sesgado**.
- No aprendemos de los datos: **The machines learn but we don't**

Solo mirar la significación estadística:

- Un modelo que predice muy mal fuera de la muestra, ¿puede dar explicaciones válidas y generalizables?
- Problemas con NHST: **discusión sobre el p-value**

"… algorithms are what statisticians do while inference says why they do them." (Efron, B., Hastie, T. (2016))

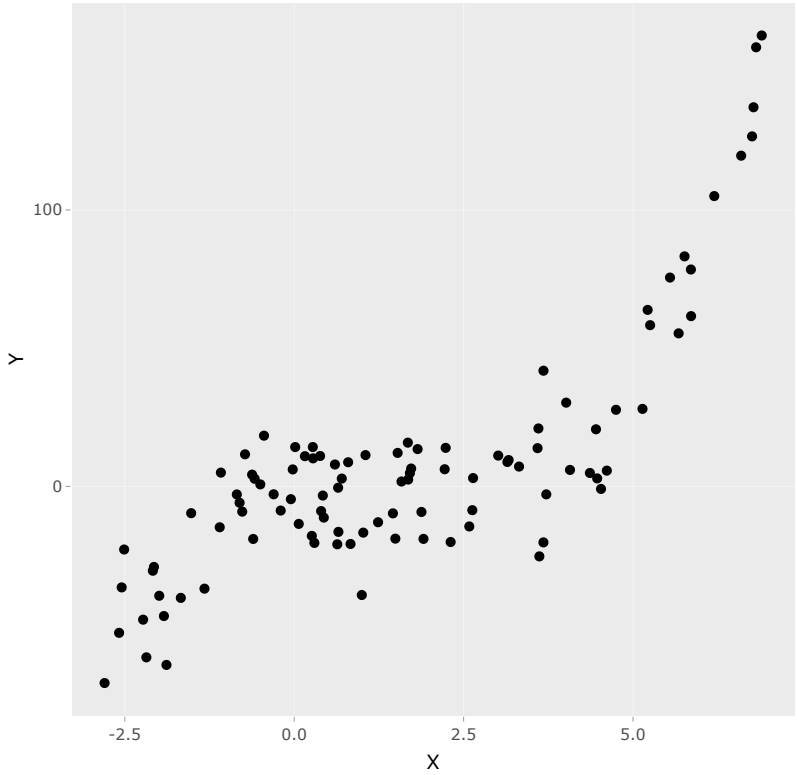
Modelo predictivo

- Asumimos una relación entre $\mathbf{X}$ e Y
- $Y = f(\mathbf{X}) + \epsilon$
- donde $f(\mathbf{X})$ es una función fija y desconocida y ϵ independiente de $\mathbf{X}$ y con media 0.
- f representa la información sistemática de $\mathbf{X}$ sobre Y

  Los datos son simulados de una f conocida.
- Simulamos 100 observaciones de $x \sim U(-3, 7)$
- Simulamos 100 observaciones de $Y = 3 + 2x - 4x^2 + x^3 + \epsilon$

Datos Simulados

Los datos son simulados de una f conocida.



Precisión en la estimación

Cuán preciso es $\hat{Y}$ para predecir Y depende de:

- **Error reducible**: error de modelización al elegir $\hat{f}$ como estimador de f
- **Error irreducible** : error aleatorio, no controlable ϵ.

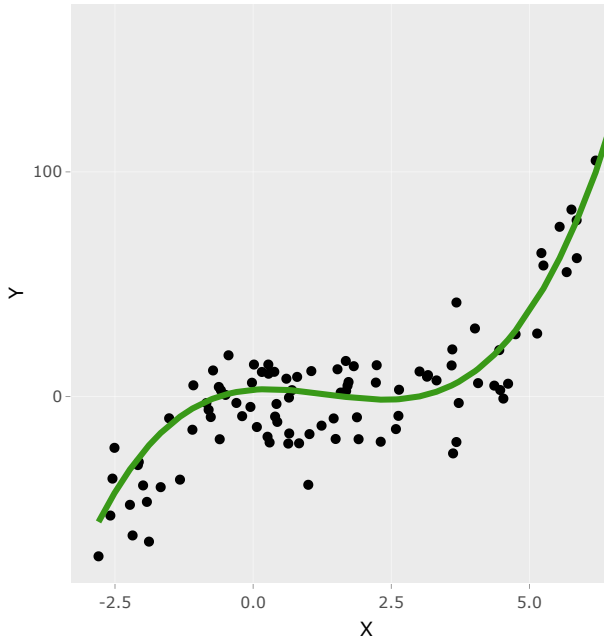
  Error irreducible puede ser mayor a cero porque ϵ puede contener variables no medidas que son útiles para la predicción de Y pero como no las medimos f no las puede usar en la predicción. También podría contener variabilidad no medida.

Precisión en la estimación

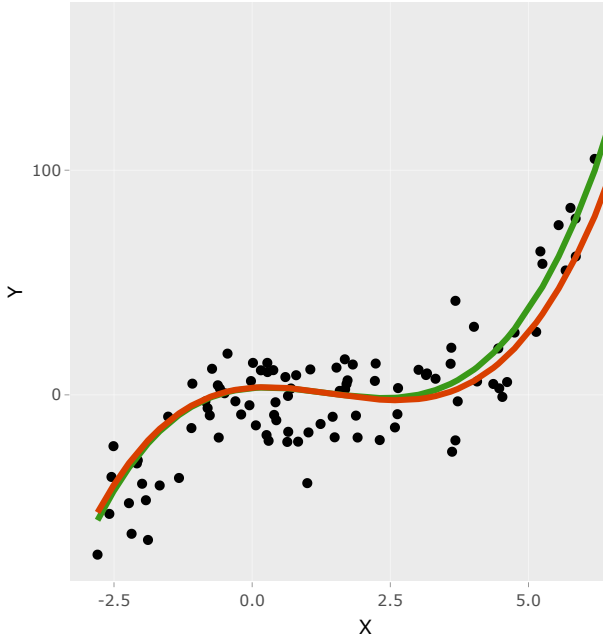
Asumiendo que $\hat{f}$ y $\mathbf{X}$ son fijos entonces la única variabilidad está dada por ϵ. Se puede probar que:

$$\begin{aligned} \mathbf{E}(Y - \hat{Y})^2 &= \mathbf{E}(f(\mathbf{X}) + \varepsilon - \hat{f}(\mathbf{X}))^2 \\ &= \underbrace{\mathbf{E}(f(\mathbf{X}) - \hat{f}(\mathbf{X}))^2}_{\text{reducible}} + \underbrace{\mathrm{Var}(\varepsilon)}_{\text{irreducible}} \end{aligned}$$

$\mathbf{E}(Y - \hat{Y})^2$ es el valor esperado de la diferencia al cuadrado del valor predicho y el valor verdadero de la respuesta. $\mathrm{Var}(\varepsilon)$ es la varianza del error.

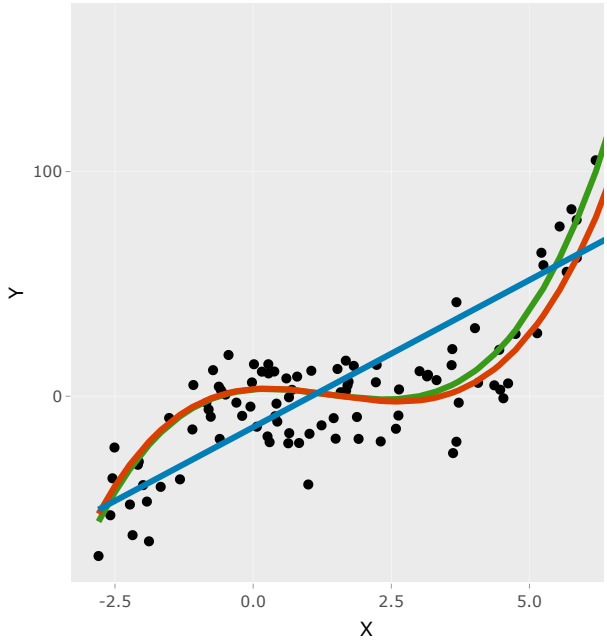


La línea verde representa la **verdadera** f. Esto es lo mejor que podemos obtener y todo el error que queda es irreducible



La línea roja representa un modelo estimado $\hat{f}$. Este ajuste es muy similar al verdero f pero aún se puede mejorar.

Suponé que ahora usamos un modelo más sencillo, modelo lineal



- La línea azul indica un modelo estimado más simple $\hat{f}$. Hay mucho espacio para mejorar en ese caso en la estimación.
- Estudiaremos técnicas para estimar f de forma tal que se minimice el error reducible.

Recapitulando

- El error irreducible es el que podemos mejorar produciendo el mejor modelo.
- En el error irreducible asociado a fluctuaciones aleatorias de muestra a muestra no sistemáticas.
- El objetivo es obtener predicciones del modelo que sean precisas para datos futuros.

¿Cómo estimamos f?

- **Métodos paramétricos:**
- Asume que el modelo tiene una forma específica.
- Ajustar el modelo implica estimar los parámetros del mismo.
- En general se considera poco flexible.
- Si los supuestos no se cumplen esperable que tengan un mal desempeño
- **Métodos no paramétricos:**
- No hay supuestos específicos.
- Permite que los datos especifiquen la forma del modelo sin ser muy irregular.
- Más flexible.
- En general son necesarias más observaciones.

Modelos paramétricos

$$f(\mathbf{X}) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

Menos flexible

$$f(\mathbf{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \ldots$$

Más flexible

No paramétricos

Ejemplo: Regresión polinómica local ajusta un modelo lineal a muchos subconjuntos de datos.

Compromiso entre precisión e interpretabilidad

- **Modelos lineales** Son modelos rígidos que resultan muy buenos para interpretar resultados pero en general no suelen ser buenos para hacer predicciones.
- **Modelos no lineales** (que son más flexibles) son complejos de interpretar, pero en general resultan ser buenos predictores, aunque debe tenerse cuidado con el sobreajuste (overfitting).
- Ejemplos modelos no lineales: modelos aditivos generalizados, árboles de regresión y clasificación, redes neuronales, Bagging, Boosting.

  Importante: no hay una técnica mejor que otra per se, sino técnicas que resultan más apropiadas que otras dependiendo del problema a resolver.

Flexibilidad vs Interpreteabilidad

Hay un compromiso entre la flexibilidad e interpretabilidad de un modelo.

Siikamäki, Juha, Matias Piaggio, Natalia da Silva, and Ignacio Alvarez. 2022. “Global Assessment of Non-Wood Forest Ecosystem Services: A Revision of a Spatially Explicit Meta-Analysis and Benefit Transfer.” Washington, DC: The World Bank. https://documents1.worldbank.org/curated/en/099850110202253173/pdf/P17727806732bb09d0a3 _gl=1*1ksx3ya*_gcl_au*MTk0MzIzNTI0LjE3MjM0NzcwNjM.

Tancredi, Bruno. 2024. “Aprendizaje Estadístico Aplicado Para Potenciar La Enseñanza de Inglés En Primaria: El Caso de Ceibal En Inglés En Uruguay.” Tesis de grado. Universidad de la República (Uruguay). https://hdl.handle.net/20.500.12008/44455.