Machine Learning Concepts

2021年1月25日 18:31

PS: Pre-Learning.

(I). Motivation: Prediction & Inference.

- Prediction → Mainly in the predicted valuable as result of input. But not each input.
  <Result>
- Inference → Mainly in the way each one of the input affect.
  <Process>

(II). Types:

- Regression → Continuous.
- Classification → Discrete.

>> all is to predict, but the type of results are different. And Both are supervised problem.

- Clustering → A set of inputs is to be divided into groups, which has the same properties.
- Density Estimation → The distribution of some space. Generally, it supposes that the samples are obey a special distribution in some hypothetic space.

(III). Kind. <Type>.

- Parametric → Based on the inputs. I can select or design a model which has been consisted by several unknown parameters. And I need to tune it to find which params. are most suited.
- Non-Parametric → it is not given by params. but it doesn't mean that it doesn't need params. It just need a large number of observations to obtain an accurate estimate for f.

(IV.) Categories:

- Supervised → The goal is to learn a general rules that input → output (maps)
  ex: classification. regression
- un-supervised → No lambel are given to the learning algorithm.
  ex: clustering.

ex: clustering.

- reinforcement Learning → The program is provided feedback in terms of rewards and punishments as it navigates its problem space.

(V). Performance Analysis.

- Confusion Matrix:

		True condition			
	Total population	Condition positive	Condition negative	Prevalence = Σ Condition positive / Σ Total population	Accuracy (ACC) = (Σ True positive + Σ True negative) / Σ Total population
Predicted condition	Predicted condition positive	True positive	False positive (Type I error)	Positive predictive value (PPV), Precision = Σ True positive / Σ Predicted condition positive	False discovery rate (FDR) = Σ False positive / Σ Predicted condition positive
	Predicted condition negative	False negative (Type II error)	True negative	False omission rate (FOR) = Σ False negative / Σ Predicted condition negative	Negative predictive value (NPV) = Σ True negative / Σ Predicted condition negative
pop. C+ C− prev. ACC / Pc+ T+ F+ PPV FDR / Pc− F− T− FOR NPV / TPR FPR LR+ DOR / FNR TNR LR−		True positive rate (TPR), Recall, Sensitivity, probability of detection = Σ True positive / Σ Condition positive	False positive rate (FPR), Fall-out, probability of false alarm = Σ False positive / Σ Condition negative	Positive likelihood ratio (LR+) = TPR / FPR	Diagnostic odds ratio (DOR) = LR+ / LR−
		False negative rate (FNR), Miss rate = Σ False negative / Σ Condition positive	True negative rate (TNR), Specificity (SPC) = Σ True negative / Σ Condition negative	Negative likelihood ratio (LR−) = FNR / TNR	F_1 score = 2 / (1/recall + 1/precision)

- Error Rate: The percentage which the samples classified false in the whole samples.
- Accuracy: The percentage which the ones classified true in the all.

	TRUE CONDITION	
	Condition Positive	Condition Negative
Predicted Positive	TRUE POSITIVE	FALSE POSITIVE
Predicted Negative	FALSE NEGATIVE	TRUE NEGATIVE

- Accuracy (ACC): $\frac{\Sigma TP + \Sigma TN}{\Sigma \text{Total Samples}}$. present the model's Accuracy.

- Precision: $P = \frac{TP}{TP + FP}$ "查准率"
- Recall: $R = \frac{TP}{TP + FN}$ "查全率" (反比).
- F1 score: $\frac{2 \times P \times R}{P + R} = \frac{2 \times TP}{\Sigma \text{Total} + TP - TN} = \frac{2}{\frac{1}{P} + \frac{1}{R}}$

↳ For importantive one of it, the f1 score can be described as:

$$F_1 \text{ score} = \frac{(1+\beta^2) \times P \times R}{(\beta^2 \times P) + R}$$

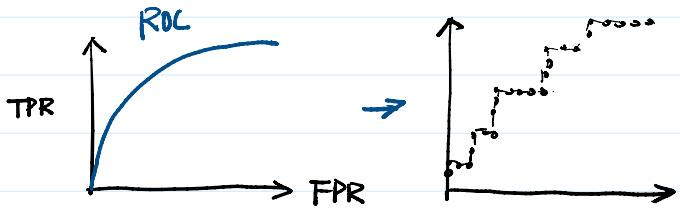
- ROC & AUC

ROC is for research the ability of generalization.

- AUC & ROC

ROC is for research the ability of generalization.

1. TPR (TRUE POSITIVE RATE)
$$= \frac{TP}{TP+FN}$$
2. FPR (FALSE POSITIVE RATE)
$$= \frac{FP}{TN+FP}$$



Problem: Realization. the ROC can't be so smooth like this. → AUC

绘图(ROC)过程:

STEP ①: 给定 m^+ Positive and m^- NEGATIVE

STEP ②: 把分类阈值设到最大。→ 把所有样例→反例。→ (0,0).

STEP ③: 把分类阈值依次设为每个样例的预测值。

若为真例。→ $(x, y+\frac{1}{m^+})$ 若为反例。→ $(x+\frac{1}{m^-}, y)$

AUC (Area of Under ROC Curve).

$$AUC = \frac{1}{2}\cdot\sum_{i=1}^{m-1}(x_{i+1}-x_i)\cdot(y_i+y_{i+1})$$ (1/2 的由来：梯形面积公式)

- MSE (Mean Squared Error)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2,$$

Measures the average of the squares of the error or deviations — that is, the difference between the estimator and what is estimated.