시험

1. Which of the following does the formula below generalize?

$$\hat{y} = \text{sign}\left(\mathbf{w}^\top \mathbf{x} + b\right) = \text{sign}\left(\sum_i w_i x_i + b\right)$$

A. Linear model for regression

B. Linear model for classification

C. Multi-class classification

D. Agglomerative clustering

- Linear model for regression

$$y(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + w_0$$

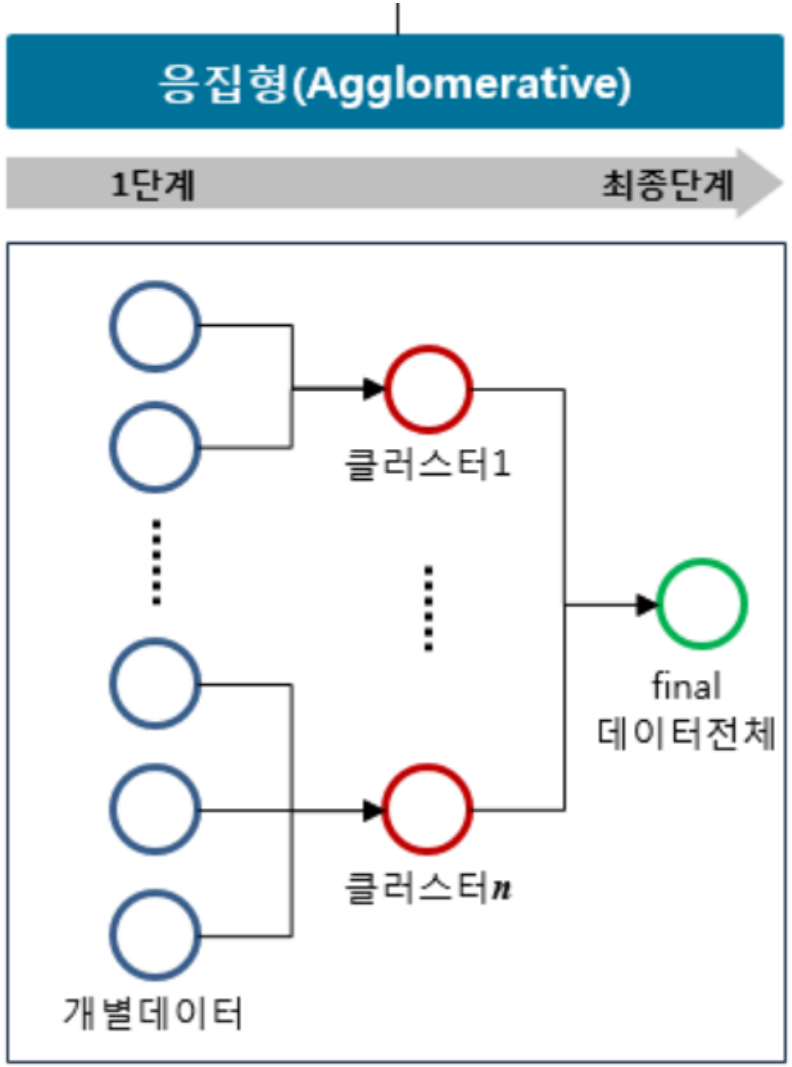
- Linear model for classification

$$\hat{y} = \text{sign}\left(\mathbf{w}^\top \mathbf{x} + b\right)$$

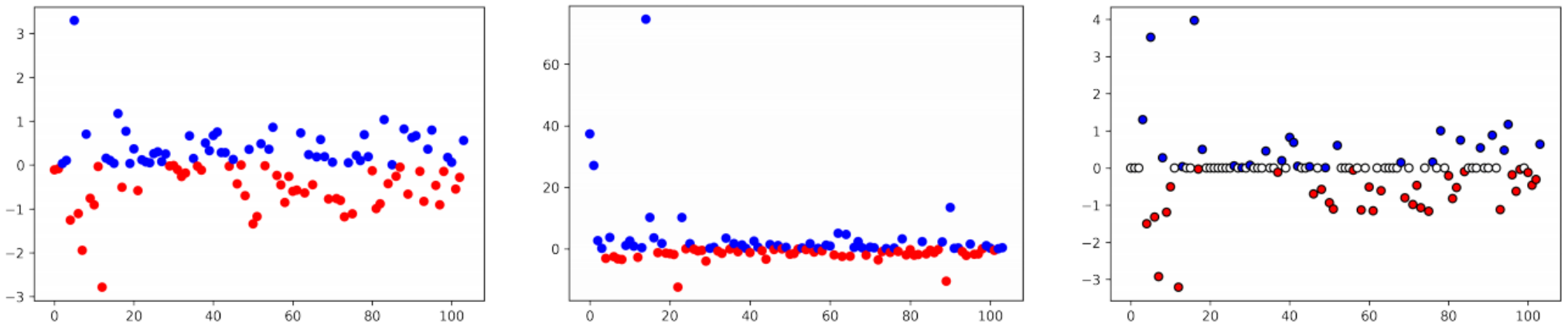
- Multi-class classification

$$\max_i h_\theta^{(i)}(x)$$

- Agglomerative clustering



2. In Figures 1a-1c the y-axis shows magnitude of coefficients w corresponding to respectively the following linear regression models:



(a) (b) (c)

Figure 1: Coefficients w corresponding to three different linear models for regression; for easier interpretation, positive values are coloured in blue and negative in red.

A. Ordinary Least Squares, Ridge, Lasso

B. Ordinary Least Squares, Lasso, Ridge

C. Ridge, Ordinary Least Squares, Lasso

D. Lasso, Ordinary Least Squares, Ridge

3. Which of the following statements are true?

(a) Ordinary Least Squares regression performs L1 regularization.

A. True

B. False

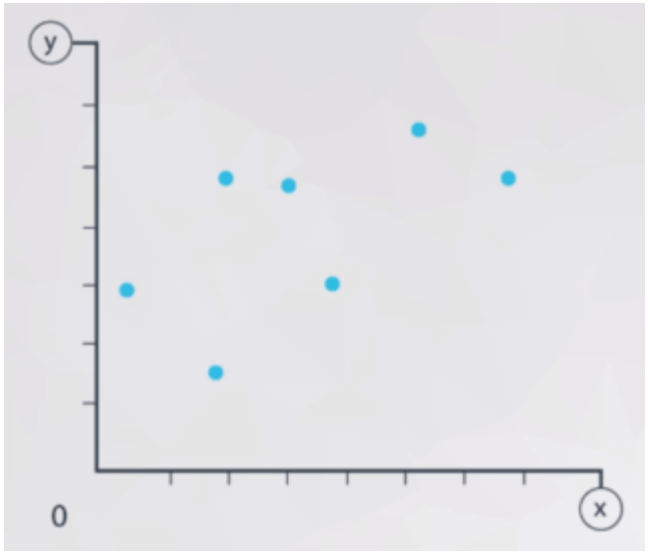
(b) Ridge regression performs L1 regularization.

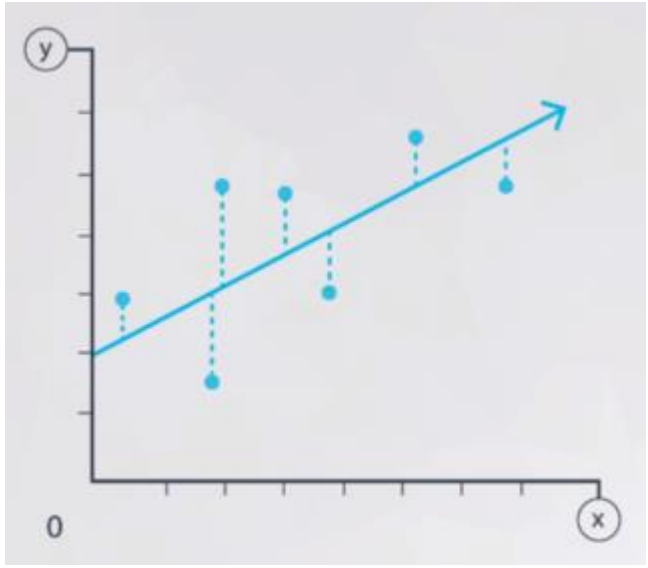
A. True

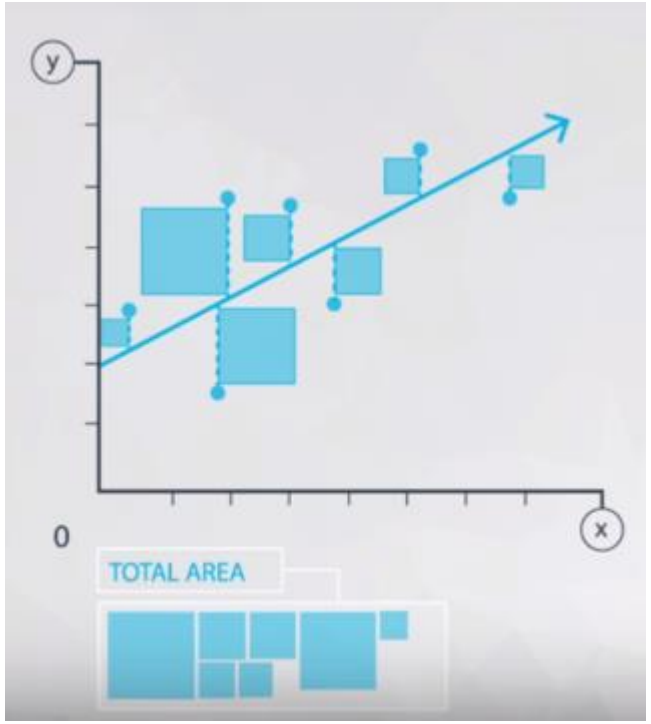
B. False

선형모델

- OLS : 최소 제곱법을 이용한 선형 모델.





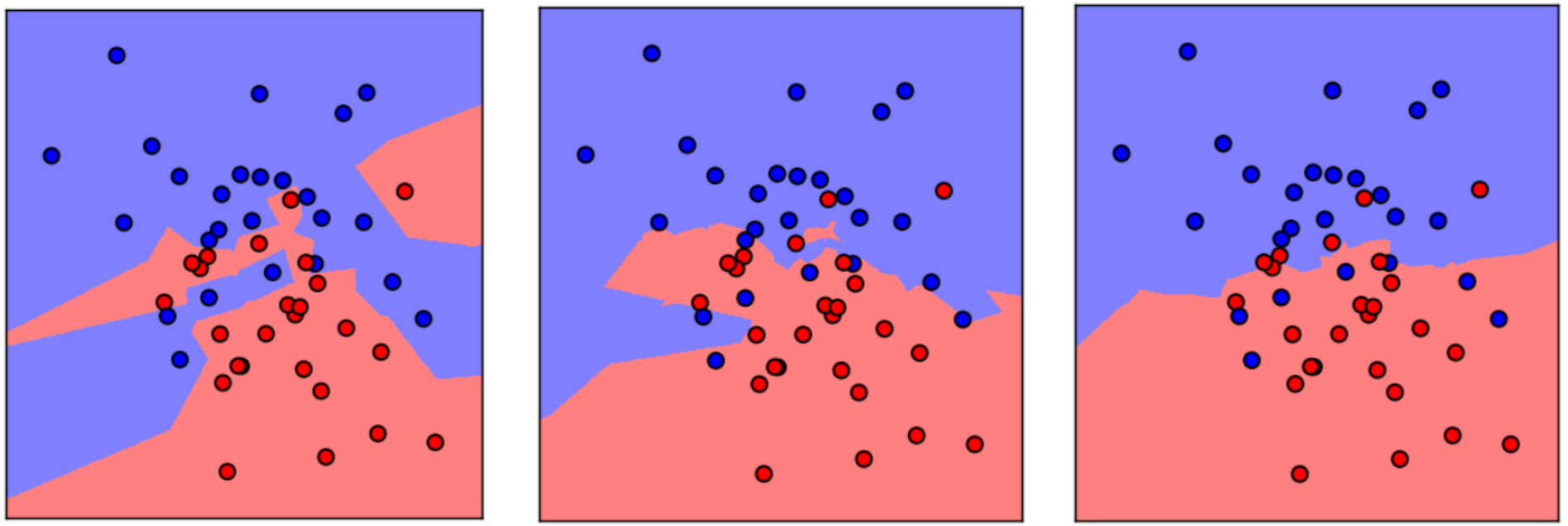


- 데이터 사이의 평균 제곱 오차를 최소화
- 데이터를 가장 잘 표현하는 회귀선을 그림.
- 매개변수가 없다는 특징이 있음.

선형모델

- Ridge
  - 제약조건이 추가됨.
  - 가중치의 절댓값을 가능한 작게 만들며 0에 가깝게 함.
  - L2 규제를 사용 (L2 norm의 제곱을 페널티로 적용, 두 벡터 사이의 직선 거리)

- Lasso
  - 제약조건이 추가됨.
  - 계수를 0에 가깝게 만드는 것은 비슷함.
  - 차이점은 Lasso는 실제로 0이 될 수 있으며 Ridge는 아님. -> 자동적으로 feature를 선택할 수 있게 함
  - L1규제를 사용 (L1 norm의 제곱을 페널티로 적용, 두 벡터 사이의 각 원소들의 차이의 절대값의 합)



(a) (b) (c)

Figure 2: Decision boundaries created by k-NN classifier for different values of parameter k.

4. Sort the plots from Figure 2 in increasing order of model complexity.

   A. $a \rightarrow b \rightarrow c$

   B. $a \rightarrow c \rightarrow b$

   C. $b \rightarrow a \rightarrow c$

   D. $c \rightarrow b \rightarrow a$

모델의 복잡도

- 분류를 잘 할수록 복잡도는 올라감 – 낮은 K값
- 하지만, Overfitting일 수도 있음.

- 반대로 분류를 잘하지 못할수록 복잡도는 낮아짐 – 높은 K값
- 하지만, Underfitting일 수도 있음.

5. Which of the following statements are true?

(a) Nearest Centroids algorithm requires higher fitting time than k-NN.

A. True

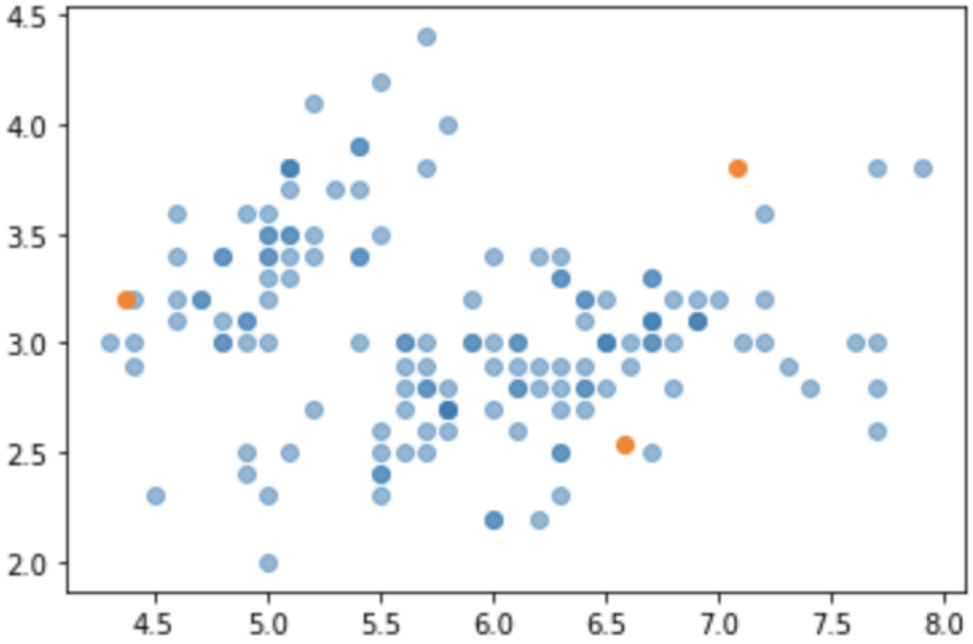
B. False

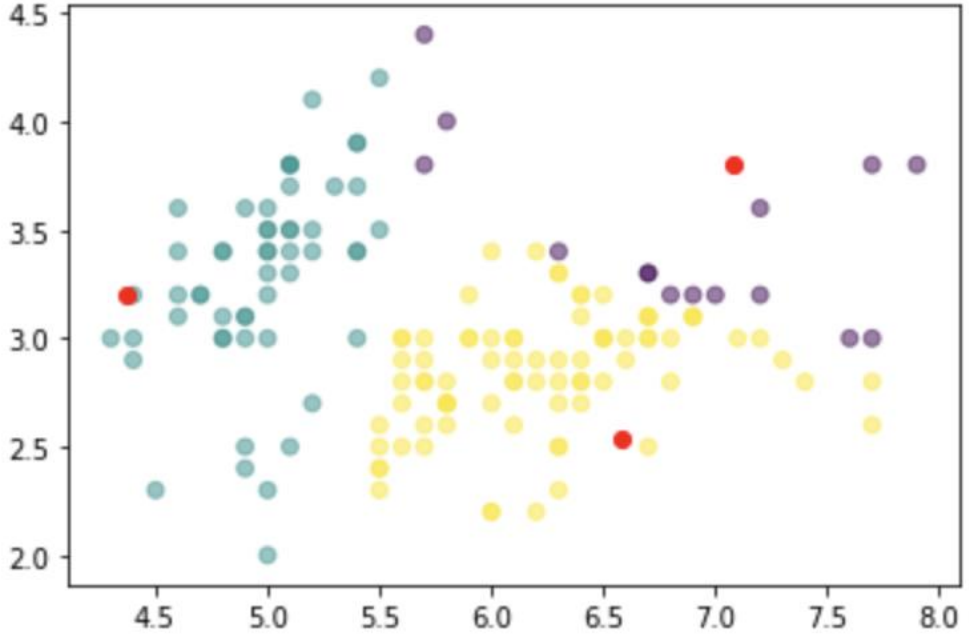
(b) Nearest centroids algorithm requires higher prediction time than k-NN.

A. True

B. False

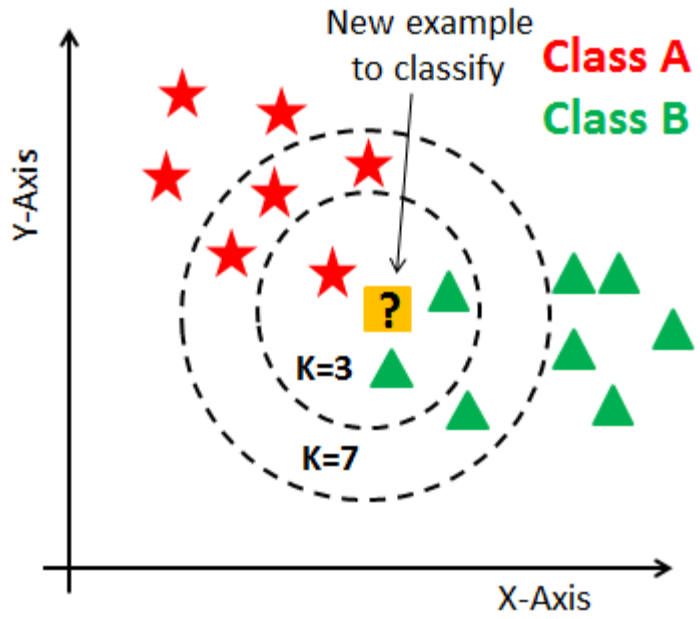
Nearest Centroid – K means Clustering





- 점 K개를 정해서 유클리드 거리를 통해 가까운 곳의 class를 정함
- 즉, fit하는 시간은 오래 걸리나 predict하는 시간은 짧음

KNN



- 새로운 샘플이 들어왔을 때 근처의 K개를 보고 class를 정함
- 즉, fit하는 시간이 필요 없음. k means에 비해 predict 느림

6. Which of the following statements are true?

   (a) Accuracy of 0.75 on the training and 0.73 on the test set is a likely sign of underfitting.

   A. True

   B. False

   (b) Accuracy of 0.95 on the training set and 0.63 on the test set is a likely sign of overfitting.

   A. True

   B. False

Overfitting vs Underfitting

- 보통 학습을 진행하는 경우 Train data와 Test data를 나눔.
- 실제의 경우 새로운 데이터가 들어왔을 때 잘 예측을 해야함.
  즉, 학습에 사용되지 않은 데이터가 필요 == Test data
- 학습 후 Train data에는 잘 예측하나 Test data에 대해 예측을 잘 못하는 경우 Overfitting
- 학습 후 Train data에 대해 잘 예측하지 못하는 경우 Underfitting으로 학습이 완료되지 않았다고 볼 수 있음.

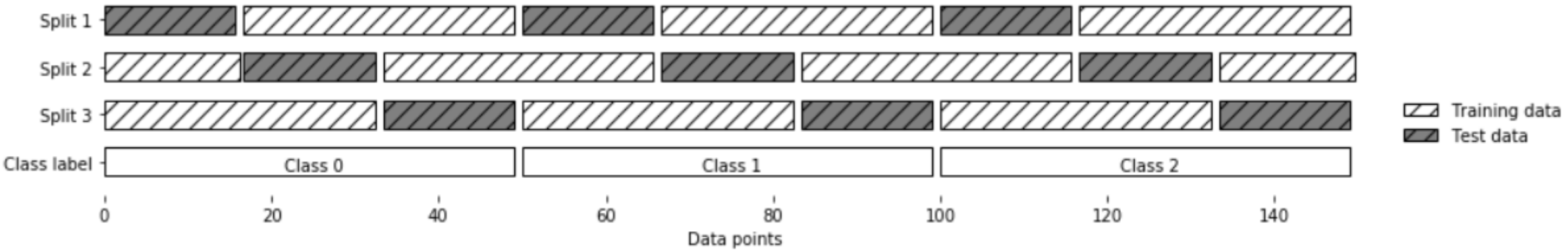


Figure 3: Illustration of a cross-validation strategy.

7. Which type of cross-validation is shown in Figure 3?

A. Standard cross-validation

B. Stratified cross-validation

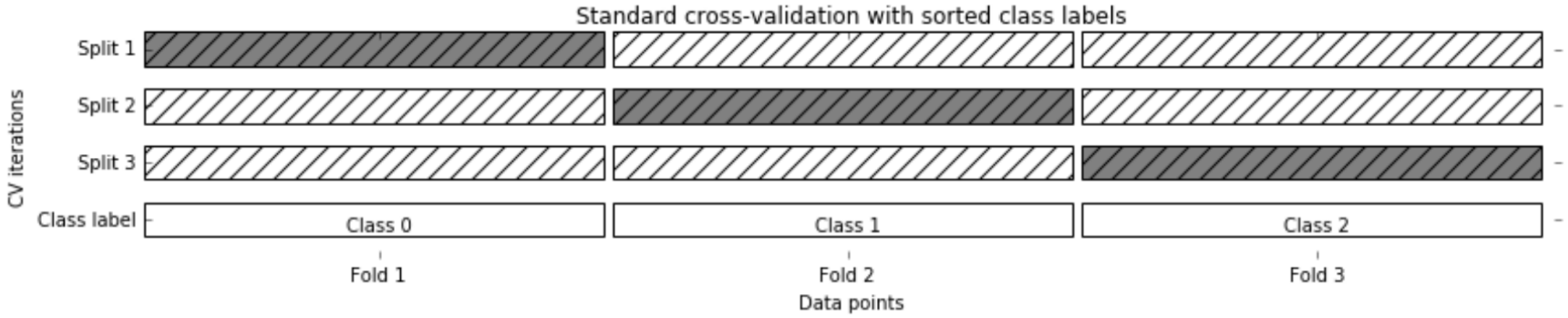
C. ShuffleSplit cross-validation

D. GroupKFold cross-validation

Cross validation

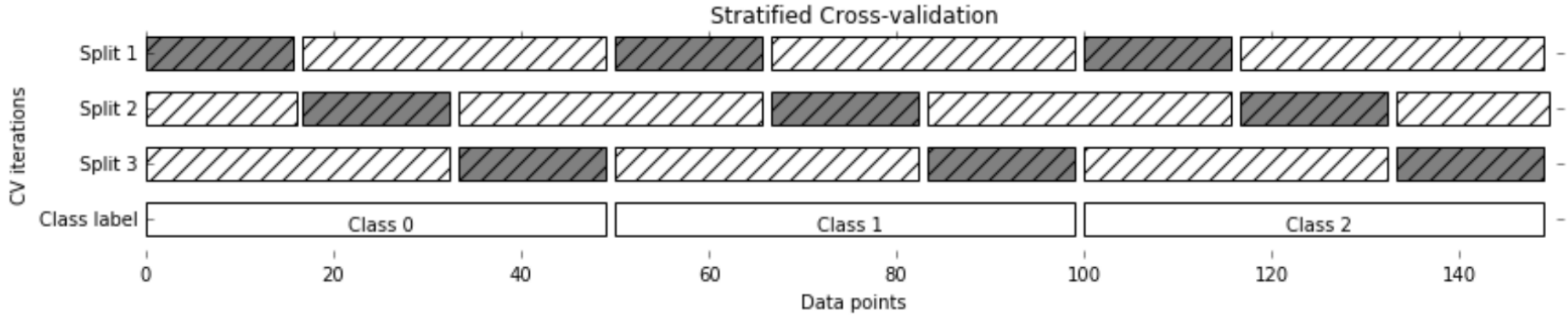
- train test 데이터를 나누는 경우 한쪽에 데이터가 치우치는 경우도 생길 수 있음 -> 이러한 경우를 없애기 위해, 일반화를 더 잘하기 위해 사용
- 다양한 전략이 존재함.

Cross validation - Standard



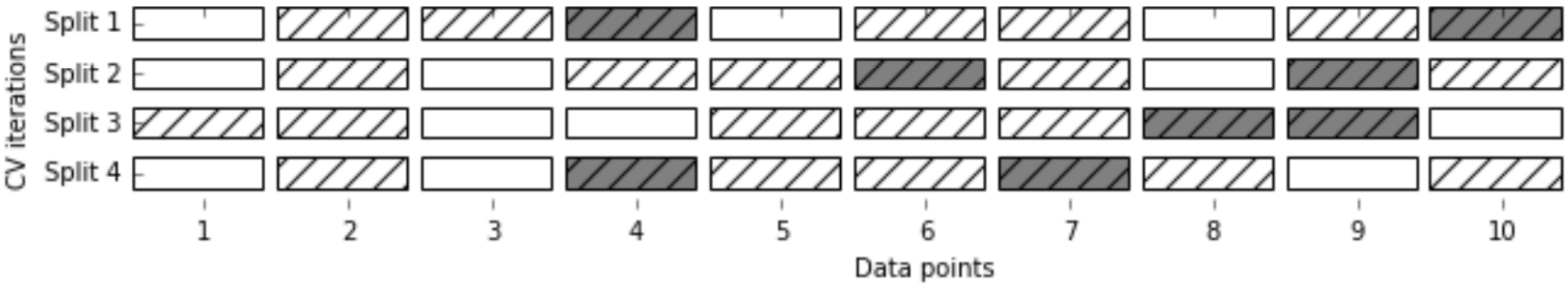
Class 하나를 test데이터로 나머지 Class를 train으로 사용.
단순하게 자른 경우 데이터가 편향되므로 좋지 않음

Cross validation - Stratified



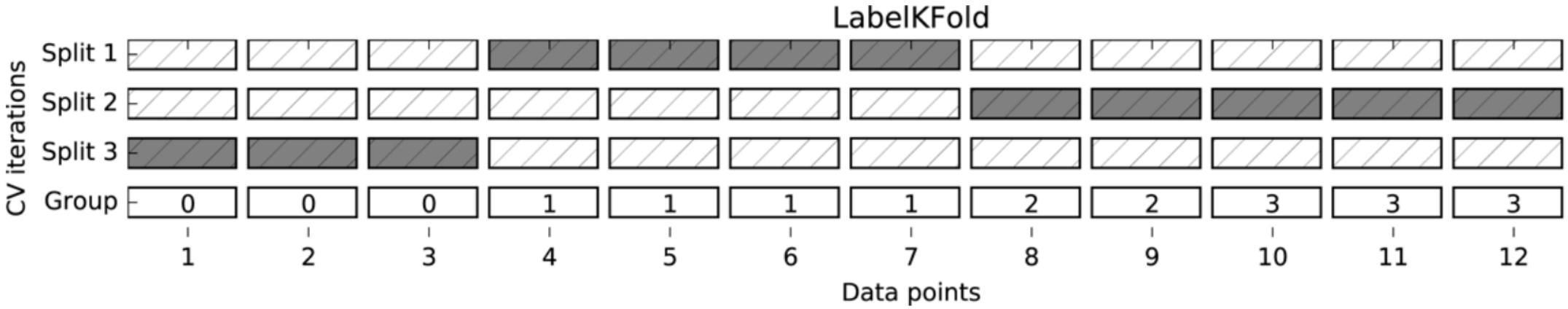
기본적으로 사용되는 방법
각 Class별 1/n 만큼 잘라서 train test 데이터를 구성

Cross validation - ShuffleSplit



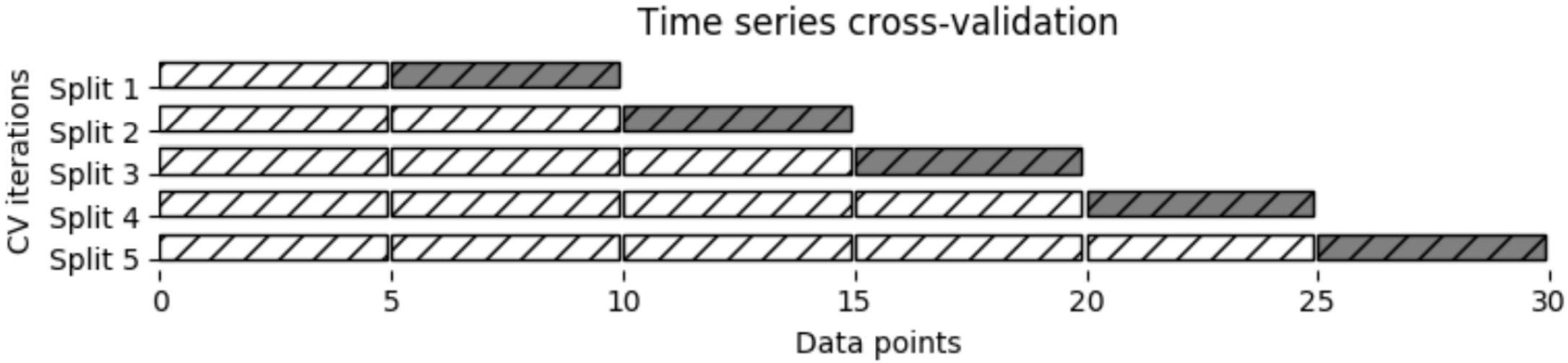
데이터 전체를 n만큼 자르고 랜덤하게 뽑음.
Train_size = 5, test_size=2, n_splits=4인 경우가 위의 그림

Cross validation – GroupKFold



데이터 안에 매우 연관된 그룹이 있는 경우, 떨어뜨리면 안되는 데이터의 경우 그룹별로 묶어서 CV를 진행

Cross validation – TimeSeriesSplit



시계열데이터의 경우 앞에서부터 잘라가는 것이 일반적

8. You are deciding on the model to use for your classification problem and you need probability estimates. Which of the following are you going to choose?

   A. Logistic Regression

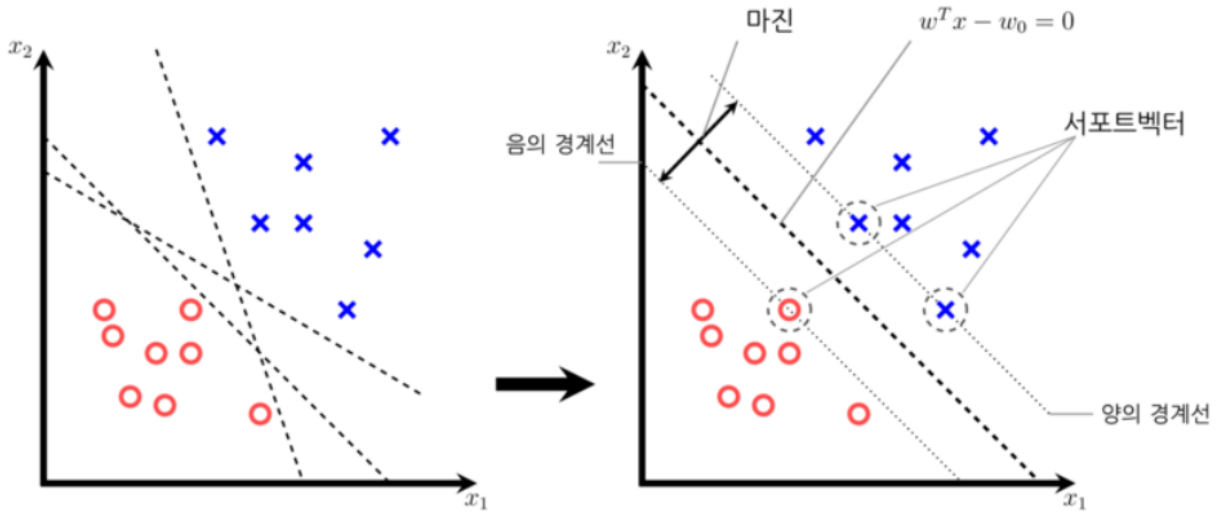
   B. SVM with a linear kernel

   C. SVM with polynomial kernel

   D. SVM with Radial Basis Function (RBF) kernel

SVM

- 확률을 제공하지 않음
- 클래스를 구분할 수 있는 초평면(결정 경계)과 가장 가까운 훈련 샘플인 마진 사이의 거리를 최대로 하는 것



9. Which linear model for classification is given by the following equation?

$$\min_{w \in \mathbb{R}^p} -C \sum_{i=1}^{n} \log \left(\exp \left(-y_i w^\top \mathbf{x}_i \right) + 1 \right) + \|w\|_2^2$$

A. Logistic Regression with L1 penalty

B. Linear SVM with L1 penalty

C. Logistic Regression with L2 penalty

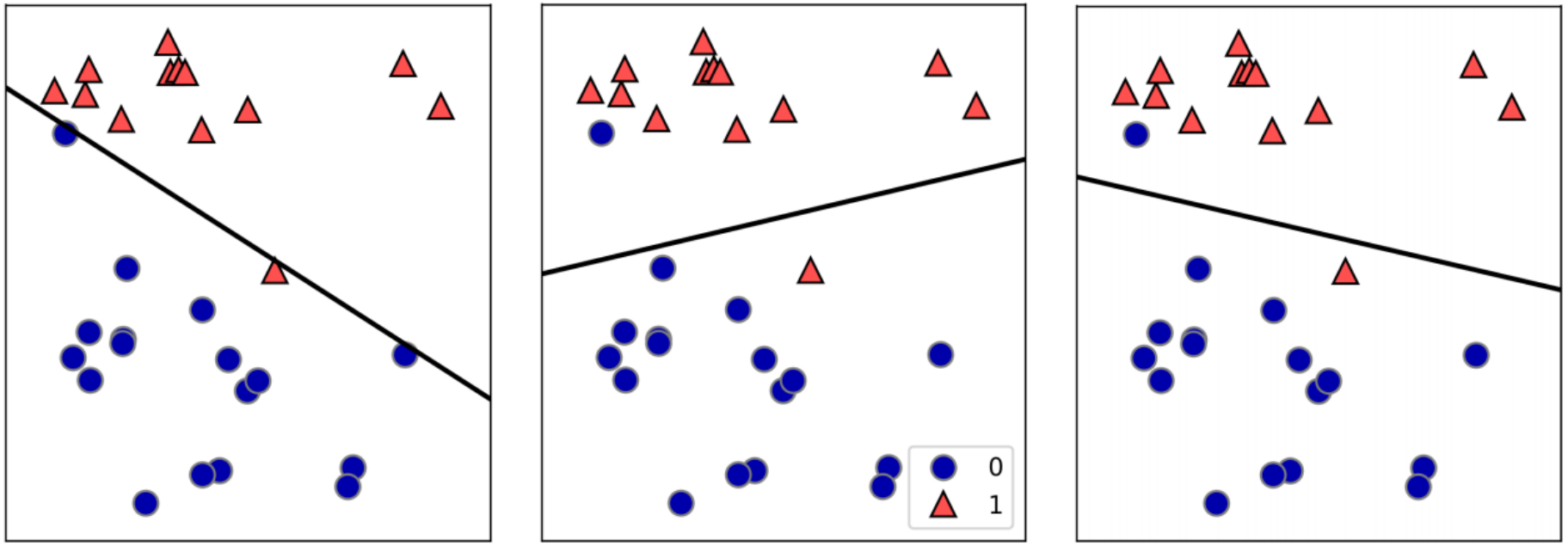
D. Linear SVM with L2 penalty

- Ridge (L2)

$$\min_{w \in \mathbb{R}^p} \sum_{i=1}^{n} ||w^T x_i - y_i||^2 + \alpha ||w||^2$$

- Lasso (L1)

$$\min_{w \in \mathbb{R}^p} \sum_{i=1}^{n} ||w^T \mathbf{x}_i - y_i||^2 + \alpha ||w||_1$$



(a) (b) (c)

Figure 4: Decision boundaries of a linear SVM classifier for different values of parameter C

10. Sort the plots from Figure 4 in increasing order of the parameter C.

A. $a \rightarrow b \rightarrow c$

B. $a \rightarrow c \rightarrow b$

C. $b \rightarrow a \rightarrow c$

D. $b \rightarrow c \rightarrow a$

11. Which of the following statements related to linear SVM are true?

(a) Large value of parameter C corresponds to a high level of regularization.

A. True

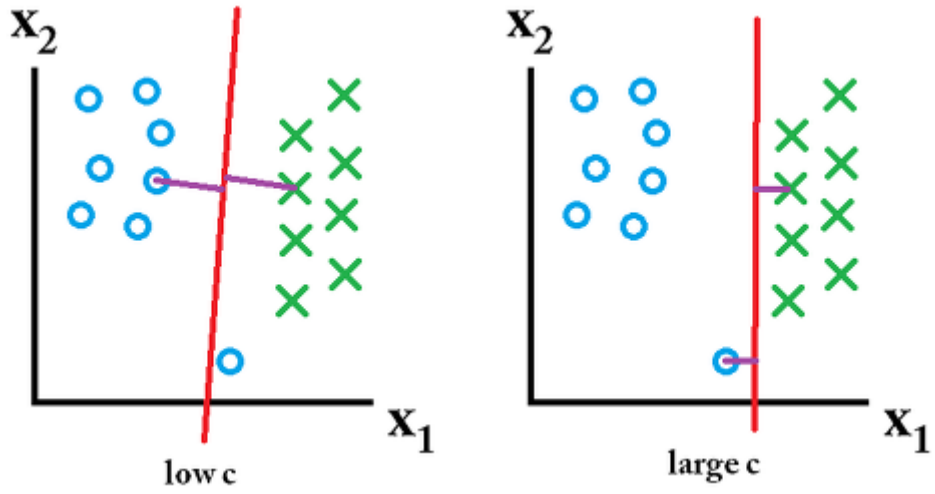
B. False

(b) High level of regularization corresponds to a more complex model.

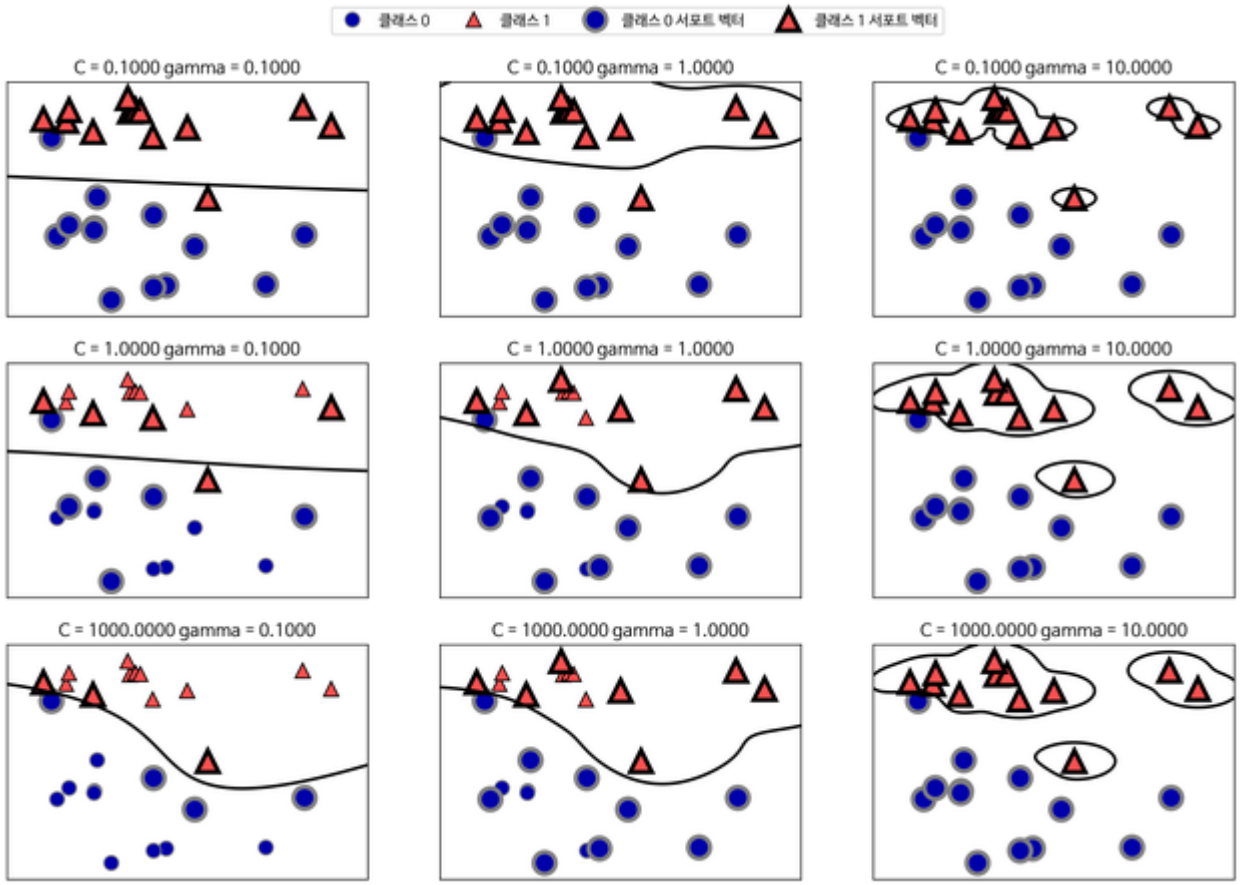
A. True

B. False

SVM에서의 C의 영향



- C는 얼마나 많은 데이터 샘플이 다른 클래스에 놓이는 것을 허용하는지 결정
- C가 낮을수록 좀 더 일반적인 결정 경계를 찾을 수 있음



12. How does StandardScaler transform the data?

A. Transforms features by removing the mean and scaling to unit variance.

B. Transforms features by scaling each feature to a given range.

C. Normalize samples individually to unit norm.

D. None of the above.

12. How does StandardScaler transform the data?

StandardScaler A. Transforms features by removing the mean and scaling to unit variance.

MinMaxScaler B. Transforms features by scaling each feature to a given range.

Normalizer C. Normalize samples individually to unit norm.

D. None of the above.

- Standard : 평균을 0으로 표준편차를 1이 되도록 scaling
- MinMax : 최댓값 1, 최소값 0으로 scaling
- Normalizer : 유클리드 거리가 1이 되도록 scaling

13. Which of the following statements are true?

(a) Model-drive imputation normally involves training of a classifier for missing values.

A. True

B. False

(b) In k-NN imputation the missing values are filled with a value obtained from related items in the dataset.

A. True

B. False

- Model driven : RF와 같은 것으로 주변 분포를 학습

- KNN의 특성 : 결측치에 대해 학습하는 것이 아닌 기존 데이터를 대상으로 예측

14. Which of the following is in general not true for feature selection?

A. Helps interpret the model

B. Helps prevent underfitting

C. Results in faster prediction and training

D. Requires less storage for model and dataset

Feature selection

- 필요한 feature들만 뽑아내는 것
  - 모델 해석에 도움
  - Overfitting을 방지 -> 일반화
  - 학습시간 단축
  - 적은 용량의 모델과 저장소

15. When choosing a feature selection algorithm, which choice we do not have to make?

A. Unsupervised vs. supervised

B. Univariate vs. multivariate

C. NMF vs. manifold learning

D. Model-based vs. not model based

Feature selection

- Unsupervised vs supervised
  - Clustering vs Logistic Regression
- Univariate vs Multivariate
  - 개개의 특성과 target 사이의 특징 vs 특성 사이의 다양한 특징을 고려
- Model based vs not model based
  - KNN vs mean, min, max, 0, etc..
- NMF, manifold는 데이터를 변환하는 것

16. Which of the following is not a motivation for introducing kernel SVMs?

A. Linear models are computationally expensive.

B. Adding non-linear features to linear models can improve their performance.

C. Kernels allow training a classifier in a high dimensional space.

D. Kernels compute distances between items for the expanded feature representation.

Kernel SVM

- 많은 비선형 특성을 추가하면 연산 비용이 커짐. 이러한 비용을 수학적 기교를 통해 학습 -> 선형모델이 연산비용이 큰 것이 아님
- 비선형 특성을 추가하는 것으로 모델을 강력하게 만듦
- 커널을 이용하여 고차원에서 분류기를 학습시킬 수 있음
- 실제로 데이터를 확장하는 것이 아닌 확장된 특성에 대한 데이터의 거리를 계산

17. Parameter γ is controlling width of the Gaussian (RBF) kernel $K(\mathbf{x}, \mathbf{x}') = \exp(-\gamma\|\mathbf{x}-\mathbf{x}'\|^2)$. Which of the following is true for γ?

(a) High γ values result in far reach of a training example.

A. True

B. False

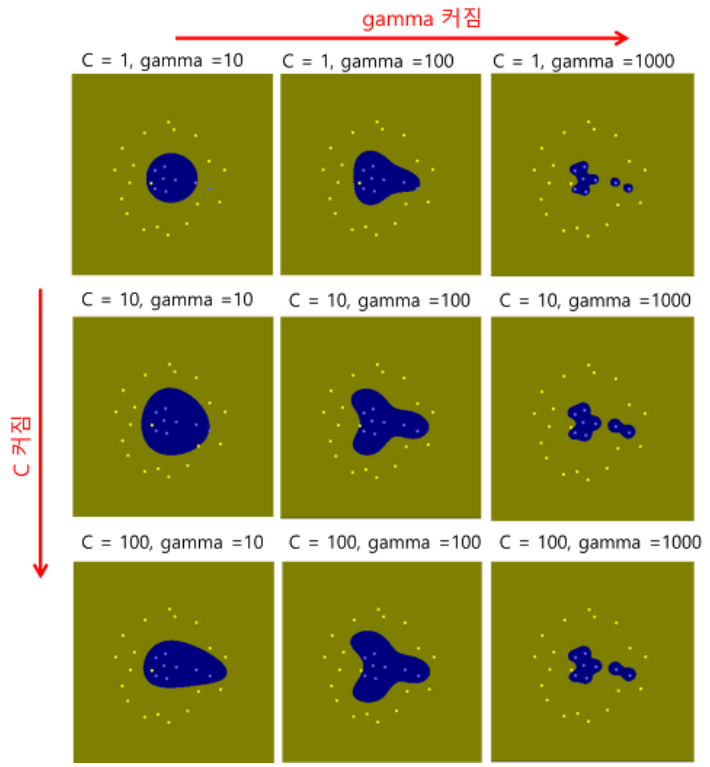
(b) High γ values result in a complex model.

A. True

B. False

커널의 종류

- 다항식 : 특성의 가능한 조합을 지정된 차수까지 모두 계산
- RBF : 가우시안 커널로도 부르며 무한한 특성 공간에 매핑하는 것으로 모든 차수의 다항식을 고려
  - Gamma : 가우시안 커널의 폭을 제어
    - 커질수록 복잡한 모델 생성



18. Which of the following is not a remedy to class imbalance problem when training a model?

A. Sampling the classes proportional to their prior probabilities

B. Undersampling of the majority class

C. Oversampling of the minority class

D. Re-weighting of the loss function

데이터 불균형

- Class가 불균형하게 되어 있는 경우 학습이 제대로 이루어지지 않음
  - 개수가 많은 Class에 대해 Undersampling으로 수를 맞추어 줌
  - 개수가 적은 Class에 대해 Oversampling으로 수를 맞추어줌
  - Loss function을 조정하여 적은 클래스에 더 민감하게 반응하도록 함

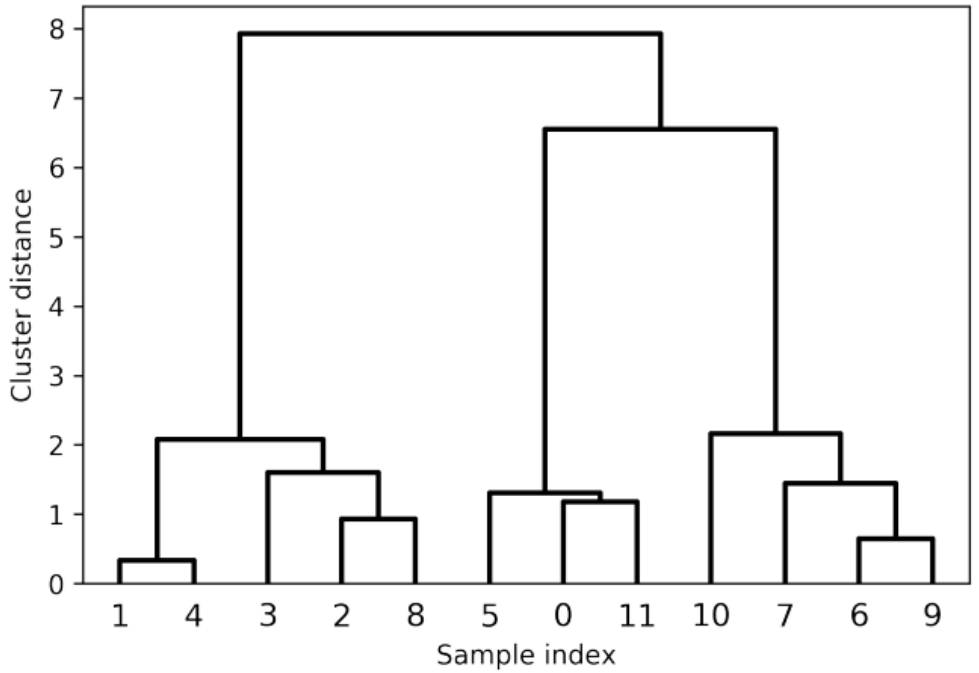


Figure 5: Cluster hierarchy built by agglomerative clustering.

19. The dendrogram in Figure 5 illustrates cluster hierarchy built by the agglomerative clustering. Based on the dendrogram, choose a logical number of clusters.

A. 8

B. 2

C. 4

D. 3

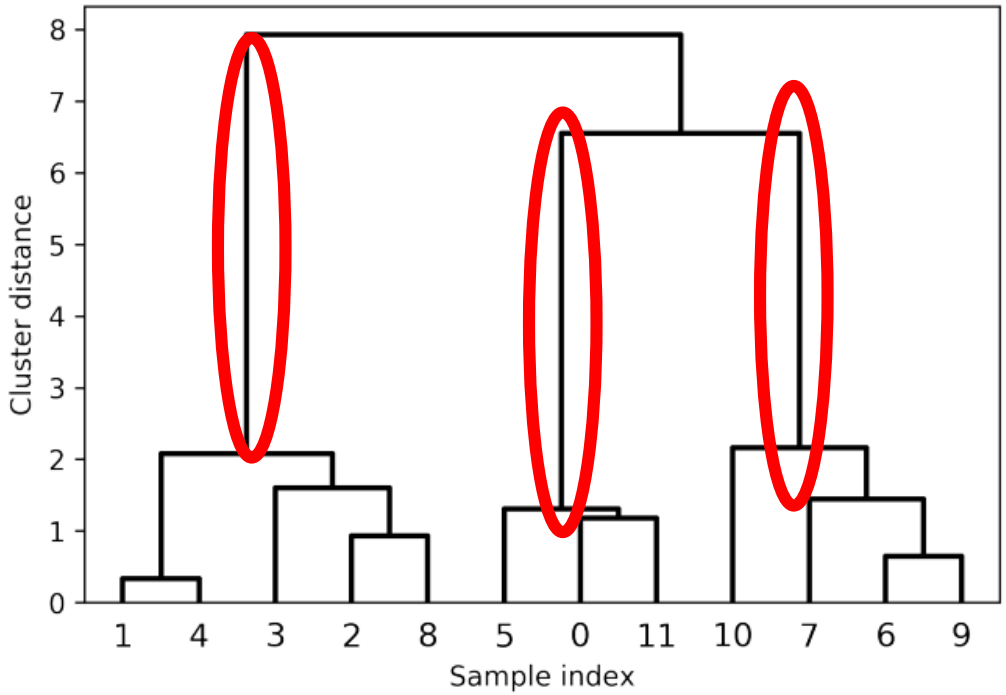


Figure 5: Cluster hierarchy built by agglomerative clustering.

- 길이가 길다는 것은 거리가 멀다고 볼 수 있음.
- 3개로 묶는 것이 덴드로그램에서 논리적으로 나눌 수 있는 클러스터 갯수

20. Which of the following isn't true for t-Distributed Stochastic Neighbor Embedding (t-SNE)?

A. Starts with a random 2D representation for each data point.

B. Makes the points that are close in the original feature space closer.

C. Makes the points that are far apart in the original feature space farther apart.

D. Preserves Mahalanobis distances between the points in the dataset.

t-SNE

- 데이터 사이의 거리를 가장 잘 보존하는 2차원 표현을 찾는 것
- 이웃 데이터 포인트에 대한 정보를 보존
- 각 데이터 포인트를 2차원에 무작위로 표현한 후 원본 특성 공간에서 가까운 포인트는 가깝게, 멀리 떨어진 포인트는 멀어지게 만듦

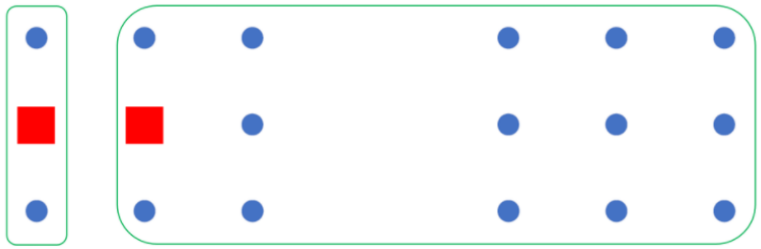
O.Q.1 K-Means Clustering (7 points)

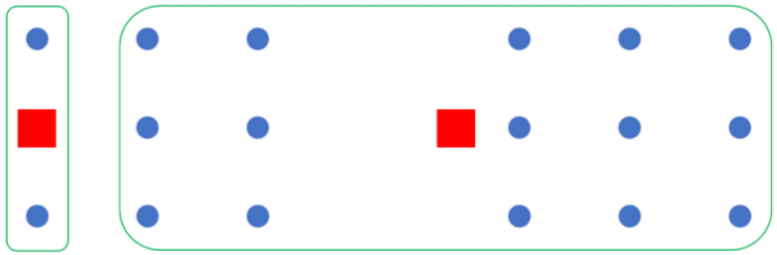
Describe the main steps of k-means clustering algorithm, its advantages, limitations and the properties of resulting clusters. How would you evaluate clustering with and without labels?

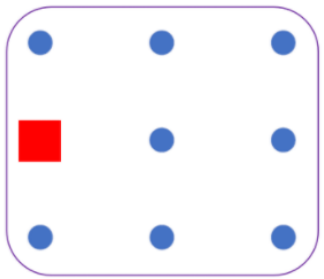
K means Clustering의 주요 단계, 장점, 제한사항, 결과 군집화 특성, 레이블 있는 것과 없는 것에 대한 평가

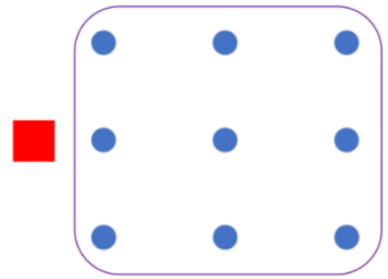
1. 주요 단계
   - K개의 점을 무작위로 초기화 (초기화)
   - 데이터 포인트를 가장 가까운 클러스터 중심에 할당 (포인트 할당)
   - 그 후, 클러스터에 할당된 데이터 포인트의 평균으로 클러스터 중심을 다시 지정 (중심 재계산)
   - 클러스터에 할당되는 데이터 포인트에 변화가 없을 때까지 반복
2. 장점
   - 이해가 쉽고 구현이 쉬우며 비교적 빠름. 대용량 데이터셋에서도 비교적 잘 작동함
3. 제한사항
   - 무작위 초기화를 사용하기 때문에 출력이 계속 바뀜. 클러스터의 모양을 가정하고 있기 때문에 활용범위가 제한적. 클러스터 개수를 지정해야 함.

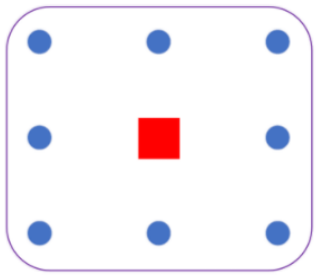
K means

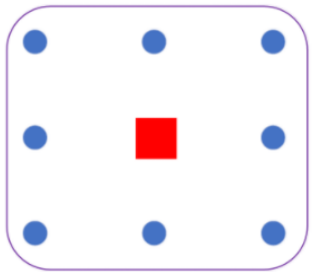












4. 결과 군집화 특성
   - 각 클러스터를 정의하는 것이 중심 하나뿐이므로 클러스터는 둥근 형태

5. 레이블이 있는 경우 evaluate
   - ARI(adjusted rand index) – 가장 최적인 경우 1, 무장위인 경우 0
   - NMI(normalized mutual information)

6. 레이블이 없는 경우 evaluate
   - silhouette coefficient – 최대 점수 1

O.Q.2 Multiclass Classification (6 points)

Multiclass classification means a classification task with more than two classes - e.g., classify a set of images of fruits which may be oranges, apples, or pears. Multiclass classification makes the assumption that each sample is assigned to one and only one label: a fruit can be either an apple or a pear but not both at the same time.

Describe two common approaches to constructing multiclass classifiers. Which approach is faster and why? How are the multiclass classifiers evaluated?

Multiclass classification 구축에 대한 두 가지 일반적 접근 방식. 각 방식 중 어떤 것이 빠르고 그 이유와 어떻게 평가하는가.

1. Multiclass 두 가지 일반적 접근방식
   - One vs Rest : 일대다 방식은 각 클래스를 다른 모든 클래스와 구분하도록 이진 분류 모델을 학습. 클래스의 수만큼 이진 분류 모델 생성
   - One vs One : multiclass를 이진 분류 문제로 바꿈.

2. 빠르고 그 이유
   - One vs Rest가 빠름.
   - 우측과 같이 One vs Rest는 n개의 분류기만 필요

3. 평가 방법
   - Confusion matrix, Micro & Macro F1
   - ROC AUC,

- **One vs Rest** ← Standard

  1v{2, 3, 4}, 2v{1, 3, 4}, 3v{1, 2, 4}, 4v{1, 2, 3}

  n binary classifiers - each on all data

- **One vs One**

  1v2, 1v3, 1v4, 2v3, 2v4, 3v4

  n * (n-1) / 2 binary classifiers - each on a fraction of the data

O.Q.3 Precision, Recall and F1-score (7 points)

True labels `t` as well as the labels predicted by three different classification models, i.e. `c1`, `c2` and `c3` are given below. Calculate precision, recall and f1-score for each model for class "1". Imagine that you are a medical specialist screening patients for a potentially deadly disease, where "1" means that a test is positive. Which model would you choose and why?

```
t = [0, 0, 0, 0, 0, 1, 1, 1, 1, 1]
c1 = [0, 1, 0, 1, 0, 1, 0, 1, 1, 1]
c2 = [1, 0, 1, 1, 0, 1, 1, 1, 1, 1]
c3 = [1, 0, 0, 1, 0, 0, 0, 1, 1, 1]
```

Precision, Recall, F1 score 계산 및 치명적인 질병에 대해 어떤 모델을 선택할 것인가.

Metric

```
t = [0, 0, 0, 0, 0, 1, 1, 1, 1, 1]
c1 = [0, 1, 0, 1, 0, 1, 0, 1, 1, 1]
c2 = [1, 0, 1, 1, 0, 1, 1, 1, 1, 1]
c3 = [1, 0, 0, 1, 0, 0, 0, 1, 1, 1]
```

- Precision
  - TP / TP + FP
- Recall
  - TP / TP + FN
- F1 score
  - 2 * (Precision * Recall) / (Precision + Recall)

음성 클래스	**TN**	**FP**
양성 클래스	**FN**	**TP**
	음성 예측	**양성 예측**

- TP : 실제 True인 클래스를 True라고 예측 (정답)
- FP : 실제 False인 클래스를 True라고 예측 (오답)
- FN : 실제 True인 정답을 False라고 예측 (오답)
- TN : 실제 False인 정답을 False라고 예측 (정답)

Metric

```
t = [0, 0, 0, 0, 0, 1, 1, 1, 1, 1]
c1 = [0, 1, 0, 1, 0, 1, 0, 1, 1, 1]
c2 = [1, 0, 1, 1, 0, 1, 1, 1, 1, 1]
c3 = [1, 0, 0, 1, 0, 0, 0, 1, 1, 1]
```

- Precision
  - 1이라고 예측한 것 중에 실제 1인 개수 / 1로 예측한 개수
- Recall
  - 실제 1중에 맞춘 개수 / 실제 1의 개수

Choice

Model c2, recall이 1이라는 것은 적어도 모든 질병을 1이라고 맞춘 것. 치명적인 질병은 무조건 확진하는 것이 좋다고 생각함

```
print(classification_report(t, c1))
              precision    recall  f1-score

           0       0.75      0.60      0.67
           1       0.67      0.80      0.73
print(classification_report(t, c2))
              precision    recall  f1-score

           0       1.00      0.40      0.57
           1       0.62      1.00      0.77
 print(classification_report(t, c3))
              precision    recall  f1-score

           0       0.60      0.60      0.60
           1       0.60      0.60      0.60
```

시험

201807

1. Which linear model for regression is given by the following equation?

$$\min_{w \in \mathbb{R}^p} \sum_{i=1}^{n} \|w^\top \mathbf{x}_i - y_i\|^2 + \alpha \|w\|_1$$

A. Logistic Regression with L1 penalty

B. Ordinary Least Squares

C. Ridge

D. Lasso

1. Which linear model for regression is given by the following equation?

$$\min_{w \in \mathbb{R}^p} \sum_{i=1}^{n} \|w^\top \mathbf{x}_i - y_i\|^2 + \alpha \|w\|_1$$

A. Logistic Regression with L1 penalty

B. Ordinary Least Squares

C. Ridge

D. Lasso

Figure 1: Illustration of a cross-validation strategy.

2. Which type of cross-validation is shown in Figure 1?

   A. Standard cross-validation

   B. Stratified cross-validation

   C. ShuffleSplit cross-validation

   D. GroupKFold cross-validation

Figure 1: Illustration of a cross-validation strategy.

2. Which type of cross-validation is shown in Figure 1?

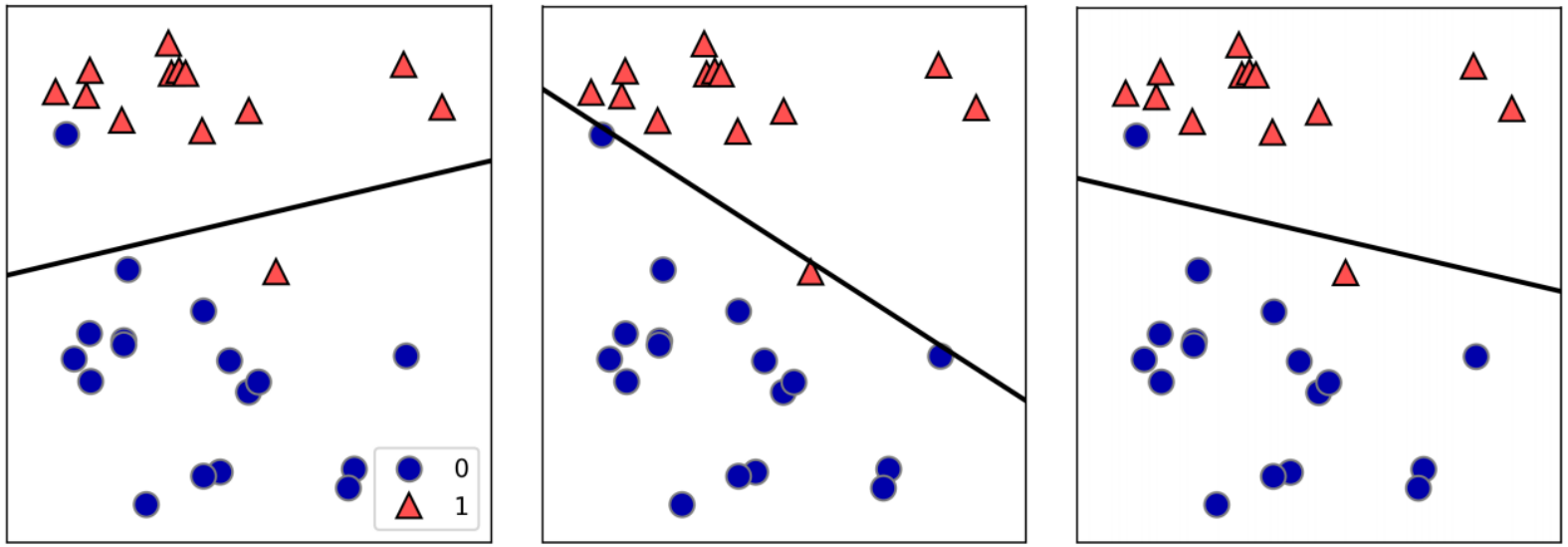
A. Standard cross-validation

B. Stratified cross-validation

C. ShuffleSplit cross-validation

D. GroupKFold cross-validation

3. Sort the plots from Figure 2 according to increasing likelihood of underfitting.

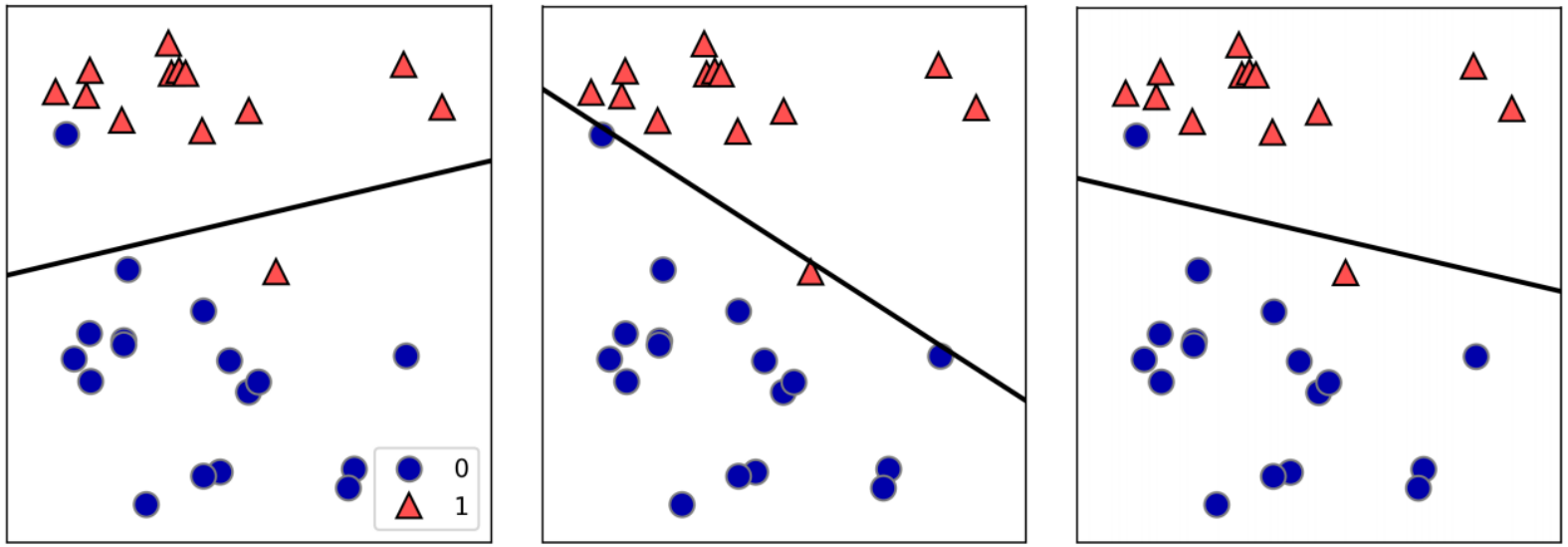


(a) (b) (c)

Figure 2: Decision boundaries of a linear SVM classifier for different values of parameter C

A. $a \rightarrow b \rightarrow c$
B. $a \rightarrow c \rightarrow b$
C. $b \rightarrow c \rightarrow a$
D. $c \rightarrow b \rightarrow a$

3. Sort the plots from Figure 2 according to increasing likelihood of underfitting.



(a) (b) (c)

Figure 2: Decision boundaries of a linear SVM classifier for different values of parameter C

A. $a \rightarrow b \rightarrow c$

B. $a \rightarrow c \rightarrow b$

C. $b \rightarrow c \rightarrow a$

D. $c \rightarrow b \rightarrow a$

4. How does MinMaxScaler transform the data?

   A. Transforms features by removing the mean and scaling to unit variance.

   B. Transforms features by scaling each feature to a given range.

   C. Normalize samples individually to unit norm.

   D. Scale each feature by its maximum absolute value.

4. How does MinMaxScaler transform the data?

A. Transforms features by removing the mean and scaling to unit variance.

B. Transforms features by scaling each feature to a given range.

C. Normalize samples individually to unit norm.

D. Scale each feature by its maximum absolute value.

5. Which of the following statements are true?

   (a) KNN imputation applies k-nearest neighbour classifier to fill in the missing values.

      A. True

      B. False

   (b) The main advantage of model-based feature selection is that it doesn't require labels.

      A. True

      B. False

5. Which of the following statements are true?

(a) KNN imputation applies k-nearest neighbour classifier to fill in the missing values.

A. True

B. False

(b) The main advantage of model-based feature selection is that it doesn't require labels.

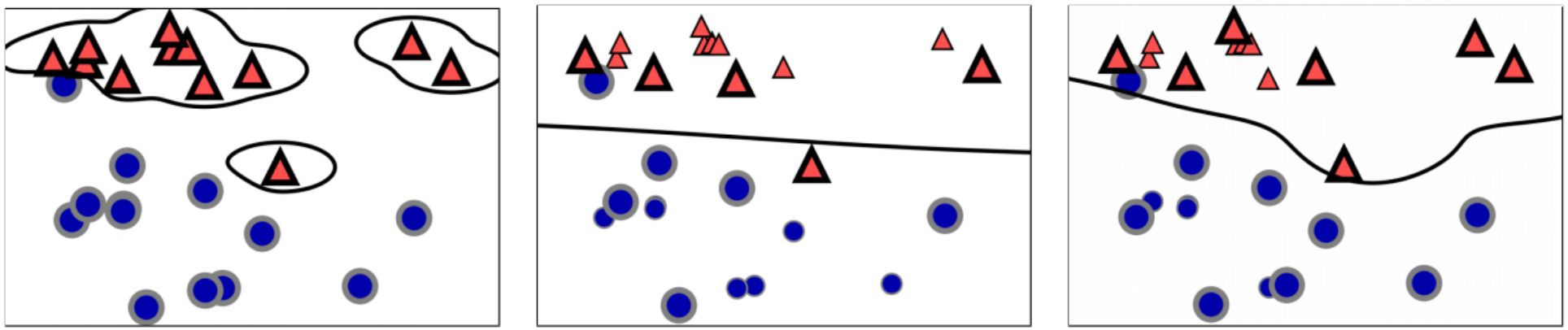
A. True

B. False

KNN imputation : 결측값이 없는 k개의 nearest neighbor를 찾고 평균을 이용하여 missing value를 채움
즉, K-nearest neighbor classifier를 적용하는 것이 아님

model-based feature selection : 모든 가능한 조합을 고려할 수 있음. Label이 필요하지 않은 경우는 unsupervised Feature selection인 경우

6. Sort the plots from Figure 3 in increasing order of parameter γ.



(a) (b) (c)

Figure 3: Decision boundaries and support vectors of an SVM with RBF kernel for different settings of the parameter γ

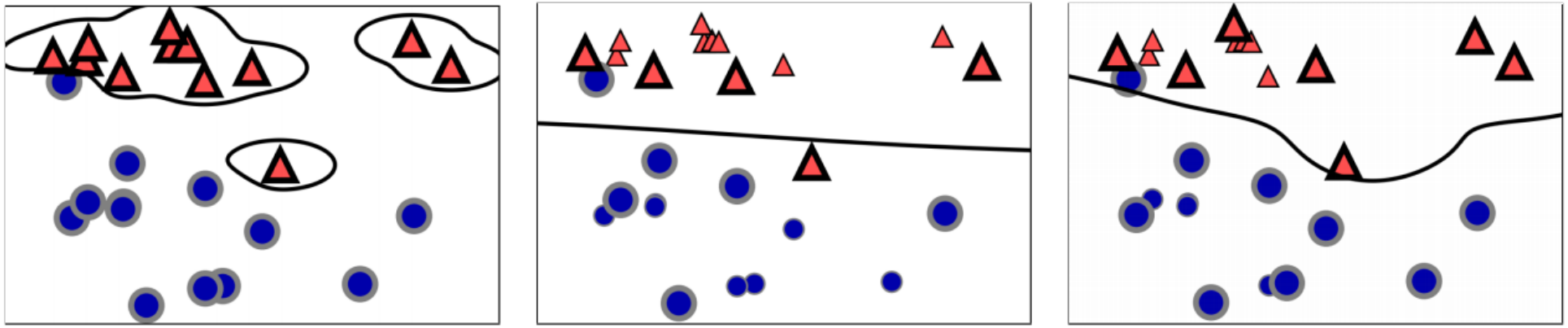
A. $a \rightarrow b \rightarrow c$

B. $a \rightarrow c \rightarrow b$

C. $b \rightarrow c \rightarrow a$

D. $c \rightarrow b \rightarrow a$

6. Sort the plots from Figure 3 in increasing order of parameter γ.



(a) (b) (c)

Figure 3: Decision boundaries and support vectors of an SVM with RBF kernel for different settings of the parameter γ

A. $a \rightarrow b \rightarrow c$

B. $a \rightarrow c \rightarrow b$

C. $b \rightarrow c \rightarrow a$

D. $c \rightarrow b \rightarrow a$

7. Which of the following is not a regression metric?

A. Accuracy score

B. Coefficient of determination

C. Mean squared error

D. Mean absolute error

7. Which of the following is not a regression metric?

A. Accuracy score

B. Coefficient of determination

C. Mean squared error

D. Mean absolute error

8. Which of the following statements are true in context of multiclass classification with imbalanced data?

   (a) Undersampling of the majority class leads to a faster model training.

      A. True

      B. False

   (b) Oversampling of the majority class helps reduce model complexity.

      A. True

      B. False

8. Which of the following statements are true in context of multiclass classification with imbalanced data?

(a) Undersampling of the majority class leads to a faster model training.

A. True

B. False

(b) Oversampling of the majority class helps reduce model complexity.

A. True

B. False

Oversampling은 minority class의 개수를 늘려주는 역할

9. Which of the following statements are true?

   (a) K-Means clustering is popular due to its ability to model complex cluster shapes.

      A. True

      B. False

   (b) Density-Based Spatial Clustering of Applications with Noise (DBSCAN) can be deployed for detecting the outliers.

      A. True

      B. False

9. Which of the following statements are true?

(a) K-Means clustering is popular due to its ability to model complex cluster shapes.

A. True

B. False

(b) Density-Based Spatial Clustering of Applications with Noise (DBSCAN) can be deployed for detecting the outliers.

A. True

B. False

- K-Means는 단순한 둥근 형태로만 나타낼 수 있음
- DBSCAN
  - 클러스터의 개수를 미리 지정하지 않음. 속도 느림.
  - 랜덤한 포인트를 선택하여 포인트로부터 일정 거리 안의 밀집 지역을 찾음. 일정 거리 안에 포인트 수가 min_samples보다 작으면 noise로 지정됨.

10. Which of the following is not true for Non-negative matrix factorization (NMF)?

    A. Data points are composed into positive sums.

    B. Positive weights can be easier to interpret.

    C. NMF computes relatively fast on large datasets.

    D. NMF can be viewed as "soft clustering": each point is positive linear combination of weights.

10. Which of the following is not true for Non-negative matrix factorization (NMF)?

A. Data points are composed into positive sums.

B. Positive weights can be easier to interpret.

C. NMF computes relatively fast on large datasets.

D. NMF can be viewed as "soft clustering": each point is positive linear combination of weights.

NMF (비음수 행렬 분해)

- feature selection의 한 기법.
- Unsupervised
- 음수가 아닌 성분과 계수 값을 찾음.

O.Q.1 Common Machine Learning Techniques (10 points)

Regression, classification and clustering are examples of common machine learning techniques. For the applications listed below indicate which of these techniques is the most applicable and briefly explain your choice.

(a) Predicting whether a patient has a contagious disease.

(b) Estimating the view count an online commercial will get.

(c) Grouping customers of an online store based on their properties.

(d) Predicting the price of real estate in a particular neighbourhood.

(e) Recognizing the type of object.

O.Q.1 Common Machine Learning Techniques (10 points)

Regression, classification and clustering are examples of common machine learning techniques. For the applications listed below indicate which of these techniques is the most applicable and briefly explain your choice.

(a) Predicting whether a patient has a contagious disease.

classification

(b) Estimating the view count an online commercial will get.

regression

(c) Grouping customers of an online store based on their properties.

clustering

(d) Predicting the price of real estate in a particular neighbourhood.

regression

(e) Recognizing the type of object.

classification

O.Q.2 Model Complexity and Accuracy (10 points)

Using the example of linear SVM model given by (2), explain the relation between accuracy and model complexity. Your answer should include at least discussion of the following concepts:

- Generalization, underfitting and overfitting
- How are they related to the number of features?
- What is the role of hyperparameter C?

$$\min_{w \in \mathbb{R}^p} C \sum_{i=1}^{n} \max\left(0, 1 - y_i w^\top \mathbf{x}_i\right) + \|w\|_1 \tag{2}$$

SVM의 정확도와 모델 복잡성 사이의 관계. 일반화, 언더피팅, 오버피팅, feature의 수 와의 관계, 하이퍼파라미터 C의 역할

- SVM의 정확도와 모델 복잡성 사이의 관계.
  - 일반적으로 기본 성능이 매우 좋은 알고리즘
  - 하지만 시간/공간 복잡도가 매우 높음

- 일반화, 언더피팅, 오버피팅,
  - Margin을 최대로 하는 것을 기본으로 하기 때문에 일반화 성능이 우수하다고 볼 수 있음.
  - Overfitting과 underfitting을 피하기 위해 hyperparameter C를 잘 조절해야 한다.
  - C가 작아질 수록 underfitting, C가 커질수록 overfitting이 될 수 있다. 적정한 C를 찾는것이 중요.

- feature의 수 와의 관계
  - Feature의 수가 적어도 복잡한 결정경계를 만들 수 있지만 많으면 제대로 동작하지 않음.

- 하이퍼파라미터 C의 역할
  - C의 값이 작을 수록 제약이 큰 모델을 만들고 각 데이터 포인트의 영향력이 작음.
  - C의 값이 커질 수록 각 데이터 포인트들이 모델에 큰 영향을 주며 결정경계를 휘어서 정확하게 분류

O.Q.3 Evaluation Metrics for Classification (10 points)

True labels `t` as well as the labels predicted by three different classification models, `c1`, `c2` and `c3` are given below. Calculate **precision**, **recall** and **f1-score** for each model for class "1". Imagine that you are an aerospace engineer trying to predict whether a new rocket system will be effective in intercepting unmanned aerial vehicles in the extremely expensive field trials, where "1" means effective. Which prediction model would you choose and why?

```
 t = [0, 0, 0, 0, 0, 1, 1, 1, 1, 1]
c1 = [1, 1, 1, 1, 0, 0, 1, 1, 1, 1]
c2 = [1, 1, 0, 0, 0, 1, 0, 1, 0, 1]
c3 = [0, 0, 0, 0, 0, 0, 0, 1, 0, 0]
```

```
 t = [0, 0, 0, 0, 0, 1, 1, 1, 1, 1]
c1 = [1, 1, 1, 1, 0, 0, 1, 1, 1, 1]
c2 = [1, 1, 0, 0, 0, 1, 0, 1, 0, 1]
c3 = [0, 0, 0, 0, 0, 0, 0, 1, 0, 0]
```

```
>>> print(classification_report(t, c1))
              precision    recall  f1-score   support

           0       0.50      0.20      0.29         5
           1       0.50      0.80      0.62         5

    accuracy                           0.50        10
   macro avg       0.50      0.50      0.45        10
weighted avg       0.50      0.50      0.45        10

>>> print(classification_report(t, c2))
              precision    recall  f1-score   support

           0       0.60      0.60      0.60         5
           1       0.60      0.60      0.60         5

    accuracy                           0.60        10
   macro avg       0.60      0.60      0.60        10
weighted avg       0.60      0.60      0.60        10

>>> print(classification_report(t, c3))
              precision    recall  f1-score   support

           0       0.56      1.00      0.71         5
           1       1.00      0.20      0.33         5

    accuracy                           0.60        10
   macro avg       0.78      0.60      0.52        10
weighted avg       0.78      0.60      0.52        10
```

O.Q.4 Agglomerative Clustering (10 points)

Describe the main steps of agglomerative clustering algorithm. What are its main advantages and drawbacks? Sketch the diagram that could be used to visualize possible clusterings and determine "logical" number of clusters? How to evaluate clustering with and without labels?

- 병합 군집화 알고리즘 주요 단계
  - 각 포인트를 하나의 클러스터로 지정.
  - 종료 조건을 만족할 때까지 비슷한 두 클러스터를 합침.
- 장단점
  - 장점 : 클러스터 개수를 지정하지 않아도 됨.
  - 단점 : 복잡한 형상은 구분하지 못함.
- 가능한 군집을 시각화하고 논리적 군집 수를 결정하는 데 사용할 수 있는 다이어그램
  - Dendrogram
- 레이블이 있거나 없는 클러스터링 평가방법
  - 레이블 O : ARI, NMI
  - 레이블 X : Silhouette coefficient

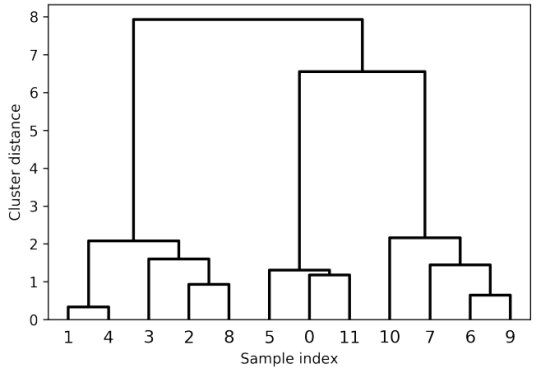


Figure 5: Cluster hierarchy built by agglomerative clustering.

O.Q.5 Supervised Feature Selection (10 points)

What is feature selection and what are the most important reasons for deploying it? Briefly describe the main families of supervised feature selection methods, including the discussion of their main advantages and disadvantages.

- Feature selection이란 무엇이며 사용하는 이유는 무엇?
  - 새로운 feature를 만드는 방법이 많아 데이터의 차원이 원본 이상으로 증가할 수 있음
  - 위의 경우 모델은 복잡해지고 overfitting이 생길 수 있음
  - 가장 유용해보이는 featur를 선택하여 모델을 간단하게 만들며 일반화 성능을 높임
  - 모델 해석에 도움
  - Overfitting을 방지 -> 일반화
  - 학습시간 단축
  - 적은 용량의 모델과 저장소

- 장단점
  - 장점 : 관련 없는 데이터 제거 가능. 중복성 제거 가능. Overfitting 방지, 학습시간 단축.
  - 단점 : feature selection이 오래 걸릴 수도 있음. 변수간 상관관계 고려 어려움.
- 종류
  - Univariate Statistic : 개개의 특성과 target 사이에 통계적 관계를 계산
  - Model-based selection : machine learning model들을 사용하여 특성의 중요도를 평가
  - Iterative selection : 하나씩 추가하거나 하나씩 제거하며 다양한 모델을 만드는 것