Estadistica Multivariada Computacional

Estadistica Multivariada Computacional 2019, basado en las clases de Mathias Bourel (IMERL).

Índice

```
0
    .1
    .2
    .3
1
    .1
    .2
2
```

0 Tipos de aprendizaje automatizado

- Aprendizaje supervisado
- Aprendizaje no supervizado
- *Aprendizaje por refuerzo*

0.1 Aprendizaje supervisado

0.2 Aprendizaje no supervizado

0.3 Aprendizaje por refuerzo

1 Bases de datos

Las bases de datos también conocidas como *datasets*, FILL THIS HERE. Podemos separarlas en dos tipos, bases de datos con etiqueta (o *label*) y sin etiquetar.

1.1 Bases de datos con etiqueta

Las bases de datos con etiqueta son utilizadas para el aprendizaje supervisado.

a_{1}	a_{2}	a_{3}	$\cdots$	a_{m}	y
		x_{1}			y_{1}
		x_{2}			y_{2}
		x_{2}			y_{3}
		$\vdots$			$\vdots$
		x_{n}			y_{n}

$a_{i=1,\ \cdots,\ m} \in A,\ A={\ atributos\ }$

$x_{i=1,\ \cdots,\ n}$ es un vector con los valores de los atributos, $x_{i}\subset X$, X son las características explicativas.

$y_{i=1,\ \cdots,\ n}$ es la variable independiente a predecir $\in Y$, puede ser una categoría o un valor continuo $\in \mathbb{R}$

Podemos describir a la base de datos como ${(x_{1},y_{1}),(x_{2},y_{2}),\cdots ,(x_{n},y_{n})}$, $\forall_{i=1,\ \cdots,\ n}\ (x_{i},y_{i})$ es una relación de la variable aleatoria multidimensional (x,y)

El objetivo del aprendizaje automatizado supervisado es encontrar $f: X\rightarrow Y$

1.2 Bases de datos sin etiqueta

Las bases de datos con etiqueta son utilizadas para el aprendizaje no supervisado.

a_{1}	a_{2}	a_{3}	$\cdots$	a_{m}
		x_{1}		
		x_{2}		
		x_{2}		
		$\vdots$		
		x_{n}		

$a_{i=1,\ \cdots,\ m} \in A,\ A={\ atributos\ }$

$x_{i=1,\ \cdots,\ n}$ es un vector con los valores de los atributos, $x_{i}\subset X$

2 Aprendizaje automatizado

El objetivo del aprendizaje supervisado

2.1 Función de perdida

La función $L(y,u)$ cuantifica cual es la perdida de decir u cuando el verdadero valores es y.

Algunos ejemplos de funciones de error para diferentes problemas:

- Clasificación: $L(y,u)=\mathbb{1_{{u\neq y}}} \left{\begin{matrix} 1 & si & u\neq y\ 0 & si & u=y \end{matrix}\right.$
- Regresión: $L(y,u)=(y-u)^2$
- No supervisado: $L(u)=-log(u)$ *(verosimilitud)*

Quiero encontrar una funcion f que minimiza el "*riesgo de perder*".

Función de riesgo teórica:

$R_{L}(f)=\mathbb{E}[L(y,f(x))]$

De donde $\mathbb{E}$ es la esperanza y $L(y,f(x))$ es la perdida, por lo tanto:

$f_{C}=ar\underset{f}gmin \ R_{L}(f)= ar\underset{f}gmin \ \mathbb{E}[L(y,f(x))]$

Buscamos f tal que minimiza la función.

Como ejemplo podemos pensar una regresión lineal simple en la que buscamos la recta perteneciente al conjunto C de polinomios de grado uno.

INTERTAR IMAGEN CON fc ADENTRO DE UN CIRCULO y f AFUERA

Donde f es la mejor función (no la conozco y no la voy a conocer) y f_{C} es la función que minimiza el error teórico. El problema con este planteamiento es que la esperanza depende de la distribución de la variable aleatoria que no la tenemos y tampoco conocemos f_{C} por lo que f_{C} tampoco la conozco y no la voy a conocer.

Como esto no se puede resolver vamos a utilizar los datos. En vez de minimizar el riesgo teórico voy a querer encontrar una función que minimice el riesgo empírico:

$R_{L,n}(f)=\frac{1}{n}\sum_{n}^{i=1}L(y_{i},f(x_{i}))$

Esto lo puedo encontrar porque conozco la función de perdida y los datos.