Python으로 머신러닝 입문

세션장: 구은아, 김은기

KMeans

KMeans 클래스 및 하이퍼파라미터

sklearn.cluster.KMeans

class sklearn.cluster.KMeans(*n_clusters=8, *, init='k-means++', n_init=10, max_iter=300, tol=0.0001, verbose=0, random_state=None, copy_x=True, algorithm='auto'*) [source]

https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

KMeans 클래스 및 하이퍼파라미터

파라미터 이름	설명
n_clusters	몇 개의 군집으로 군집화할 것인지 결정 디폴트는 8
init	첫번째 군집의 중심점을 결정할 방식 디폴트는 k-means++
max_iter	최대 반복 횟수 디폴트는 300

실루엣 계수

실루엣 계수를 구하는 클래스

sklearn.metrics.silhouette_score

sklearn.metrics.silhouette_score(*X*, *labels*, *, *metric='euclidean'*, *sample_size=None*, *random_state=None*, ***kwds*) [source]

https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html

sklearn.metrics.silhouette_samples

sklearn.metrics.silhouette_samples(*X*, *labels*, *, *metric='euclidean'*, ***kwds*) [source]

https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_samples.html

평균 이동

평균이동 클래스 및 하이퍼파라미터

sklearn.cluster.MeanShift

class sklearn.cluster.MeanShift(*, *bandwidth=None*, *seeds=None*, *bin_seeding=False*, *min_bin_freq=1*, *cluster_all=True*, *n_jobs=None*, *max_iter=300*) [source]

https://scikit-learn.org/stable/modules/generated/sklearn.cluster.MeanShift.html

파라미터 이름	설명
bandwidth	KDE의 대역폭. 클수록 과적합 방지 효과가 증가함. 디폴트는 None

GMM

GMM 클래스 및 하이퍼파라미터

sklearn.mixture.GaussianMixture

class sklearn.mixture.GaussianMixture(*n_components=1, *, covariance_type='full', tol=0.001, reg_covar=1e-06, max_iter=100, n_init=1, init_params='kmeans', weights_init=None, means_init=None, precisions_init=None, random_state=None, warm_start=False, verbose=0, verbose_interval=10*) [source]

https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html

파라미터 이름	설명
n_components	몇 개의 가우시안 분포로 군집화할 것인지 결정 디폴트는 1

과제 안내

과제: Customer Segmentation

마지막 과제!

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom

1) 이 데이터에는 문제점이 몇가지 있습니다. 이를 해결하여 데이터를 깔끔하게 정제해봅시다. (꼭 retail_df에서 진행해주세요.)
2) RFM 기반 분석이므로 주문 데이터를 활용하여 최근 상품 구입일에서 오늘까지 기간(Recency), 상품 구매 횟수(Frequency), 총 구매 금액(Monetary) 열을 만들어야 하지만 과제 파일에 미리 만들어놓았습니다. cust_df를 사용하여 분석을 해주세요.
3) cust_df에는 왜곡된 분포를 가진 열이 있습니다. StandardScaler를 사용하여 이를 해결해봅시다.
4) 실루엣 계수를 사용하여 가장 적절한 군집 개수를 구해봅시다. (Kmeans를 사용해주세요)

수고하셨습니다!
과제 열심히 하시고 다음 주에 뵈어요~