Análisis Matemático para Inteligencia Artificial

Martín Errázquin (merrazquin@fi.uba.ar)

Especialización en Inteligencia Artificial

Optimización en ML, aprendizaje supervisado

Optimización en ML

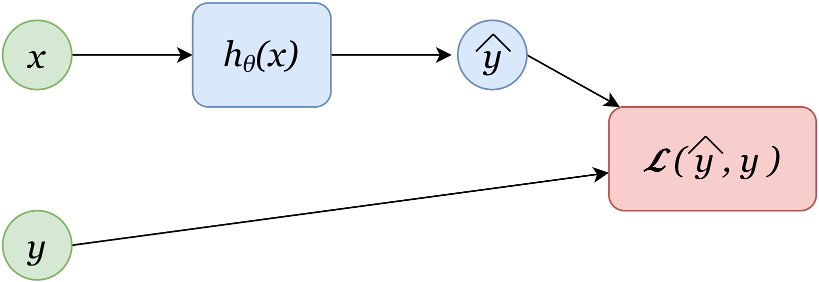
Convención: Todos los casos van a asumirse de minimización, sin pérdida de generalidad ya que maximizar f equivale a minimizar $f' = -f$.

Optimización en general: buscamos minimizar $J(\theta)$, tenemos toda la información necesaria disponible.

Optimización en ML: buscamos minimizar $J(\theta)$, sólo disponemos de un $\hat{J}(\theta)$ basado en el dataset disponible.

Conclusión: **no** son el mismo problema.

Aprendizaje supervisado: esquema



Dada una observación (x, y) fija, entonces la predicción $\hat{y} = h_\theta(x)$ depende puramente de los parámetros θ del modelo, y por lo tanto también la pérdida/error.

Para un dataset $(x_1, y_1), \ldots, (x_n, y_n)$ fijo, definimos entonces una función de costo $J(\theta)$ que sólo depende de los parámetros del modelo, y queremos minimizarla.

Proxy target/surrogate loss

Importante: Definida una función de pérdida por observación $\mathcal{L}(\hat{y}, y)$, la función de costo típicamente se define como

$$J(\theta) = \mathbb{E}[\mathcal{L}(\hat{y}, y)]$$

de donde

$$\hat{J}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(\hat{y}_i, y_i)$$

Denominamos *proxy* o *surrogate* a una función f' que queremos minimizar como *medio* para minimizar otra función f que es la que verdaderamente nos interesa.

El esquema entonces resulta:

- *aprendemos* vía train set $\rightarrow$ *necesitamos* minimizar $J_{train}(\theta)$
- *predecimos* vía test set $\rightarrow$ *queremos* minimizar $J_{test}(\theta)$

Ejemplo

Supongamos un caso de clasificación binaria donde definimos la función de pérdida como el *accuracy*, definido como

$$\mathcal{L}(\hat{y}, y) = 1\{\hat{y} \neq y\}$$

Como podemos ver, esta función de pérdida es *muy mala* para minimizar.

Planteamos entonces entrenar sobre la *cross-entropy loss*

$$\mathcal{L}_{train}(\hat{y}, y) = -y \cdot log(\hat{y}) - (1 - y) \cdot log(1 - \hat{y})$$

que nos permite ya no sólo trabajar con $\hat{y} \in \{0, 1\}$ sino todo el rango continuo $[0, 1]$ de probabilidades, además de, especialmente, ser **derivable respecto de** $\hat{y}$.

Taxonomía

Ahora que nuestro problema es minimizar $J_{train}(\theta)$, podemos separarlo en varios casos: