Introduccion

Matias Bajac

2024-09-20

##Notacion##

n numero de observaciones **p** numero de variables Una matriz de datos:

$$\mathbf{X}_{n\times p} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

Las filas de X la escribiremos como $x_1, x_2, ..., x_n$

Donde x_i es un vector de longitud p, con p variables medidas para la i - esima observacion.

La i - esima observacion se denota como:

$$\mathbf{x}_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{bmatrix}$$

Las columnas de X, la escribiremos como $x_1, x_2, ..., x_p$

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_p^\top \end{bmatrix}$$

donde cada x_j es un vector de longitud n

la j - esima variable se denota como:

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

y_i es la i - esima observacion de la variable que queremos **predecir**

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

Tambien usaremos Y para representar la variable de **respuesta** o la variable **dependiente**. Los datos observados

$$D = \{(x_i, y_i)\}^n = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$$

con un x_i un vector de longitud p

Aprendizaje estadistico

Variable de interes Y y un conjunto de p predictores diferentes

$$X = (x_1, x_2, ..., x_p)$$

Sea $X \in R^p$, Y puede ser numerica o categorica dependiendo del problema y Pr(X,Y) la distribucion de probabilidad conjunta.

Nos interesa estudiar la relacion entre Y y $X = (x_1, x_2, ..., x_p)$

$Y = f(X) + \epsilon$

Componente deterministico : (Y,f(X)): Describe el comportamiento medio **Componente aleatorio** (ϵ): Describe desviaciones del comportamiento medio

El aprendizaje estadistico refiere un conjunto de aproximaciones para estimar

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

Tipos de aprendizaje estadistico

Nos interesa estudiar la relacion entre Y y X, $Y = f(x) + \epsilon$ Tipos de problemas:

Supervisado, la variable de respuesta y_i **disponible** para todas las x_i. Problemas de regresion si y_i es numerica, o clasificacion si y_i es categorica

No supervisado, y_i **no** esta disponbile para ninguna x_i

Semi-supervisado y_i disponible para algunas x_i

Importante identificar el tipo de problema de aprendizaje nos estamos enfrentando para identificar posibles metodos.

Por que estimamos f?

Prediccion

En muchas ocasiones **X** puede estar disponible, pero **Y** puede ser dificil de recolectar.

Entonces nos gustaria usar **X** para predecir un nuevo valor de **Y**

No estamos muy preocupados si f es dificil de entender solo que haga un buen trabajo para predecir nuevos valores de **Y**

$\hat{Y} = \hat{f}(X)$

$\hat{Y}$ prediccion para **Y**

$\hat{f}(X)$ estimador de f

En este contezxto usualmente no importa la forma de f **(black box)**, sino laa precision de las predicciones del modelo.

Inferencia

Muchas veces estamos interesados e entender la asociacion (por ejemplo lineal) entre Y y X, en este contexto el objetivo es estimar f pero no necesariamente predicciones de Y.

En este caso $\hat{f}$ no puede ser **black box** ya que queremos responder a preguntas como:

Que predictor/es estan asociados con la variable de respuesta?

Cual es la relacion entre la variable de respuesta y cada predictor?

La relacion de la variable de respuesta con cada predictor puede ser adecuadamente resumida con una relacion lineal o la relacion es mas complicada?

Aprendizaje supervisado

- Se propone un algoritmo o modelo no - parametrico.
- Seleccion sistematica de modelo guiada por datos.
- **Error de prediccion** en datos nuevos
- Minimizar la **perdida esperada**
- Performance en simulaciones, datos reales simples/conocidos y datos reales complejos.
- Disponibilidad de la implementacion

Explicar o predecir

Muchas veces se plantea que:

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

Explicar o predecir es la cuestion. Hay que hacer todo

En el ejemplo de la riqueza en los bosques de paises

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

En el ejemplo de **Ceibal** en ingles

- Entender factores que incrementan la probabilidad de alcanzar el nivel de ingles esperado

Predecir si un estudiante alcanza el nivel de ingles con datos hasta julio.

Si solo miramos la performance **predictiva**

-No consideramos los posibles **Sesgos** en los datos **NPL sesgado**

- No aprendemos de los datos: **The machines learn but we do not**

Por otro laado, solo mirar la **significacion** estadistica (interpretabildiad)

- Puede suceder que el modelo **prediga** muy mal fuera de la muestra, se puede dar explicaciones vailidas y generalizables?
- Discusiones sobre p -value

Modelo predicitivo

- Asumimos una relacion entre $\mathbf{X}$ y $\mathbf{Y}$
- $\mathrm{Y} = f(\mathbf{X}) + \epsilon$
- donde $\mathbf{X}$ es una funcion fija y desconocida y ϵ independiente de $\mathbf{X}$ y con media 0

- f representa la info sistematica de $\mathbf{X}$ sobre Y

ejemplo:

Los datos son simulados sobre una f conocida

- simulamos 100 observaciones de $x \sim U(-3, 7)$
- Simulamos 100 obs de un polinomio de grado tres mas el componente aleatorio

Precision en la estimacion

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

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

El error irreducible puede ser mayor a 0 porque ϵ puede contener variables no medidas que son utiles para predecir Y pero como no la medimos f no las puede usar en la prediccion. Tambien podria contener variabilidad no medida.

Asumiendo que $\hat{f}$ y $\mathbf{X}$ son fijos, entonces la unica variabilidad esta dada por ϵ. se puede probar que:

(asumimos una distribucion para Y)

$$E(Y - \hat{Y})^2 \ = E(f(\mathbf{X}) \ + \epsilon \ - \hat{f}(\mathbf{X}))^2 =$$

$$E(f(x) + \epsilon - \hat{f}(\mathbf{X})) \ (f(x) + \epsilon - \hat{f}(\mathbf{X})) =$$

todo lo que no dependa del comoponente aleatorio va hacia un lado ya que es fijo , esto seria f(x) y su estimacion

Aplico distributiva

$$E[(f(x) - \hat{f}(x))^2 \ + \ \epsilon^2 + f(x)\epsilon + f(x)\epsilon - \hat{f}(x)\epsilon - f\hat{(}x)\epsilon] =$$

reagrupamos

$E[(f(x) - \hat{f}(x)^2) - 2(-f(x) + \hat{f}(x))\epsilon + E(\epsilon^2)]$

$$E(f(x) - \hat{f}(x))^2 + Var(\epsilon)$$