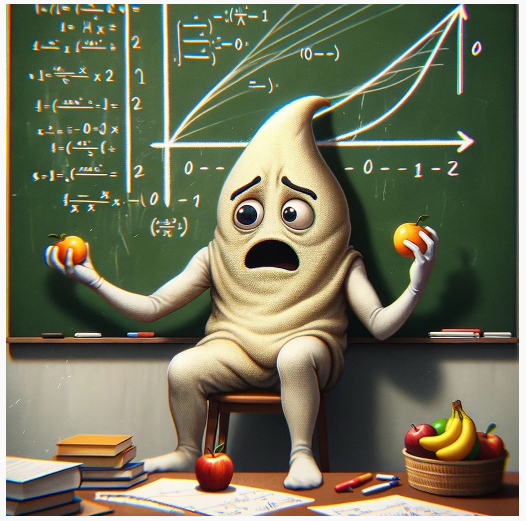
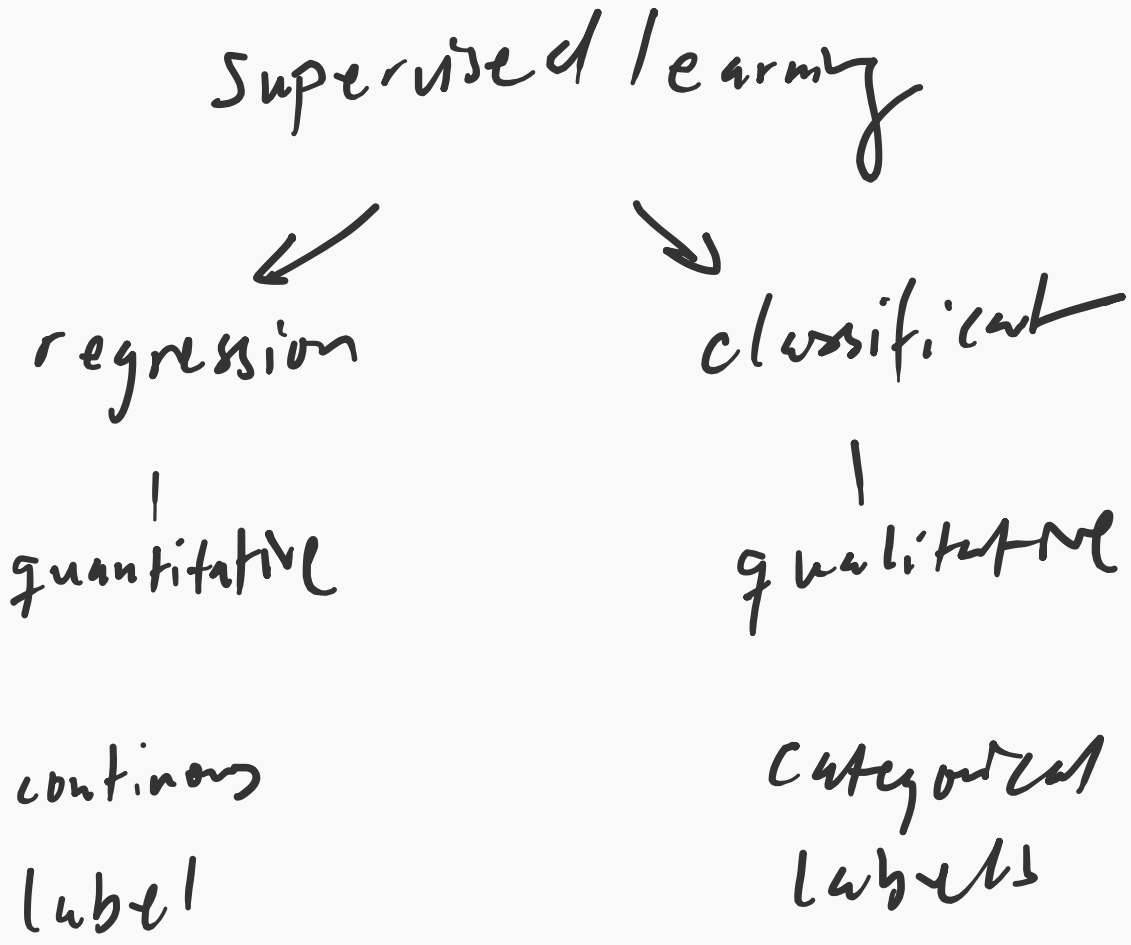
kokchun giang

logistic regression to classify data



when we have labels we are dealing with

supervised learning



linear regression can deal with categorical features, but not with categorical labels

categorical labels → need classification algorithms

categorical features require **encoding**

ex.

animal
rabbit
hare
fish

→ one-hot encoding

rabbit	hare	fish
1	0	0
0	1	0
0	0	1

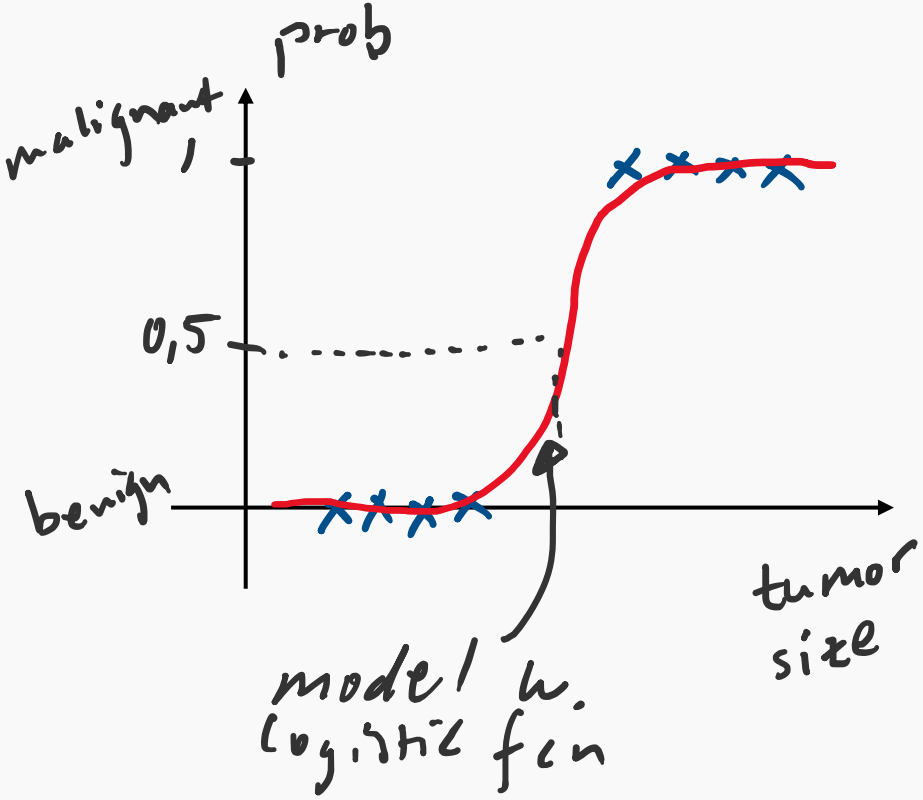
→ dummy encoding

rabbit	hare
1	0
0	1
0	0

← inferred to 3rd category => fish

model categorical label with **logistic regression**

$$y = \begin{cases} 1 & \text{malignant} \\ 0 & \text{benign} \end{cases}$$



the logistic fcn has an S-curve and takes values from 0 to 1, which can model probabilities

$$\hat{p}(x) = \frac{e^{\hat{w}_0 + \hat{w}_1 x}}{1 + e^{\hat{w}_0 + \hat{w}_1 x}}$$

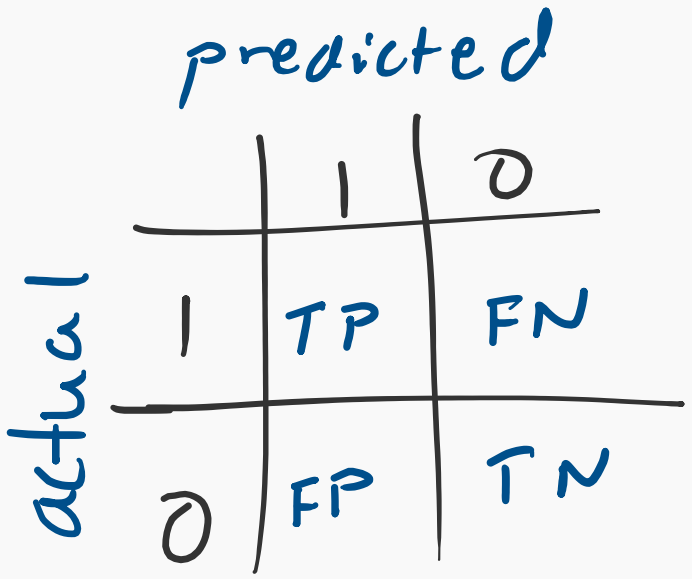
training estimates $\hat{w}_0$ & $\hat{w}_1$

Predict the class

$$\hat{y} = \begin{cases} 1 & \hat{p} \geq \text{threshold} \\ 0 & \hat{p} < \text{threshold} \end{cases}$$

evaluate a classification model

use a confusion matrix



$$\text{accuracy} = \frac{\#TP + \#TN}{\text{total}}$$

- bad for unbalanced data set

$$\text{precision} = \frac{\#TP}{\#TP + \#FP}$$

- good when we have low tolerance for FP

e.g. ham/spam

$$\text{recall} = \frac{\#TP}{\#TP + \#FN}$$

- good when low tolerance for FN

e.g. quick covid test

$$f1 = 2\frac{\text{prec} \cdot \text{rec}}{\text{prec} + \text{rec}}$$

harmonic mean of prec & rec.

→ overall performance

- good for unbalanced

multinomial logistic regression for many classes

Ex

Iris 0

Versicolor 1

Virginica 2