Quiz Week 2: Supervised machine learning

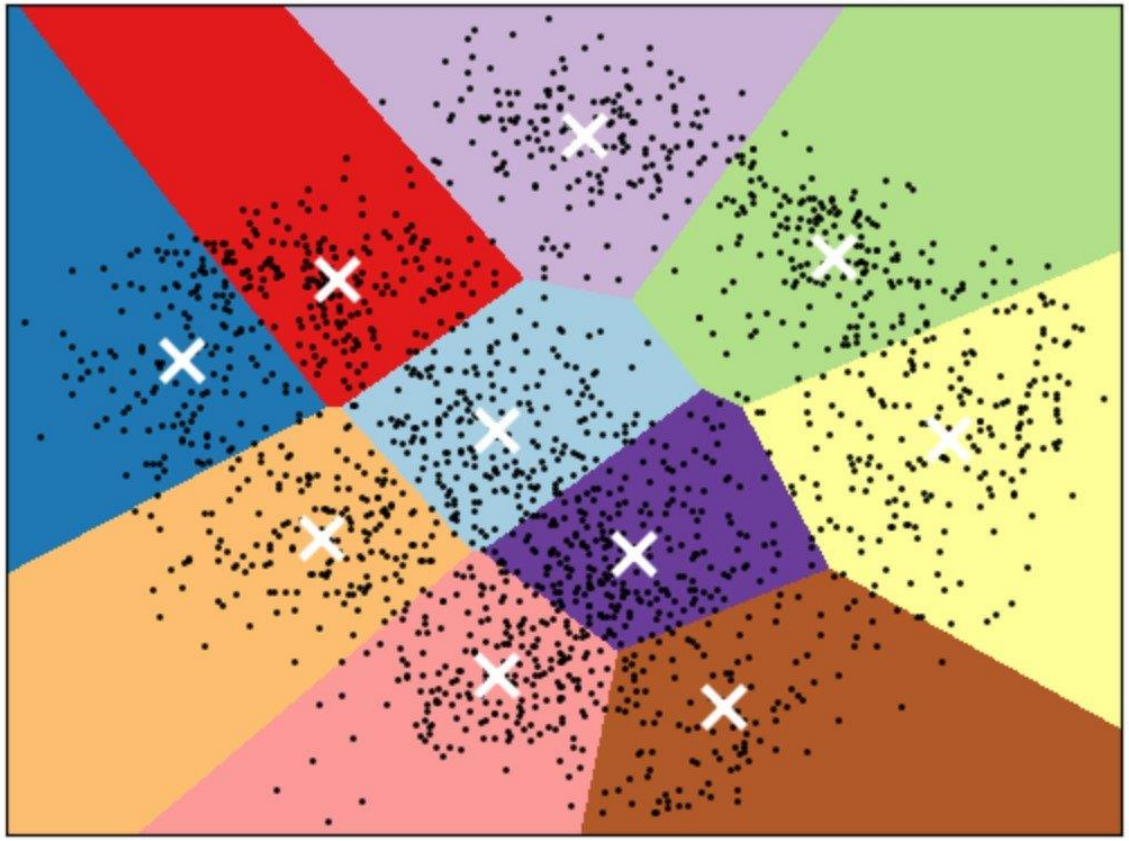
Willow Liu
22/01/2025

Q1: There are two types of problems in supervised learning. Which of the following statements are correct?

a. Clustering is one supervised-learning task
b. Regression is one supervised-learning task
c. Classification is one supervised-learning task
d. Many models can be used for both supervised learning tasks

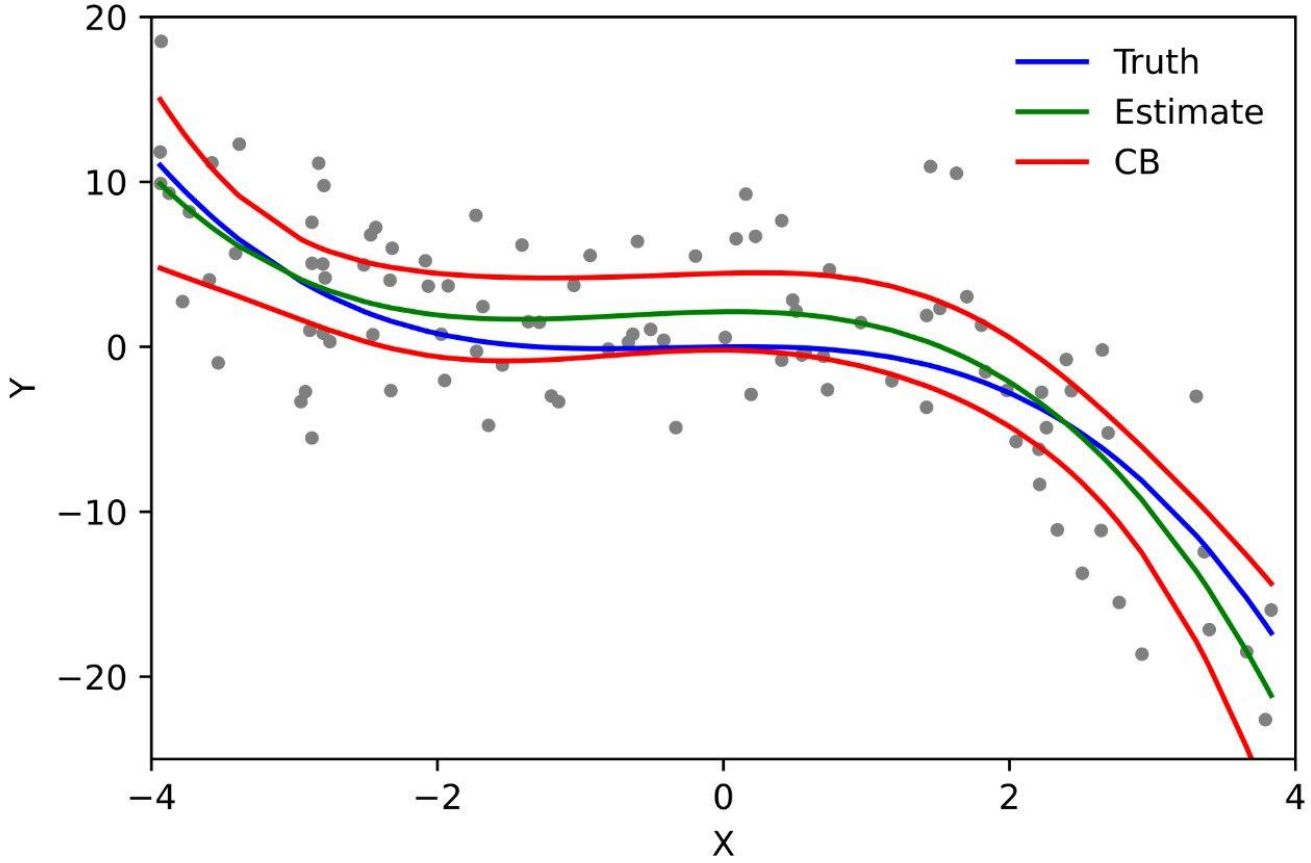
a. Clustering is one supervised-learning task ✗

it deals with unlabelled data and aims to discover hidden patterns or groupings within the data.



b. Regression is one supervised-learning task ✓

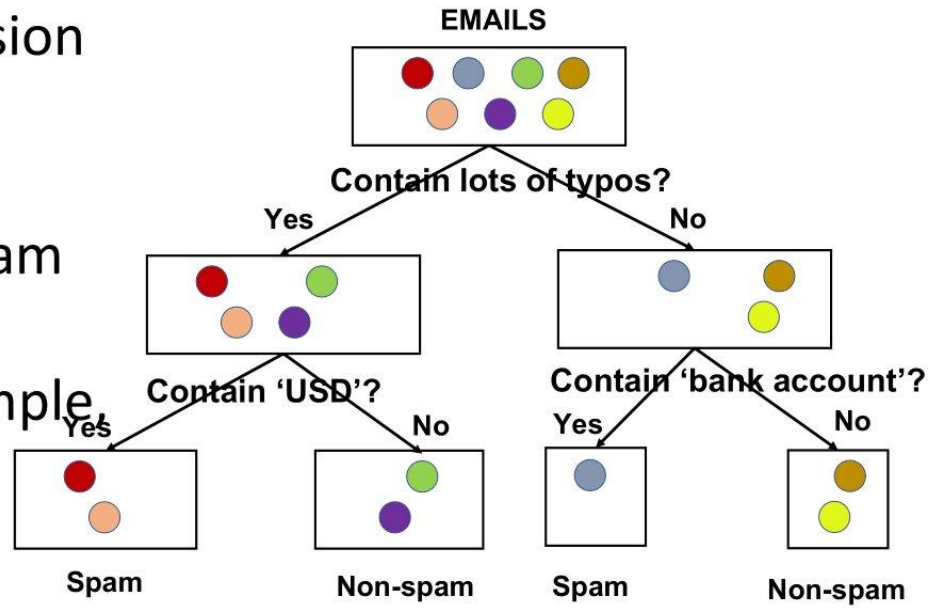
training a model on labeled data where the output (or target) variable is already known.



c. Classification is one supervised-learning task ✓

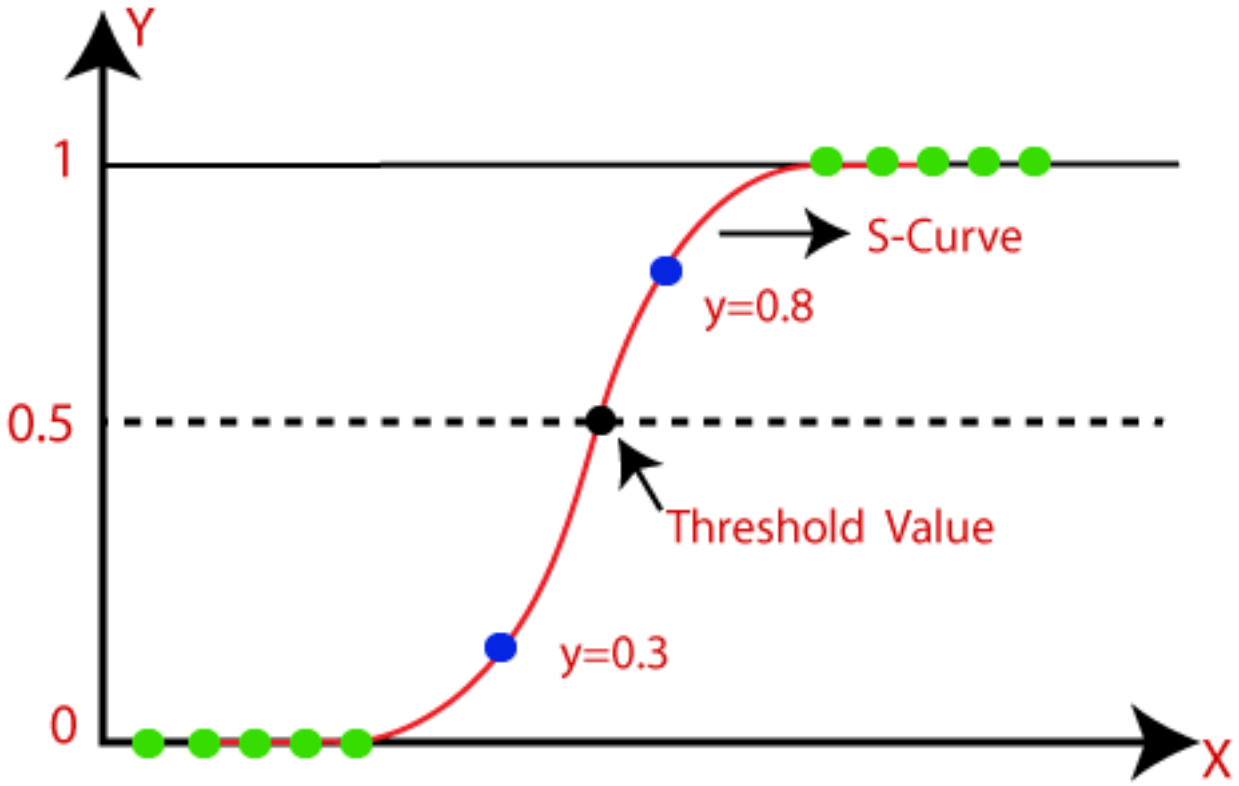
training a model to predict a discrete output (class label) based on labelled data.

- Algorithm: the idea of decision tree
- Model training: to find the optimal ‘criteria’ to split spam and non-spam
- Prediction: given a new sample,
- spam or non-spam



d. Many models can be used for both supervised learning tasks ✓

Both classification and regression aim to learn the relationship between input features (X) and an output (y).



Q2: Which of the following statements are true about supervised learning?

a. In regression desired output consists of one or more continuous variables

b. In classification, the training data consists of a set of input vectors x without any corresponding target values

c. In classification, samples belong to two or more classes and we want to learn from already labelled data how to predict the class of unlabelled data.

d. Training a model means to find the value of a set of parameters that best explain the given the data

a. In regression desired output consists of one or more continuous variables ✓

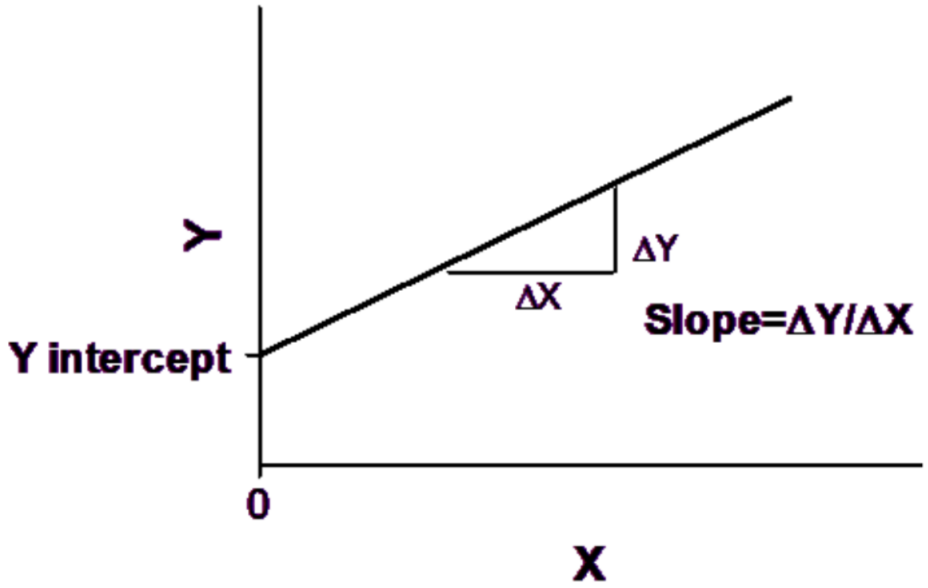
b. In classification, the training data consists of a set of input vectors x without any corresponding target values ✗

the training data **always includes corresponding target values** (also known as **labels**), which are used to teach the model the relationship between the input features (x) and the target classes.

c. In classification, samples belong to two or more classes and we want to learn from already labelled data how to predict the class of unlabelled data. ✓

d. Training a model means to find the value of a set of parameters that best explain the given the data ✓

finding the optimal **parameters** (also known as weights or coefficients) allows the model to make accurate predictions on new, unseen data



Q3: While fitting data in a supervised-learning problem, overfitting is an important challenge. Which of the following are correct?

a. Overfitting causes a low accuracy on the training set
b. Overfitting causes a low accuracy of the testing set
c. Overfitting means that the model lacks generality (i.e. it won't predict accurately unseen data points)
d. Comparing the model performance on the training and testing data will reveal the overfitting problem.

a. Overfitting causes a low accuracy on the training set ✗

b. Overfitting causes a low accuracy of the testing set ✓

Overfitting occurs when a machine learning model **learns not only the underlying patterns** in the training data but also **memorizes the noise** or **irrelevant details** (such as outliers, random fluctuations, etc.).

1. High Complexity Model: High-complexity models (e.g., deep neural networks, high-degree polynomials, deep decision trees) can **memorize** training data

2. Noise and Outliers: The model may capture **random noise** or **outliers** in the training data.

3. Lack of generalization: Overfitted models are **too specialized** to the training data./ Over-sensitivity to small variations or irrelevant details hurts performance on unseen data.

c. Overfitting means that the model lacks generality (i.e. it won't predict accurately unseen data points)

d. Comparing the model performance on the training and testing data will reveal the overfitting problem.

High Training Accuracy + Low Test Accuracy: Overfitting

Small Gap in Performance (Indicates Good Generalization): Good Generalization

Q4: Some models have hyper-parameters. Which of the following statements are correct?

a. The values of the hyper-parameters are inferred from the data via the learning process (training)

b. Cross-validation can be used to find the optimal values of the hyper-parameters

c. Hyper-parameters are parameters whose values are used to control the learning process and cannot be inferred while fitting the machine to the training set

a. The values of the hyper-parameters are inferred from the data via the learning process (training)

Hyperparameters are **not learned from the data** during training.

Hyperparameters vs. Model Parameters:

- **Model Parameters**: model learns during the training process.
- **Hyperparameters**: These are the values set **before** training begins and control the learning process itself.

Q: What are hyperparameters of a decision tree?

A:

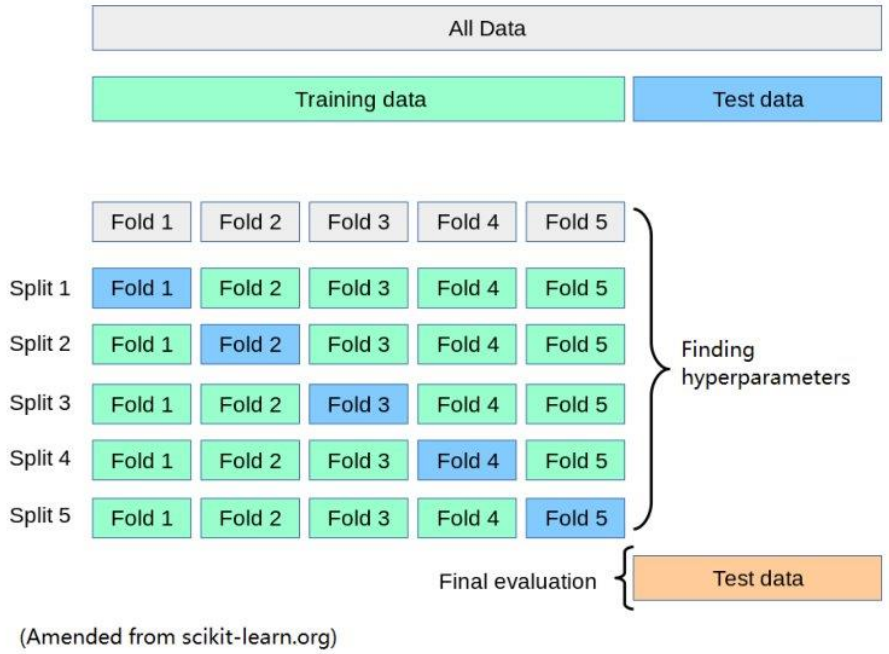
```
from sklearn.tree import DecisionTreeClassifier

model = DecisionTreeClassifier(
    max_depth=10,
    min_samples_split=5,
    min_samples_leaf=4,
    criterion="gini",
    max_features="sqrt",
    random_state=42
)

model.fit(X_train, y_train)
```

b. Cross-validation can be used to find the optimal values of the hyper-parameters

Cross validation: better use of data for model training



(Amended from scikit-learn.org)

Example: to find the optimal tree heights (TH range: 5, 10, 15) using 5-fold CV.
To get performance of TH=5: 5 models are trained.

Model	1	2	3	4	5
Training	Fold 2/3/4/5	Fold 1/3/4/5	Fold 1/2/4/5	Fold 1/2/3/5	Fold 1/2/3/4
Evaluation	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5

Performance of TH=5: average of these five models

Pick up TH with best performance, then train a model using this TH on the whole training data

21

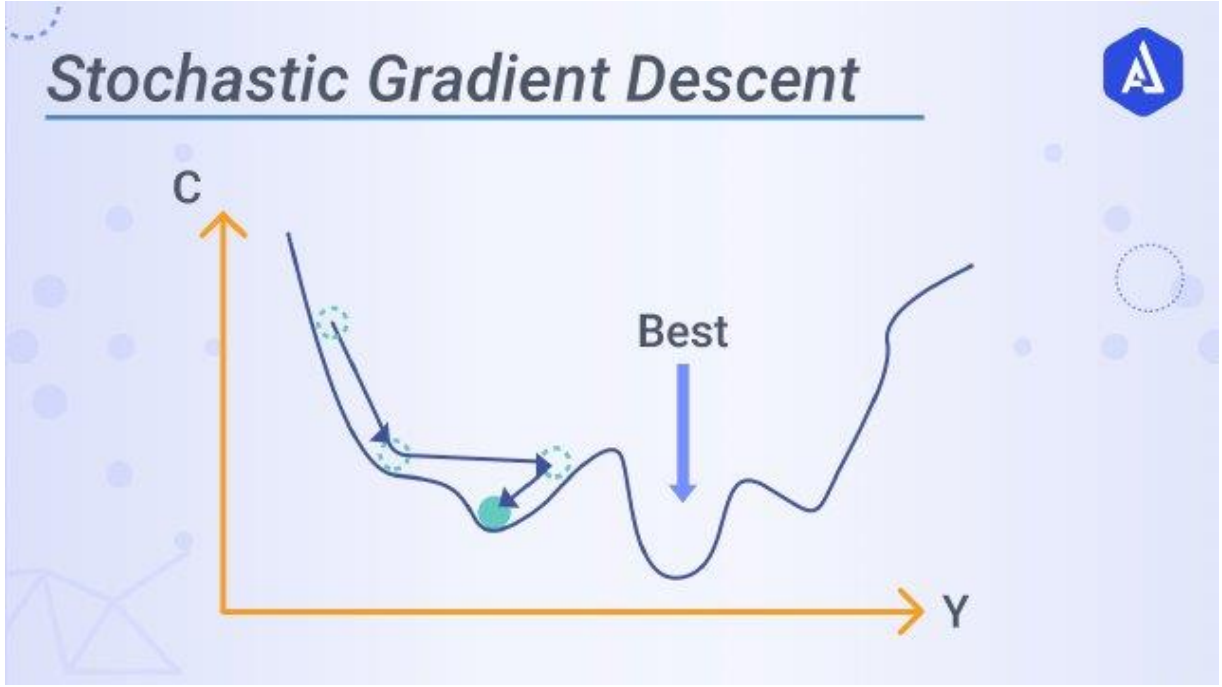
c. Hyper-parameters are parameters whose values are used to control the learning process and cannot be inferred while fitting the machine to the training set

Q5: Which of the following statements are *NOT* true about randomness in supervised learning?

a. There is randomness in data splitting process (train vs test sets)
b. There is randomness while learning the values of the model's parameters
c. In scikit-learn, given some data and a method, we will always get different results while training a model, as there is no way for the user to control the randomness of the model training.
d. To mitigate randomness, we should perform multiple runs and report average and standard deviation of performance

a. There is randomness in data splitting process (train vs test sets) ✓

b. There is randomness while learning the values of the model's parameters ✓



c. In scikit-learn, given some data and a method, we will always get different results while training a model, as there is no way for the user to control the randomness of the model training.

We **can control the randomness** during model training by setting a **random seed**.

d. To mitigate randomness, we should perform multiple runs and report average and standard deviation of performance of the model's parameters

```
[1]: from sklearn.linear_model import SGDClassifier
     from sklearn.datasets import make_classification
     import numpy as np

[1]: array([[ 6.70814003,  5.25291366, -7.55212743,  5.18197458,  1.37845099]])

[3]: rng = np.random.RandomState(0)
     X, y = make_classification(n_features=5, random_state=rng)
     sgd = SGDClassifier(random_state=rng)

     sgd.fit(X, y).coef_

[3]: array([[ 8.85418642,  4.79084103, -3.13077794,  8.11915045, -0.56479934]])

[5]: rng = np.random.RandomState(0)
     X, y = make_classification(n_features=5, random_state=rng)
     sgd = SGDClassifier(random_state=rng)

     sgd.fit(X, y).coef_

[5]: array([[ 8.85418642,  4.79084103, -3.13077794,  8.11915045, -0.56479934]])

[ ]: sgd.fit(X, y).coef_
```