Machine Learning

Noise and Error

Kristóf Karacs

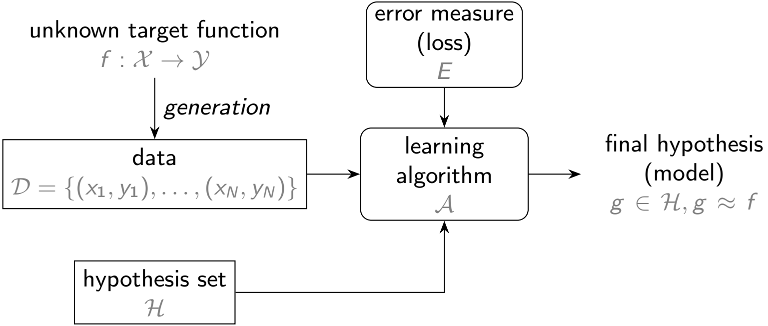
fides et ratio

On today's menu

Where do these error measures come from?

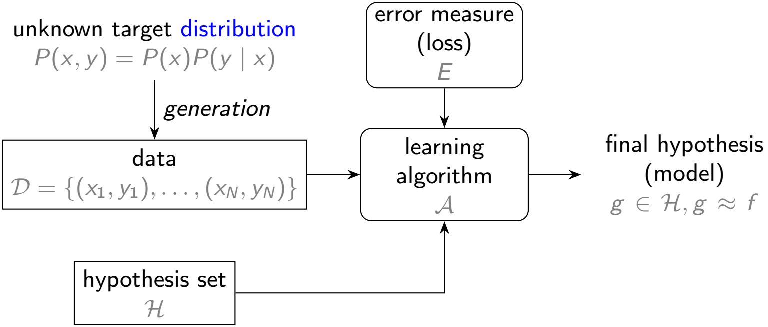
How to introduce uncertainty?

Remember: the learning flow



What if $f(x)$ is not exact?
(inaccurate information, measurement error)

Remember: the learning flow



E.g.: target = ideal mini-target + noise

Probabilistic data generation

- Suppose data is generated by $P(x, y) = P(y \mid x)P(x)$
  - $x \sim P(x)$
  - $y \sim P(y \mid x)$
  - "mini-target": $f = \arg\max_y P(y \mid x)$ (usually)
- Special case: deterministic target (no noise)
  - $P(y \mid x) = \mathbb{1}\{y = f(x)\}$

Goal of learning

Predict ideal mini targets (w.r.t. $P(y \mid x)$)
on often seen inputs (w.r.t. $P(x)$)

VC holds for $x \sim P(x)$, $y \sim P(y \mid x)$

Fun Time

Let's revisit PLA/pocket. Which of the following claim is true?

1. In practice, we should try to compute if $\mathcal{D}$ is linearly separable before deciding to use PLA.
2. If we know that $\mathcal{D}$ is not linearly separable, then the target function f must not be a linear function.
3. If we know that $\mathcal{D}$ is linearly separable, then the target function f must be a linear function.
4. None of the above.

There were a bunch of possible error measures

- 0/1 error (opposite of *accuracy*:

$$E_{out}(g) = \mathbb{E}_x \left\{ \mathbb{1}\{g(x) \neq f(x)\} \right\} \approx \frac{1}{N} \sum_{(x,y) \in \mathcal{D}} \mathbb{1}\{g(x) \neq y\}$$

- Mean Squared Error

$$E_{out}(g) = \mathbb{E}_x\{(g(x) - f(x))^2\} \approx \frac{1}{N} \sum_{(x,y) \in \mathcal{D}} (g(x) - y)^2$$

- (Binary) Cross Entropy error

$$E_{out}(g) = \mathbb{E}_x\{-\log \Pr{}_{g(x)}(f(x))\} \approx \frac{1}{N} \sum_{(x,y) \in \mathcal{D}} -\log \Pr{}_{g(x)}(y)$$

All of these are *pointwise*: $E_{out}(g) = \mathbb{E}_x \left\{ \text{err}(\text{g(x), f(x)}) \right\}$.
(Not every error is pointwise.)

Error measure with mini-targets

Suppose we only have $P(y \mid x)$. (No f.)
What should the model learn?

$$E_{out}(g) = \mathbb{E}_x \{\mathrm{err}(g(x), f(x))\}$$

Error measure with mini-targets

Suppose we only have $P(y \mid x)$. (No f.)
What should the model learn?

$$E_{out}(g) = \mathbb{E}_{x \sim P(x)} \mathbb{E}_{y \sim P(y|x)} \{\text{err}(g(x), y)\}$$

Error measure with mini-targets

Suppose we only have $P(y \mid x)$. (No f.)
What should the model learn?

$$E_{out}(g) = \mathbb{E}_{x \sim P(x)} \mathbb{E}_{y \sim P(y|x)} \{\text{err}(g(x), y)\} \approx \frac{1}{N} \sum_{(x,y) \in \mathcal{D}} \text{err}(g(x), y)$$

Error measure with mini-targets

Suppose we only have $P(y \mid x)$. (No f.)
What should the model learn?

$$E_{out}(g) = \mathbb{E}_{x \sim P(x)} \mathbb{E}_{y \sim P(y|x)} \{\mathrm{err}(g(x), y)\} \approx \frac{1}{N} \sum_{(x,y) \in \mathcal{D}} \mathrm{err}(g(x), y)$$

Depends on the error measure.

With noise, $E_{in} = 0$, but also $E_{out} = 0$ may not even be possible.

Minimising models for error measures

For an input x, model outputs prediction $\hat{y} = g(x)$

1/0 error: $\mathrm{err}(\hat{y}, y) = \mathbb{1}\{\hat{y} \neq y\}$

$g^*(x) = \arg\max_{y \in \mathcal{Y}} P(y \mid x)$

MSE: $\mathrm{err}(\hat{y}, y) = (\hat{y} - y)^2$

$g^*(x) = \sum_{y \in \mathcal{Y}} y \cdot P(y \mid x) = \mathbb{E}_{y \in P(y|x)} y$