ML Lecture 10 - Hypothesis Evaluation

Train Error — Test Error

Usually, there are observed differences. If we form a hypothesis based on evidence from training data, we need to validate it on the test data. This validation is done by computing error margins on the hypothesis formed initially. Consider a few random samples, R_1, R_2 & R_3 such that

$R_1 \xrightarrow{h}$ provides 75% accuracy
$R_2 \xrightarrow{h}$ — 80% —
$R_3 \longrightarrow$ — 70% —

} the hypothesis h boasts of accuracy in the range of 70%-80%.

75% ± 5%

Accuracy — margin of error $\Longleftrightarrow$ Confidence Interval

? Can we estimate accuracy on true data from sample data?

Def: (A) Sample accuracy is computed on sample data (usually sliced from prior population data)

(B) True Accuracy → computed on true data (population data)

Sample Error Vs true error:

Sample contains n instances → 40; hypothesis h commits 12 errors

$$\text{Error}_s = 12/40 \leftarrow \text{sample error}$$
$$= 0.3$$
$$\text{Accuracy}_s = 1 - 0.3 = 0.7$$

Next,

$$\text{Error}_{true} = \text{Error}_s \pm Z_n \sqrt{\text{errors}_s (1 - \text{errors}_s)/n}$$

C.I ⟹ set of values from the look-up table (z) based on the quantitative limit of z

Ex: 50% CI, 68%, ... ⟸

* Sample error and sample accuracy would be used interchangeably.

Consider a 95% C.I ($Z_n = 1.96$)

Hence, $$\text{Error}_{\text{true}} = 0.3 \pm \sqrt{(0.3)(0.7)/40} * 1.96$$
$$= 0.3 \pm \underbrace{0.1372}_{\text{margin of error}}$$

$$\text{Accuracy}_{\text{true}} = 0.7 \pm 0.1372$$

Baye's theorem:

Probabilistic Model;
sequential inference;
doesn't care about the decision boundary (linear/non-linear)

posterior probability

$$P(h/D) = \frac{P(D/h)\, P(h)}{P(D)}$$

Prob (hypothesis)

Prob (Data on given hypothesis)

Prob (Data)

Class label prediction ⟺ prob. of hypothesis given Data, D

(can be easily vectorized to accommodate high-dimensional data)

(Prior probability)

(Class - conditional density)

: Fish Example:

Example (from T. Mitchell)

A patient takes the cancer test and result comes back positive. The test returns correct positive results in 98% of the cases in which the disease is actually present & returns correct results in 97% of the cases in which the disease is not present. Furthermore, 0.008% of the entire population has cancer, according to data. What is the prob that the patient has cancer?

Binary classification — $\oplus$ class – cancer; $\ominus$ class – $\neg$ cancer

instances \ prediction	$\ominus$	$\oplus$	
h: $\ominus$ (TN)	0.97	0.03	FP
h: $\oplus$ (FN)	0.02	0.98	TP

Next, $P(h)$;

Prob (cancer) = 0.008

Prob ($\neg$ cancer) = 0.992

$P(D/h)$; Data, hypothesis

$P(\ominus \mid \neg \text{cancer}) = 0.97$

$P(\ominus \mid \text{cancer}) = 0.02$

$P(\oplus \mid \neg\text{cancer}) = 0.03$

$P(\oplus \mid \text{cancer}) = 0.98$

$$P(D) = \sum_{i=1}^{2} P(D/h_i)\, P(h_i)$$

(binary class)

$$P(\oplus) = P(\oplus \mid \neg\text{cancer})\, P(\neg\text{cancer}) + P(\oplus \mid \text{cancer})\, P(\text{cancer})$$

$P(\text{cancer}) = ?$ → look at the $\oplus$ data for the particular patient

Therefore, we compute

$$\boxed{P(\text{cancer} \mid \oplus) = \frac{P(\oplus / \text{cancer})\, P(\text{cancer})}{P(\oplus)}}$$

hypothesis "h"

$P(\oplus/\text{cancer}) = 0.98$; $P(\text{cancer}) = 0.008$ &

$P(\oplus) = 0.03 * 0.992 + 0.98 * 0.008 = 0.0376$

$$\therefore P(\text{cancer}/\oplus) = \frac{0.98 * 0.008}{0.0376}$$

Prob that the ~~cance~~ patient has cancer = 0.2085 //