Analiza skupień metodą k-medoids (PAM)

Zofia Galla,Ramon Dyzman,Jakub Konieczny

Uniwersytet Adama Mickiewicza

434684, 415366, 470607

28 czerwca 2021

K-medoids kod

```
# initialize medoids (at random)
medoids = initialize_medoids(num_medoids=num_clusters, data=df_scaled)

# assign data points to the medoids
assignments = assign_points_to_medoids(data=df_scaled, medoids=medoids)

# fit
new_medoids = reassign_medoids(data=df_scaled, assignments=assignments, initial_medoids=medoids)
while not is_finished(old_medoids=medoids, new_medoids=new_medoids):
    medoids = new_medoids
    new_medoids = reassign_medoids(data=df_scaled, assignments=assignments, initial_medoids=medoids)
```

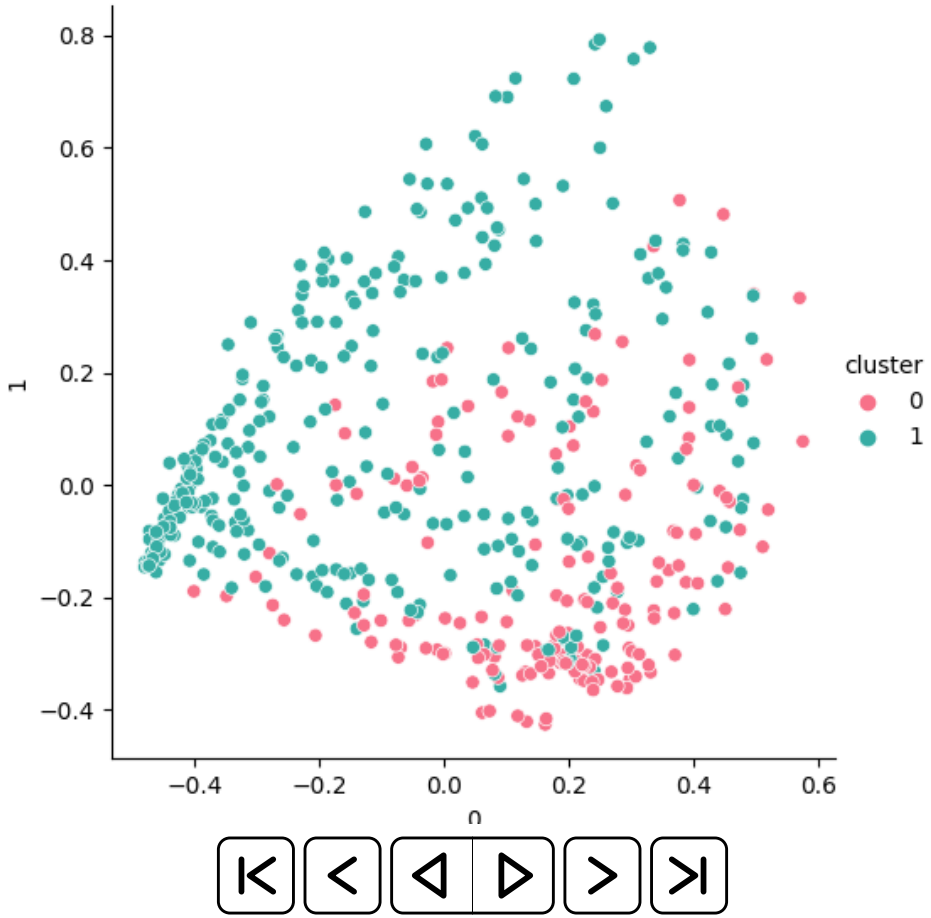
K-medoids kod

```python
def initialize_medoids(num_medoids, data):
    return [data.iloc[idx] for idx in choice(len(data), size=num_medoids, replace=False)]


def assign_points_to_medoids(data, medoids):
    return [np.argmin([distance_vec2vec(point[1], medoid) for medoid in medoids]) for point in data.iterrows()]


def distance_vec2vec(a, b) -> np.float64:
    return sum([(abs(a[i] - b[i]) ** 2) for i in range(len(a))])
```

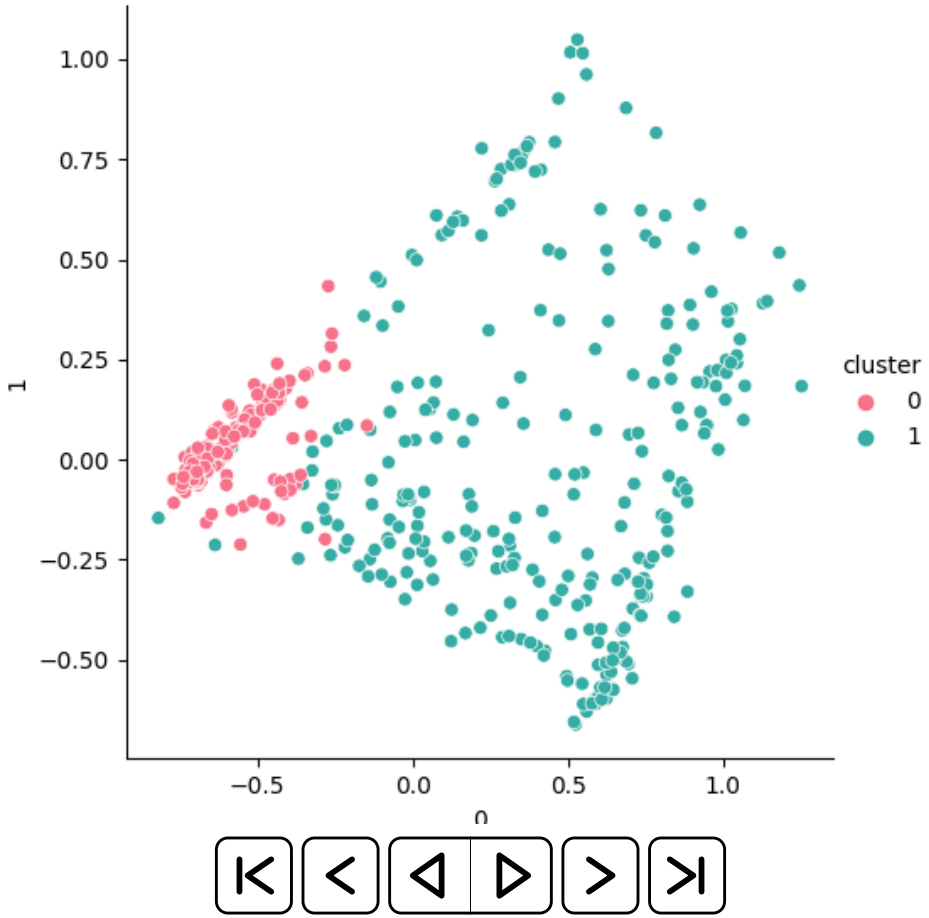
K-medoids kod

```
def reassign_medoids(data, assignments, initial_medoids):
    new_medoids = []
    for idm, medoid in enumerate(initial_medoids):
        new_medoid = medoid
        medoid_score = sum([distance_vec2vec(medoid, x[1]) if assignments[idx] == idm else 0
                            for idx, x in enumerate(data.iterrows())])
        for point in data.iterrows():
            point_score = sum(sum([distance_vec2vec(point, x[1]) if assignments[idx] == idm else 0
                                   for idx, x in enumerate(data.iterrows())]))
            if medoid_score > point_score:
                new_medoid = point
        new_medoids.append(new_medoid)
    return new_medoids


def is_finished(old_medoids, new_medoids):
    return set([tuple(om) for om in old_medoids]) == set([tuple(nm) for nm in new_medoids])
```

K-medoids z regularyzacją



K-medoids z normalizacją



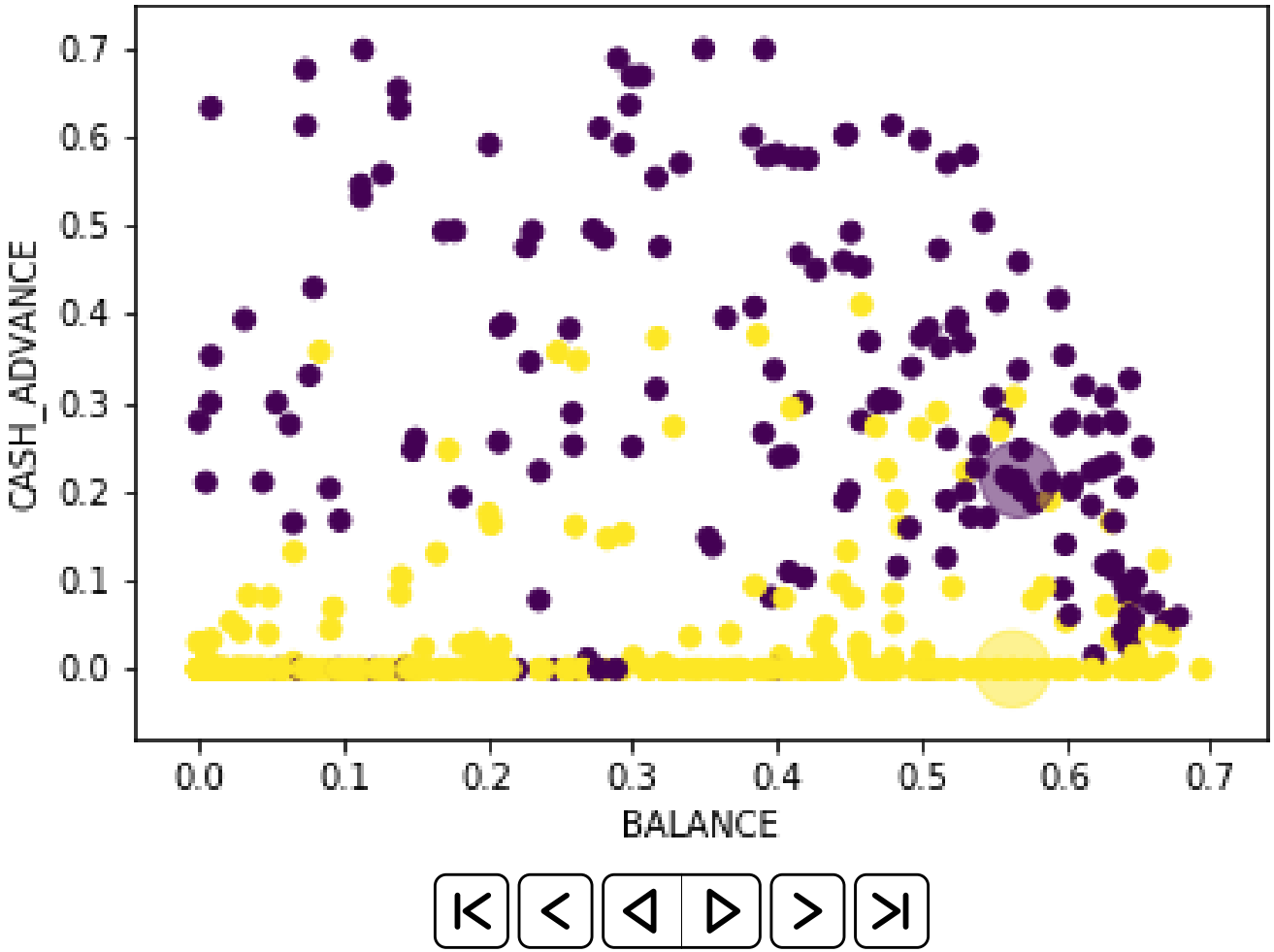
K-Medoids: Zachowania posiadaczy kart kredytowych

Opisuje zachowanie około 9000 aktywnych posiadaczy kart kredytowych w ciągu ostatnich 6 miesięcy zapisane przy pomocy 18 parametrów.
My wybraliśmy 500 rekordów, a na potrzeby wizualizacji 3 najbardziej istotne parametry za pomocą algorytmu PCA (najbardziej istotne składowe zmiennych):

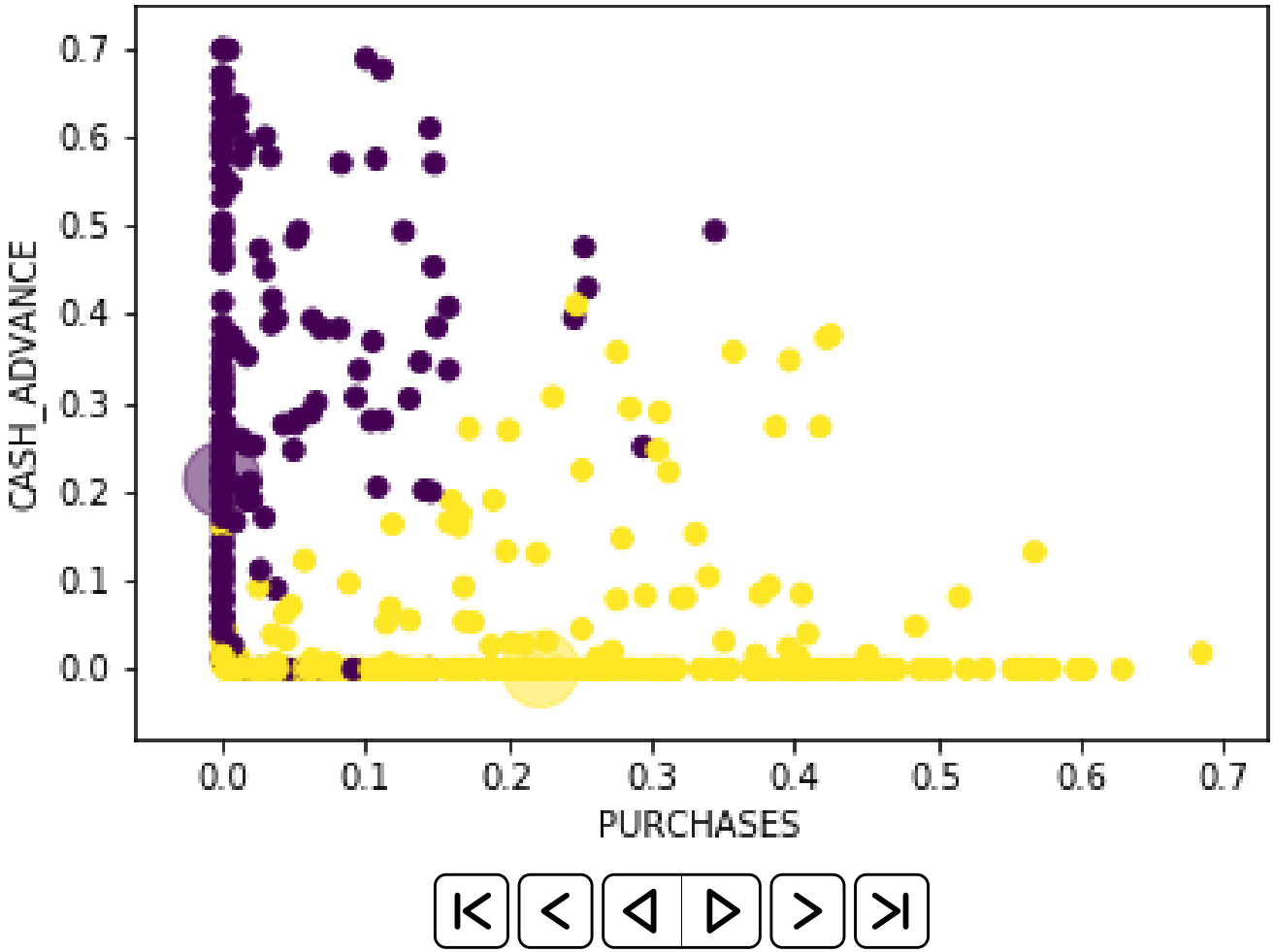
- Saldo na koncie
- Liczba wypłat z bankomatu za pomocą karty kredytowej
- Liczba płatności kartą kredytową

https://www.kaggle.com/arjunbhasin2013/ccdata

Saldo a wypłaty z bankomatu



Liczba płatności a wypłaty z bankomatu



Liczba płatności a saldo