¿Que es el aprendizaje supervisado?

Curso Reconocimiento de Patrones

LCC/UNISON

Julio Waissman

Ejemplo

Decidir si otorgar un crédito a un cliente

- *Cliente*: edad, genero, estado civil, ingreso mensual, otros creditos, lugar donde vive, número de hijos, ...
- *Salida esperada*: Sí / No

Otro Ejemplo

Decidir limite de crédito para un cliente

- *Cliente*: edad, genero, estado civil, ingreso mensual, otros creditos, lugar donde vive, número de hijos, ...
- *Salida esperada*: Un número real

Entradas

- Cliente es una *instancia* x.
- Si tenemos un conjunto de instancias entonces $x^{(i)}$.

$$x \in X = X_1 \times X_2 \times \cdots \times X_n$$

$$x = (x_1, x_2, \ldots, x_n)$$

Salidas

- Salida $y \in Y$.
  - $Y = \mathbb{R}$ regresión,
  - $Y = \{F, V\}$ clasificación binaria,
  - $Y = \{C_1, C_2, \ldots, C_k\}$ clasificación.

Y porque aprendizaje supervisado

Porque asumimos que puedo contestar a las preguntas, si yo cuento con un conjunto de clientes previamente clasificados

$$CA = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(M)}, y^{(M)})\}$$

¿Y eso que significa?

Formalizando

Asumimos que existe una función

$$f : X \to Y$$

desconocida.

Formalizando

Del conjunto de todas las posibles instancias X, tenemos una muestra

$$X_T \subset X$$

$$X_T = \{x^{(1)}, x^{(2)}, \dots, x^{(M)}\}$$

los cuales provienen de un muestreo con distribución desconocida.

Formalizando

Asumimos que los valores $y^{(i)}$ que conocemos provienen de

$$y^{(i)} = f(x^{(i)}) + e$$

donde e es una variable aleatoria de distribución desconocida.

$$CA = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(M)}, y^{(M)})\}$$

Formalizando

Y para el aprendizaje vamos a decidir usar algún modelo particular. Esto es, vamos a buscar una *hipótesis*:

$$h : X \times \Theta \to Y$$

donde $\theta = (\theta_1, \ldots, \theta_p)$ son los parámetros del modelo de aprendizaje.

Aprendizaje supervisado

El aprendizaje supervisado consiste en seleccionar un modelo de aprendizaje, y ajustar un conjunto de parámetros θ^* tal que:

$$h^* \approx f$$

que significa que

$$h_{\theta^*}(x) \approx f(x), \quad \forall x \in X$$

Conjunto de hipótesis

Notese que si θ es fija (constante), entonces

$$h_\theta : X \to Y$$

y por cada θ diferente hay una función de un *conjunto de hipótesis*:

$$\mathcal{H} = \{h_\theta | \theta \in \Theta\}$$

y el aprendizaje consiste en seleccionar un $h^* \in \mathcal{H}$

Ejemplo

Si

$$X = \mathbb{R}$$

y

$$h_i(x) = w_i x + b_i$$

- ¿Cual es el vector θ?
- ¿Cual es la dimensión del conjunto $\mathcal{H}$?

Aquí viene la bronca

¿Que significa $h^* \approx f$?

Función de pérdida

$$loss : Y \times Y \to \mathbb{R}$$

que permite calcular la diferencia en tre lo medido y lo estimado

$$loss(y, \hat{y})$$

idealmente

$$loss(f(x), h_\theta(x))$$

Ejemplos de funciones de pérdida

- MSE:

$$loss(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

- MAE:

$$loss(y, \hat{y}) = |y - \hat{y}|$$

- 0/1-loss:

$$loss(y, \hat{y}) = 0 \text{ si } y = \hat{y}, \text{en otro caso } 1$$

Error fuera de muestra

Decimos que $f \approx h$ si $E_o \approx 0$ donde

$$E_o = \mathbf{E}_X[loss(f(x), h(x))]$$

Pequeños detallitos:

- No conocemos f
- No conocemos todos los valores de $x \in X$

Error en muestra

Lo que podemos medir es lo que sí conocemos

$$E_i = \frac{1}{M}\sum_{i=1}^{M} loss(y^{(i)}, h(x^{(i)}))$$

Formalizando el aprendizaje

Decimos que $f \approx h^*$ ssi

$$E_i(f, h^*) \approx 0$$

y

$$E_o(f, h^*) \approx E_i(f, h^*)$$

$E_i(f, h^*) \approx 0$

- Problema de optimización
- Encontrar h^* equivale a encontrar el vector de parámetros θ^* tal que

$$\theta^* = \arg\min_{\theta \in \Theta} \frac{1}{M} \sum_{i=1}^{M} loss(y^{(i)}, h_\theta(x^{(i)}))$$

$$E_o(f, h^*) \approx E_i(f, h^*)$$

- Generalización
- Diferencia entre aprendizaje y optimización
- Vamos a usar una noción que parece una broma:

Aprendizaje Probablemente Aproximadamente Correcto (PAC Learning)

Desigualdad de Hoeffding

$$\Pr[|E_o - E_i| \geq \epsilon] \leq \exp(-2\epsilon^2 N)$$

donde M es el número de datos y ϵ la diferencia entre el error en muestra y el error fuera de muestra impuesto.

Entonces, el planteamiento $E_o \approx E_i$ es PAC

¿Algún problema con la desigualdad de Hoeffding?

- Si lanzo una moneda 10 veces, ¿Cual es la probabilidad de obtener águila las 10 veces?
- Si 1000 personas lanzan una moneda 10 veces, ¿Cual es la probabilidad que *alguna* de las personas obtengan águila las 10 veces?