¿En que sentido el aprendizaje supervisado es posible?

Curso Inteligencia Artificial
2025-1

Julio Waissman

Recapitulando

Decimos que $f \approx h^*$ ssi

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

y

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

$E_i(h^*) \approx 0$

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

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

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

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

Aprendizaje Probablemente Aproximadamente Correcto (PAC Learning)

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

Hoy vamos a dedicarnos a ver en que sentido es posible que un modelo ajustado por aprendizaje supervisado *generalice*:

Desigualdad de Hoeffding

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

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 $E_o(h^*) \approx E_i(h^*)$ es PAC

¿Algún problema con la desigualdad de Hoeffding?

- Supongamos un problema de clasificación binaria con 10 instancias en el conjunto de entrenamiento.
- *Algoritmo de aprendizaje*: clasificar en forma aleatoria.
- ¿Cual es la probabilidad de clasificar bien las 10 instancias?
- ¿Cual es la probabilidad de clasificar bien las 10 instancias en *alguna* iteración, si el algoritmo se entrena con un máximo de 1000 *epoch*?

Traduciendo

$$\Pr[|E_o(h^*) - E_i(h^*)| \geq \epsilon] \leq \Pr[\bigcup_{h \in \mathcal{H}} |E_o(h) - E_i(h)| \geq \epsilon]$$

$$\Pr[|E_o(h^*) - E_i(h^*)| \geq \epsilon] \leq \sum_{h \in \mathcal{H}} \Pr[|E_o(h) - E_i(h)| \geq \epsilon]$$

Y el problema de aprendizaje queda como...

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

donde N es el número de hipótesis posibles en el conjunto $\mathcal{H}$.

¿Entonces no es posible el aprendizaje?

Tranquilos, esta es una *cota superior* muy superior, vamos a tratar de hacerla más chiquita.

- Vamos a bosquejar el problema de generalización sólo para la clasificación binaria
- El procedimiento se puede generalizar a regresión pero ya se usa otra caja de herramientas en matemáticas que se sale de los alcances de este curso.

Clasificación binaria

- $h_\theta : \mathcal{X} \to \{-1, 1\}, \quad h_\theta \in \mathcal{H}$
- Una gran cantidad de translapes entre diferentes hipótesis
- Respecto al conjunto de aprendizaje, muchas hipótesis son iguales

Dicotomías

- Hipótesis $h : \mathcal{X} \rightarrow \{-1, 1\}, \quad h_\theta \in \mathcal{H}$
- Dicotomía
  $h : \{x^{(1)}, \ldots, x^{(M)}\} \rightarrow \{-1, 1\}, \quad h_\theta \in \mathcal{H}(x^{(1)}, \ldots, x^{(M)})$
- $|\mathcal{H}| = N$, muy seguramente infinito
- $|\mathcal{H}(x^{(1)}, \ldots, x^{(M)})| \leq 2^M$

La función de crecimiento

$$m_{\mathcal{H}}(M) = \max_{x^{(1)},\ldots,x^{(M)} \in \mathcal{X}} |\mathcal{H}(x^{(1)},\ldots,x^{(M)})|$$

- Acotado a $m_{\mathcal{H}}(M) \leq 2^M$

Un poco mejor

$$\Pr[|E_o(h^*) - E_i(h^*)| \geq \epsilon] \leq 2m_{\mathcal{H}}(M)\exp(-2\epsilon^2 M)$$

pero todavía no lo suficiente, necesitamos más

¿Cual es la idea?

$$\Pr[|E_o(h^*) - E_i(h^*)| \geq \epsilon] \leq 2m_{\mathcal{H}}(M)\exp(-2\epsilon^2 M)$$

- Probar en que casos $m_{\mathcal{H}}(M)$ es polinomial,
- Conforme M aumente (cantidad de datos en el conjunto de aprendizaje), $m_{\mathcal{H}}(M)$ crece más lento que lo que $\exp(-2\epsilon^2 M)$ decrece.
- Entonces, el aprendizaje es posible con el $\mathcal{H}$ correcto, y un número de datos de entrenamiento suficientemente alto.

La dimensión VC

$d_{VC}(\mathcal{H})$ es el valor más grande de M para el cual $m_{\mathcal{H}}(M) = 2^M$

- Para cualquier conjunto $\{x^{(1)}, \ldots, x^{(M)}\} \in \mathcal{X}$
- Para cualquier asignación $f(x) \in \{-1, 1\}$

El problema del aprendizaje

Si $d_{VC}(\mathcal{H})$ finito, entonces $m_{\mathcal{H}}(M)$ es $\mathcal{O}(M^{d_{VC}})$

$$\Pr[|E_o(h^*) - E_i(h^*)| \geq \epsilon] \leq 4m_{\mathcal{H}}(2M)\exp(-\frac{1}{8}\epsilon^2 M)$$

La desigualdad de Vapnik--Chervonenkis

¿Como calcular la d_{VC}

- Una posibilidad es con los *grados de libertad*
- No es un calculo correcto, pero es una aproximación que suele ser adecuada
- Cuidado con el conjunto $\mathcal{H}$

¿Y cuantos datos se necesitan para que el aprendizaje exista?

- Vamos a simplificar la desigualdad VC

$$\Pr[|E_o(h^*) - E_i(h^*)| \geq \epsilon] \leq \delta$$

$$\delta \approx M^{d_{VC}} e^{-M}$$

La regla de oro para la generalización

$$M \geq 10 d_{VC}(\mathcal{H})$$

- ¿Siempre aplica?