Quiz week 2

Q1

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

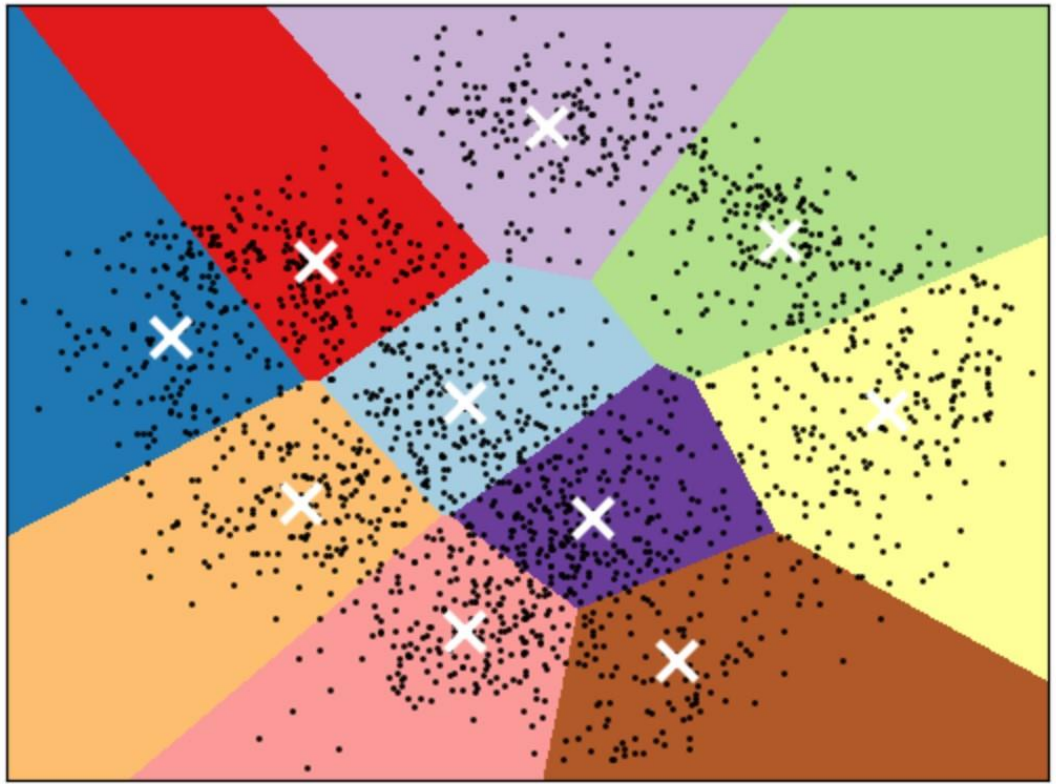
1. Clustering is one supervised-learning task
2. Regression is one supervised-learning task
3. Classification is one supervised-learning task
4. Many models can be used for both supervised learning tasks

https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_digits.html#sphx-glr-auto-examples-cluster-plot-kmeans-digits-py

Q1

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

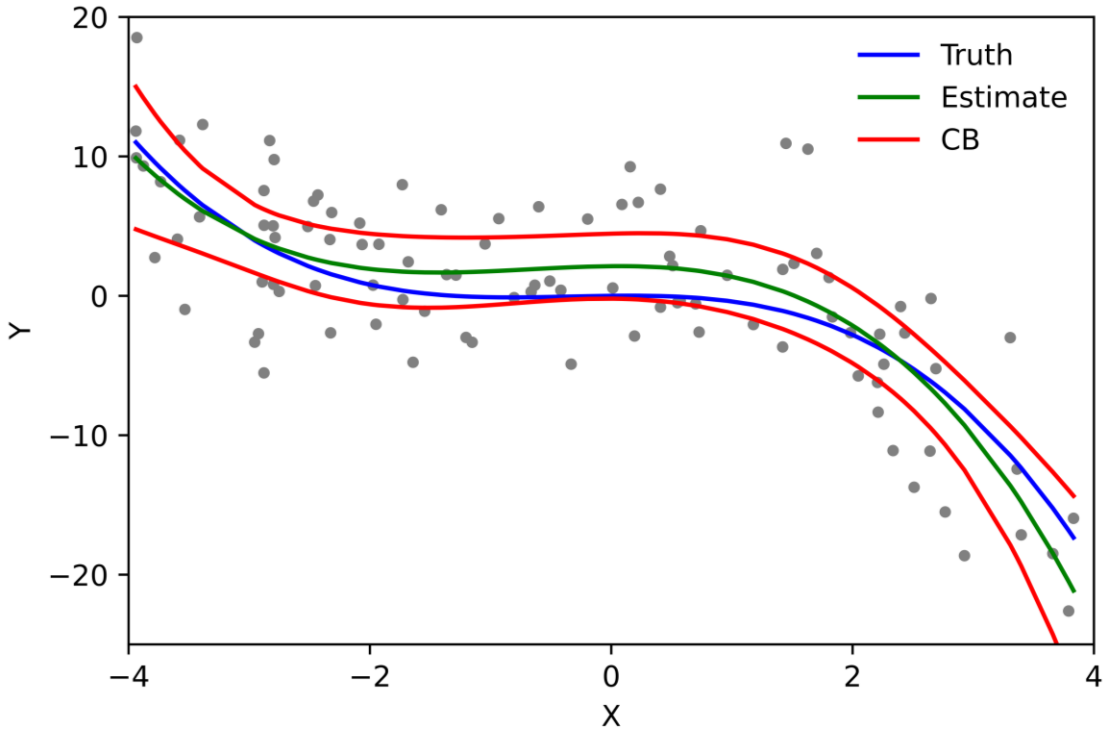
Clustering is one supervised-learning task -> NO



Q1

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

Regression is one supervised-learning task -> YES

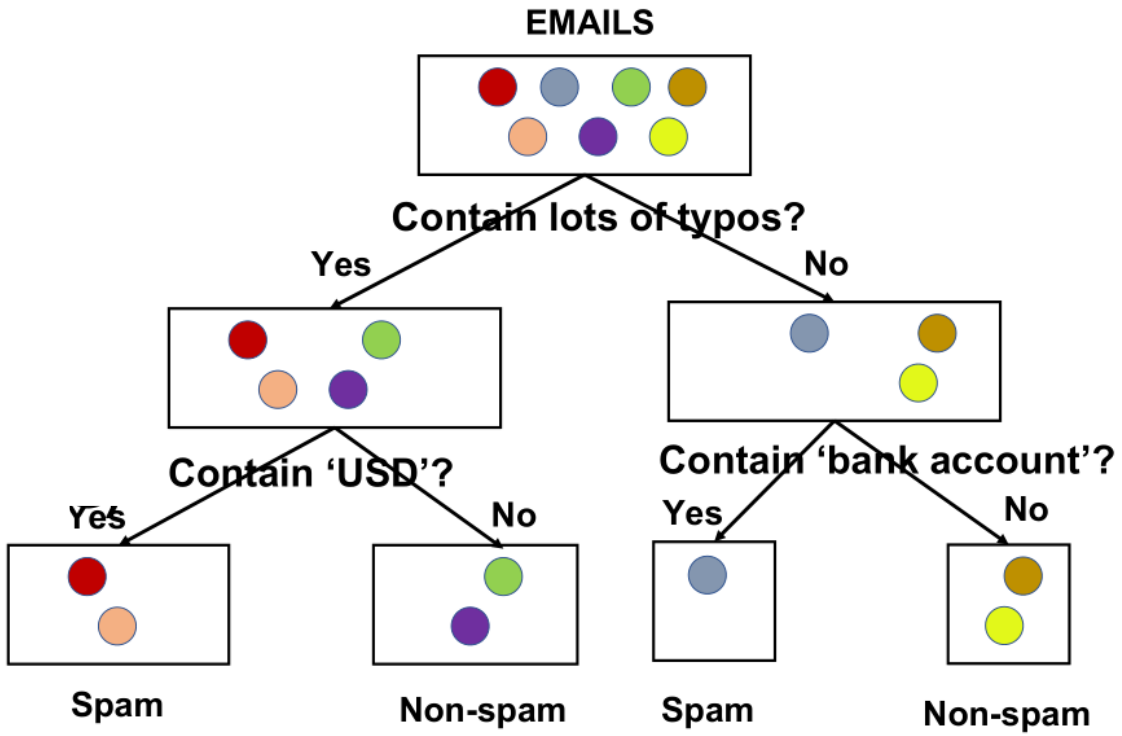


Q1

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

Classification is one supervised-learning task -> YES

1. Model training: to find the optimal ‘criteria’ to split spam and non-spam
2. Prediction: given a new sample, is it spam?

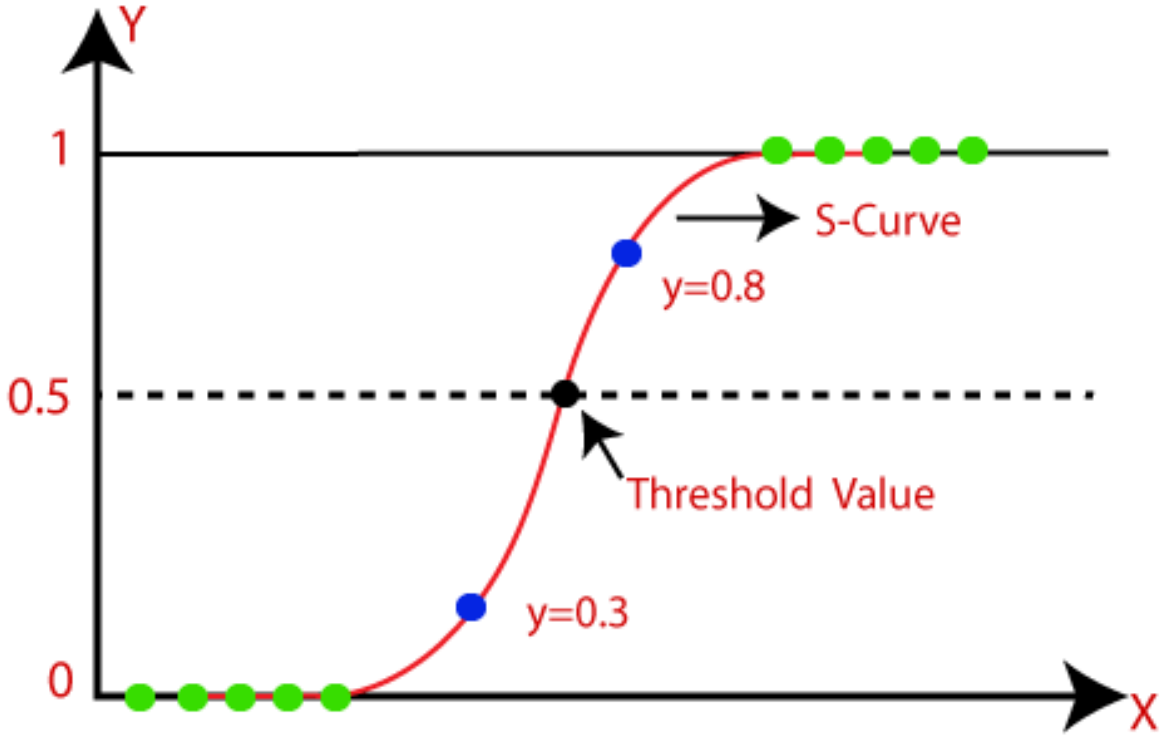


https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_digits.html#sphx-glr-auto-examples-cluster-plot-kmeans-digits-py

Q1

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

Many models can be used for both supervised learning tasks -> YES



Q1

There are two types of problems in supervised learning. Which of the following statements are correct? **(2,3,4)**

1. Clustering is one supervised-learning task
2. **Regression is one supervised-learning task**
3. **Classification is one supervised-learning task**
4. **Many models can be used for both supervised learning tasks**

Q2

Which of the following statements are true about supervised learning?

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

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

3. 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.

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

Q2

Which of the following statements are true about supervised learning?

In regression desired output consists of one or more continuous variables -> YES

Q2

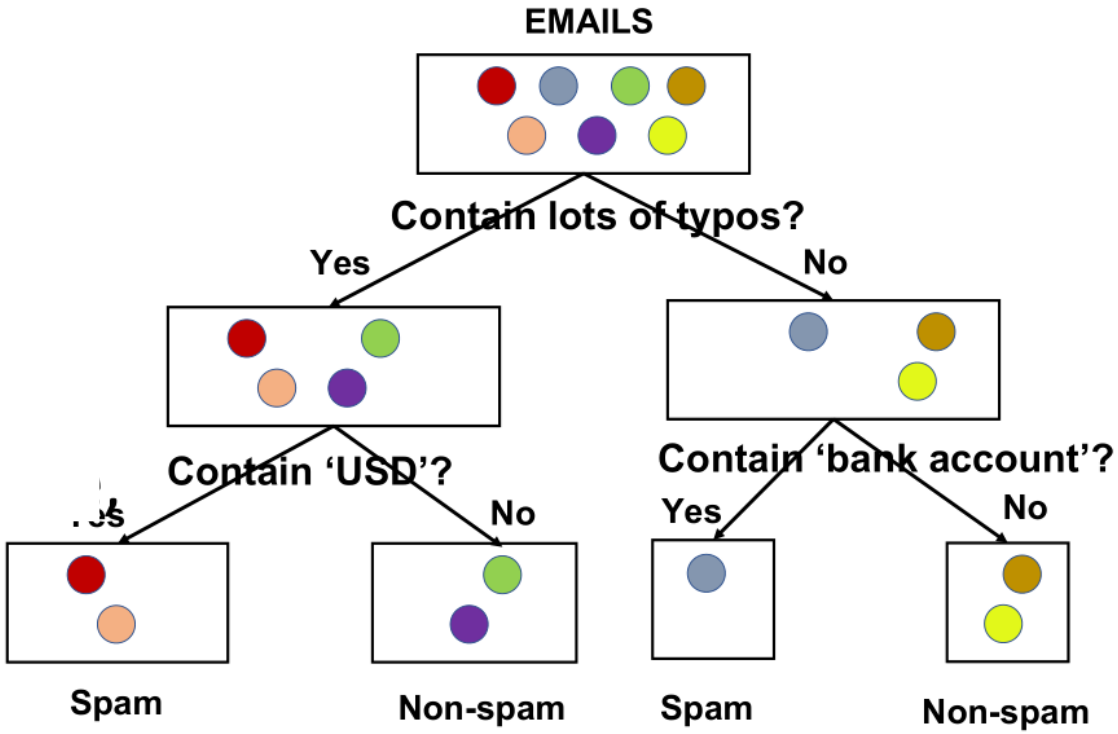
Which of the following statements are true about supervised learning?

In classification, the training data consists of a set of input vectors x without any corresponding target values -> NO

Q2

Which of the following statements are true about supervised learning?

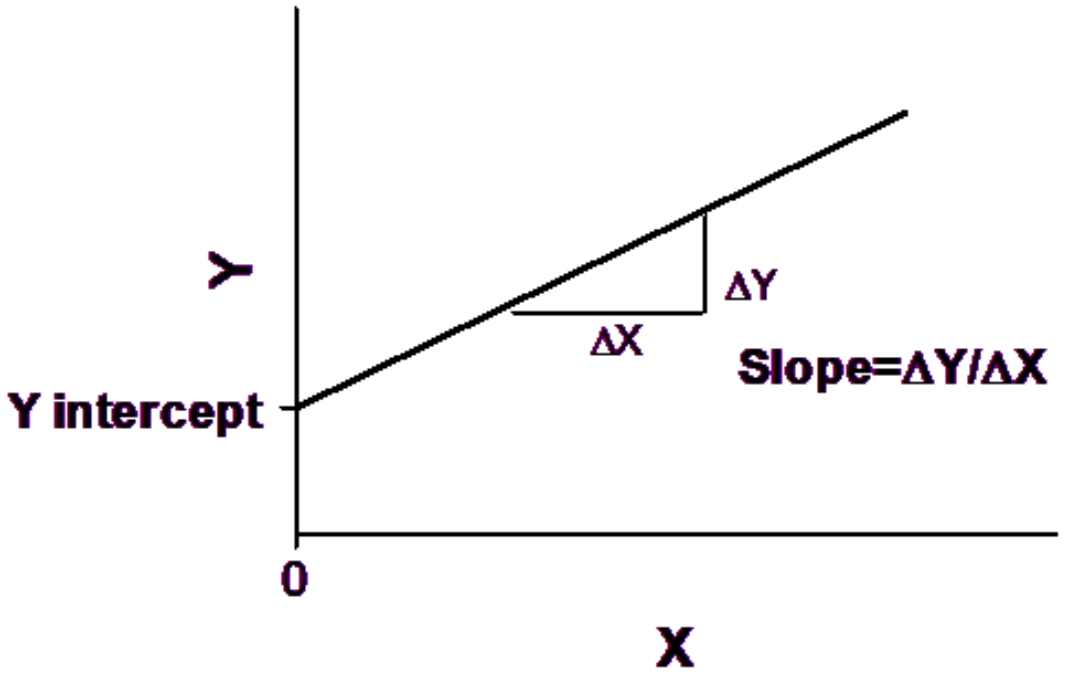
In classification, samples belong to two or more classes and we want to learn from already labelled data how to predict the class of unlabelled data -> YES



Q2

Which of the following statements are true about supervised learning?

Training a model means to find the value of a set of **parameters** that best explain the given the data -> YES



Q2

Which of the following statements are true about supervised learning? **(1,3,4)**

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

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

3. 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.

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

Q3

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

1. Overfitting causes a low accuracy on the training set

2. Overfitting causes a low accuracy of the testing set

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

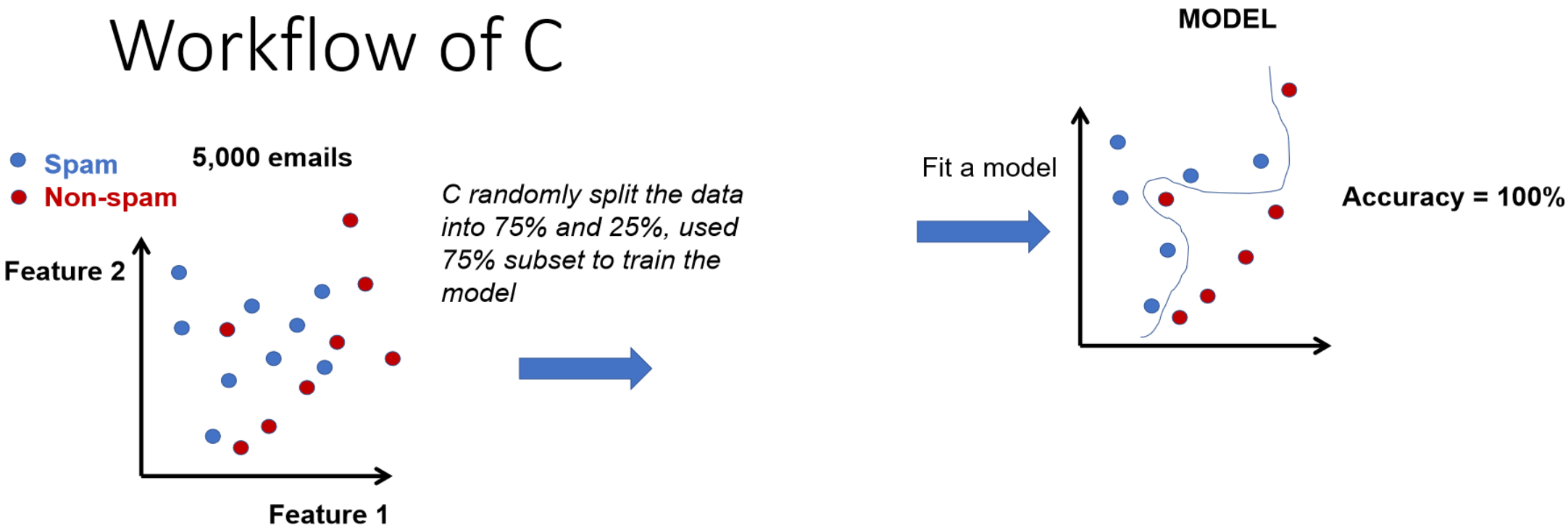
4. Comparing the model performance on the training and testing data will reveal the overfitting problem.

Q3

3. While fitting data in a supervised-learning problem, overfitting is an important challenge. Which of the following are correct? **(see also next slide)**

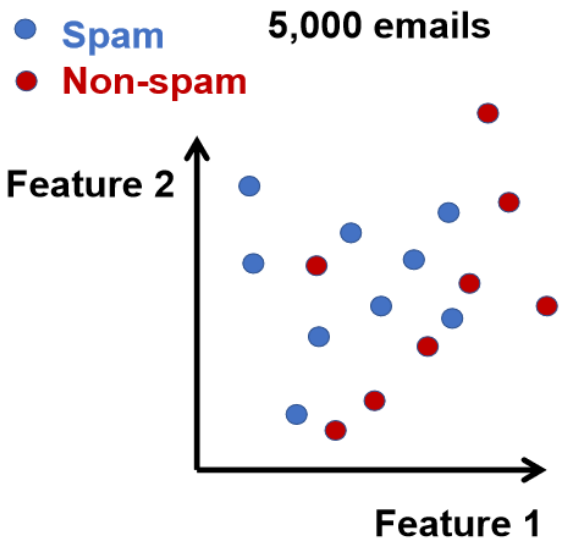
1. Overfitting causes a low accuracy on the **training set -> NO**
2. Overfitting causes a low accuracy of the **testing set -> YES**
3. Overfitting means that the model lacks generality (i.e. it won't predict accurately unseen data points) -> **YES**
4. Comparing the model performance on the training and testing data will reveal the overfitting problem -> **YES**

Workflow of C

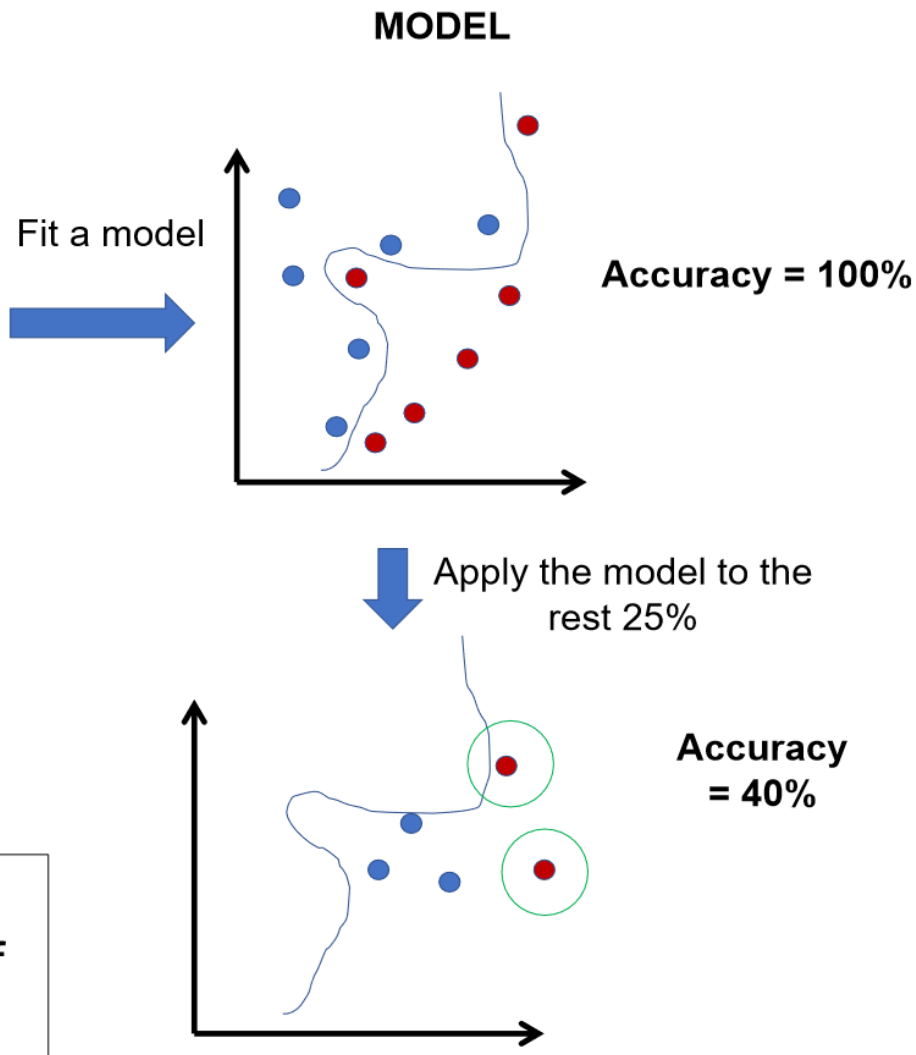


Q3

Workflow of C



C randomly split the data into 75% and 25%, used 75% subset to train the model



- Can we recognise overfitting in the workflow?
- Can we give an unbiased performance metric of this trained model?

Q3

3. While fitting data in a supervised-learning problem, overfitting is an important challenge. Which of the following are correct? **(2,3,4)**

1. Overfitting causes a low accuracy on the training set

2. Overfitting causes a low accuracy of the testing set

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

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

Q4

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

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

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

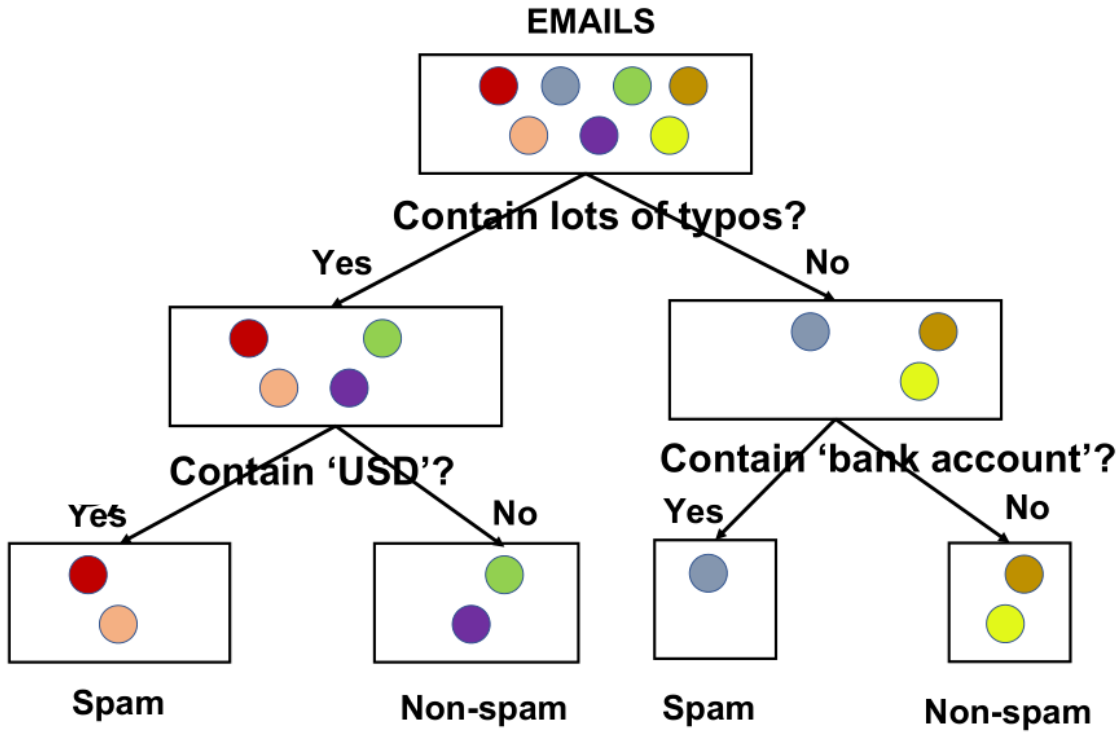
3. Hyper-parameters a parameters whose values are used to control the learning process and cannot be inferred while fitting the machine to the training set

Q4

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

1. The values of the hyper-parameters are inferred from the data via the learning process (training) -> NO

Q: What are hyperparameters of a decision tree?



Q4

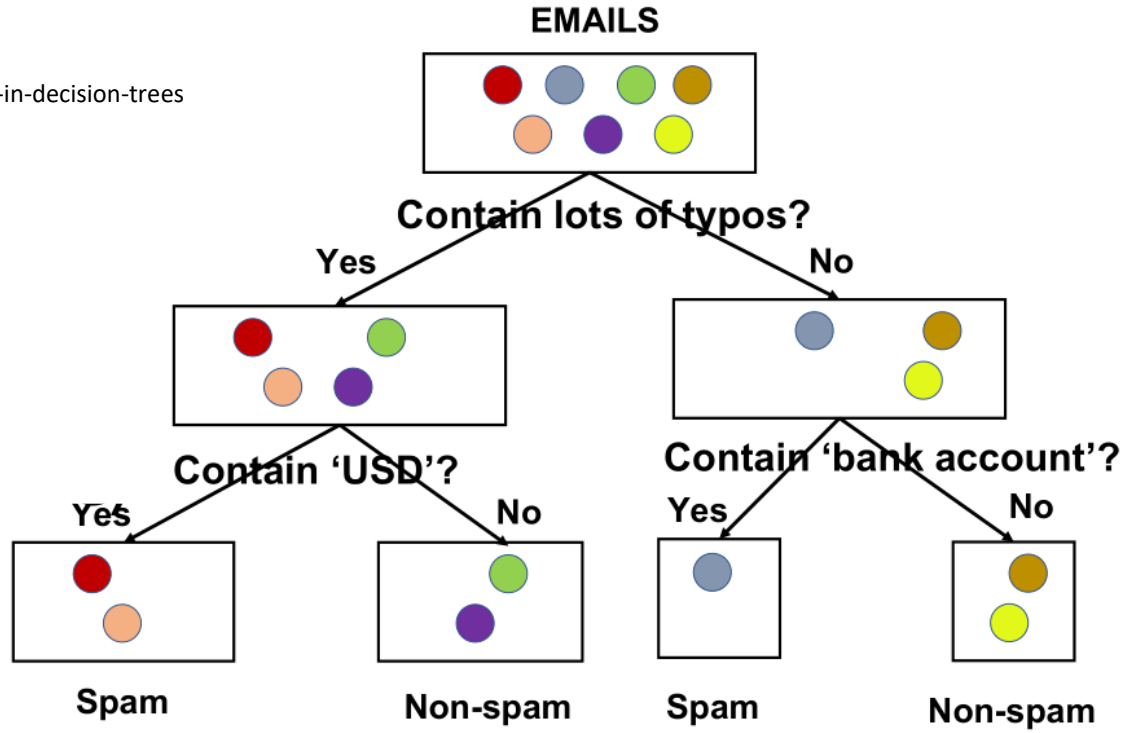
4. Some models have hyper-parameters. Which of the following statements are correct?

1. The values of the hyper-parameters are inferred from the data via the learning process (training) -> NO

https://inria.github.io/scikit-learn-mooc/python_scripts/trees_hyperparameters.html#other-hyperparameters-in-decision-trees

Q: What are hyperparameters of a decision tree?

A: *max_depth*, *min_samples_per_leaf*, *min_samples_split*, *max_leaf_nodes*, or *min_impurity_decrease*



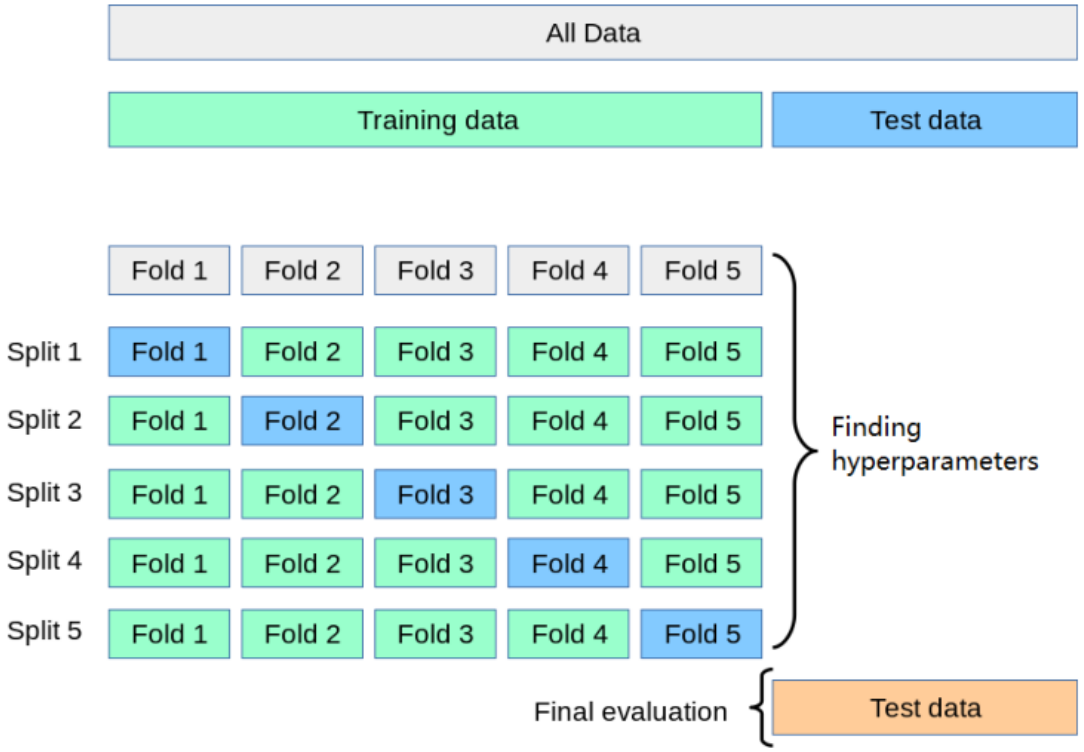
Q4

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

2. Cross-validation can be used to find the optimal values of the hyper-parameters -> YES

Q4

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

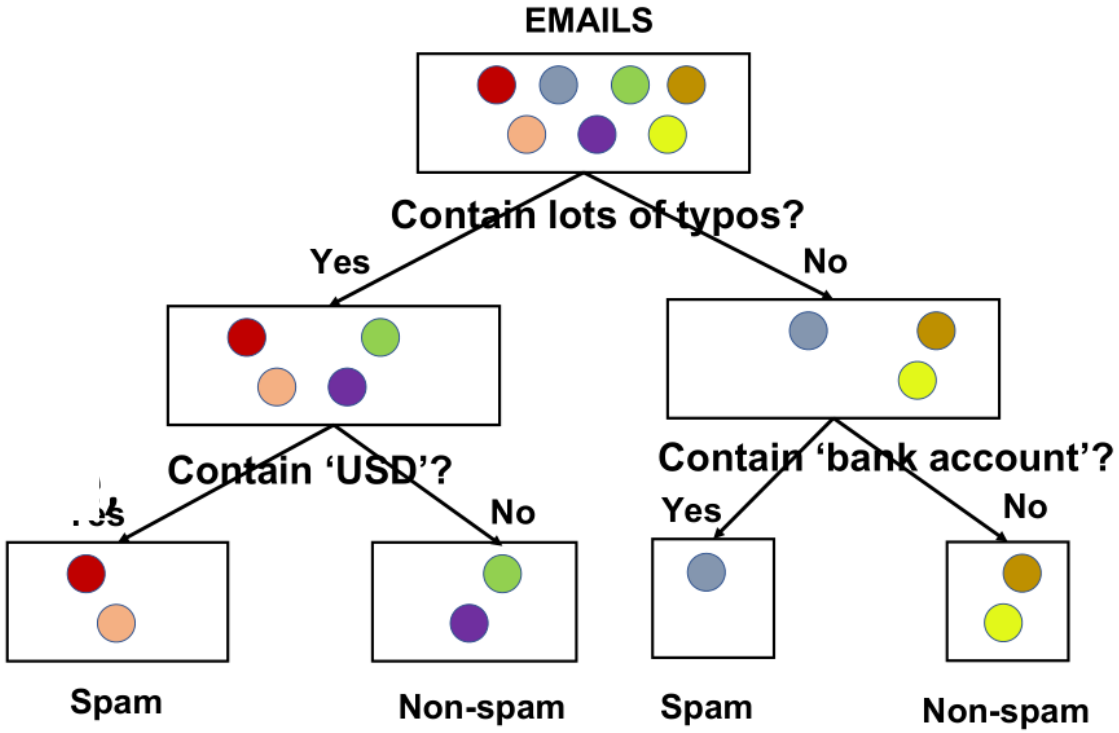
Performance of TH=5: average of these five models

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

Q4

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

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



Q4

4. Some models have hyper-parameters. Which of the following statements are correct? **(2,3)**

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

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

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

Q5

5. Which of the following statements are *NOT* true about randomness in supervised learning?

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

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

3. 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.

4. To mitigate randomness, we should perform multiple runs and report average and standard deviation of performance

Q5

5. Which of the following statements are *NOT* true about randomness in supervised learning?

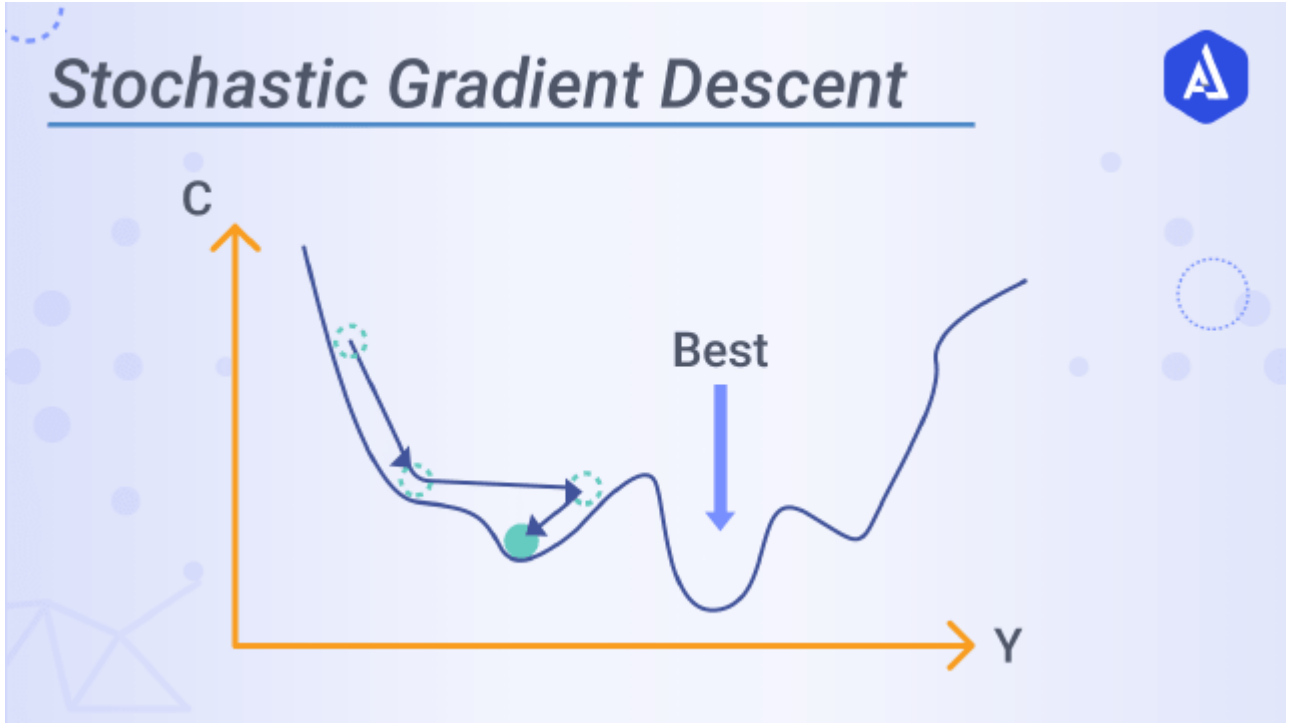
1. There is randomness in data splitting process (train vs test sets)

Important because it helps eliminate inherent biases and is a best practice to ensure you're building a generalized machine learning model, e.g. cross-validation

Q5

5. Which of the following statements are *NOT* true about randomness in supervised learning?

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



Q5

5. Which of the following statements are *NOT* true about randomness in supervised learning?

3. 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 -> NO

[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_
```

Q5

5. Which of the following statements are *NOT* true about randomness in supervised learning?

3. 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 -> NO

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

[2]:
```
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_
```

[2]:
```
array([[ 8.85418642,  4.79084103, -3.13077794,  8.11915045, -0.56479934]])
```

[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]])
```

[4]:
```
sgd.fit(X, y).coef_
```

[4]:
```
array([[ 6.70814003,  5.25291366, -7.55212743,  5.18197458,  1.37845099]])
```

Q5

5. Which of the following statements are *NOT* true about randomness in supervised learning?

4. To mitigate randomness, we should perform multiple runs and report average and standard deviation of performance

Q5

5. Which of the following statements are *NOT* true about randomness in supervised learning? **(3)**

1. There is randomness in data splitting process (train vs test sets)
2. There is randomness while learning the values of the model's parameter
3. **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**
4. To mitigate randomness, we should perform multiple runs and report average and standard deviation of performance