Кластеризация данных, целевая функция

***- Обычный подход***

 Oбучение сети сводится к подбору параметров сети (весовых коэффициентов *w* и смещений *b* для нейронов сети). Параметры сети подбираются так, чтобы минимизировать отклонение результата работы сети $F(w,b)$ от контрольного набора данных:

$$\min\_{w,b}F(w,b)$$

Функцию $F(w,b)$ обычно называют функцией потерь или функцией ошибок. В качестве такой функции можно использовать сумму квадратичных отклонений

$$F\left(w, b\right)=\frac{1}{N}\sum\_{n}^{}\left(x^{(n)}-y^{(n)}\right)^{2}$$

или какую либо другую функцию:



 Существенным моментом здесь является использование известного контрольного набора данных. В этом случае задача обработки данных сводится к классификации данных (обучение с учителем в терминологии машинного обучения).

 Кластеризации данных (обучение без учителя) не требуется контрольный набор. Кластеризация данных выявляет и использует внутреннюю структуру набора данных и на основе этого осуществляет фрагментацию данных.

 Применение нейросетей и использование стандартных библиотек (Keras, tensorflow, …) для кластеризации данных можно свести к подбору соответствующей функции, минимизация которой приводит к кластеризации исходного набора данных.

функция ошибок, функция потерь, целевая функция

[NumPy ops](https://keras.io/api/ops/numpy/#numpy-ops) - набор функций для работы с тензорами в представлении tensorflow.

import keras.ops as ko

Пример предопределённой функции MSE:

def mse\_func(y\_true, y\_pred):

 return ko.mean(ko.square(y\_pred-y\_true), axis = -1)

Использование центров кластеров – аналог K-means

$$F\left(w, b\right)=min\left(\frac{1}{N}\sum\_{n}^{}\sum\_{i}^{}\left(x\_{i}^{\left(n\right)}-C\_{i}\right)^{2}\right)$$

здесь *n*-номер объекта во входном потоке данных, *i*-номер кластера.

Координаты центра кластера находим как среднее соотв. координат объектов образующих кластер:

$$X\_{i}=\frac{1}{n\_{c}}\sum\_{C\_{i}}^{}x^{(n)}$$

Для этого случая:

Нейросеть:

model=Sequential()

model.add(Dense(30,input\_dim=DIM\_N,activation='relu'))

model.add(Dense(DIM\_N\*CLUST\_N,activation=None))

- выходные нейроны выдают координаты центров кластеров, для двумерного случая и трёх кластеров – массив *Cen* в программе:

$$\left(X\_{0},Y\_{0},X\_{1},Y\_{1},X\_{2},Y\_{2}\right)$$

Создаём представление входного набора данных data:

$$\left[\left[x^{(0)},y^{(0)}\right],\right.$$

$$\cdots $$

$$\left[x^{(i)},y^{(i)}\right],$$

$$\cdots $$

$$\left.\left[x^{(N)},y^{(N)}\right]\right]$$

DataN = np.repeat(np.expand\_dims(data, axis=1), CLUST\_N, axis=1)

$$\left[\left[x^{(0)},y^{(0)}\right],\left[x^{(0)},y^{(0)}\right],\left[x^{(0)},y^{(0)}\right],\right.$$

$$\cdots $$

$$\left[x^{(i)},y^{(i)}\right],\left[x^{(i)},y^{(i)}\right],\left[x^{(i)},y^{(i)}\right],$$

$$\cdots $$

$$\left.\left[x^{(N)},y^{(N)}\right],\left[x^{(N)},y^{(N)}\right],\left[x^{(N)},y^{(N)}\right]\right]$$

def my\_err\_func(y\_true, y\_pred):

 global DataN

 yy = ko.array(ko.split(ko.mean(y\_pred,axis=0),3))

 z1 = ko.square(ko.subtract(DataN,yy))

 z2 = ko.sum(z1,axis=2)

 zz = ko.amin(z2,axis=1)

 return zz

По функциям:

$$y\\_pred=$$

$$\left[\left[X\_{0}^{(0)},Y\_{0}^{(0)},X\_{1}^{(0)},Y\_{1}^{(0)},X\_{2}^{(0)},Y\_{2}^{(0)}\right],\right.$$

$$\cdots $$

$$\left[X\_{0}^{(i)},Y\_{0}^{(i)},X\_{1}^{(i)},Y\_{1}^{(i)},X\_{2}^{(i)},Y\_{2}^{(i)}\right],$$

$$\cdots $$

$$\left.\left[X\_{0}^{(N)},Y\_{0}^{(N)},X\_{1}^{(N)},Y\_{1}^{(N)},X\_{2}^{(N)},Y\_{2}^{(N)}\right]\right]$$

ko.mean(y\_pred,axis=0)

$$\left[\overbar{X}\_{0},\overbar{Y}\_{0},\overbar{X}\_{1},\overbar{Y}\_{1},\overbar{X}\_{2},\overbar{Y}\_{2}\right],$$

ko.split(ko.mean(y\_pred,axis=0),3)

$$\left[\left[\overbar{X}\_{0},\overbar{Y}\_{0}\right],\left[\overbar{X}\_{1},\overbar{Y}\_{1}\right],\left[\overbar{X}\_{2},\overbar{Y}\_{2}\right]\right]$$

yy = ko.array(ko.split(ko.mean(y\_pred,axis=0),3))

* Создаёт тензор в представлении tensorflow

z1 = ko.square(ko.subtract(DataN,yy))

z2 = ko.sum(z1,axis=2)

$$\left[\cdots \right.$$

$$\left[\left[\left(x^{(i)}-\overbar{X}\_{0}\right)^{2}+\left(y^{(i)}-\overbar{Y}\_{0}\right)^{2}\right],\left[\left(x^{(i)}-\overbar{X}\_{1}\right)^{2}+\left(y^{(i)}-\overbar{Y}\_{1}\right)^{2}\right],\left[\left(x^{(i)}-\overbar{X}\_{2}\right)^{2}+\left(y^{(i)}-\overbar{Y}\_{2}\right)^{2}\right]\right],$$

$$\left.\cdots \right]$$

* перечень расстояний каждой точки до центров кластеров

zz = ko.amin(z2,axis=1)

- расстояния до центра ближайшего кластера для каждой точки

Последняя строка программы вычисляет значение искомой целевой функции:

$$f\left(w, b\right)=\sum\_{n}^{}\sum\_{i}^{}\left(x^{(n)}-C\_{i}\right)^{2}$$

 Шаги итерации нейросети, аналогичные обучению при обычном подходе, подбирают значения $\left(X\_{0},Y\_{0},X\_{1},Y\_{1},X\_{2},Y\_{2}\right)$, соответствуют поиску минимума целевой функции:

$$F\left(w, b\right)=min\left(\frac{1}{N}\sum\_{n}^{}\sum\_{i}^{}\left(x\_{i}^{\left(n\right)}-C\_{i}\right)^{2}\right)$$

Полный текст соответствующей программы:

import numpy as np

import matplotlib.pyplot as plt

from sklearn import datasets

from sklearn import preprocessing

import random

POINT\_N = 300

DIM\_N = 2

CLUST\_N = 3

data, targ, Center = datasets.make\_blobs(n\_samples=POINT\_N, centers=CLUST\_N, n\_features=DIM\_N, center\_box=(-10,10), random\_state=0, return\_centers=True)

plt.figure(figsize=(7,7))

plt.scatter(data[:,0], data[:,1], c='black', marker='o')

plt.scatter(Center[:,0], Center[:,1], c='red', marker="\*", s=100)

plt.show()



# ---------- Neural Network ----------

import warnings

warnings.filterwarnings("ignore")

import tensorflow as tf

import tensorflow.keras.backend as tfb

import keras

import keras.ops as ko

from keras.models import Sequential

from keras.layers import Dense

model=Sequential()

model.add(Dense(30,input\_dim=DIM\_N,activation='relu'))

model.add(Dense(DIM\_N\*CLUST\_N,activation=None))

DataN = np.repeat(np.expand\_dims(data, axis=1), CLUST\_N, axis=1)

def my\_err\_func(y\_true, y\_pred):

 global DataN

 yy = ko.array(ko.split(ko.mean(y\_pred,axis=0),3))

 z1 = ko.square(ko.subtract(DataN,yy))

 z2 = ko.sum(z1,axis=2)

 zz = ko.amin(z2,axis=1)

 return zz

model.compile(loss=my\_err\_func, optimizer='adam', metrics=['accuracy'], run\_eagerly=True)

model.summary()

Cen = np.array([data[random.randint(0,POINT\_N-1)] for i in range(CLUST\_N)])

out = np.zeros ((POINT\_N,CLUST\_N), dtype=int)

history = model.fit(data, out, batch\_size=data.shape[0], epochs=300, verbose=1)



plt.plot(history.history['loss'], color='b', label="validation loss")

plt.xlabel("Number of Epochs")

plt.ylabel("Loss")

plt.legend()

plt.show()

pred=model.predict(data)

cen = np.array(np.split(np.mean(pred, axis=0), 3))

dd = np.sum(np.square(np.subtract(DataN,cen)), axis=2)

res = np.argmin(dd,axis=1)

colors = ['blue', 'green', 'purple', 'brown', 'black']

plt.figure(figsize=(7,7))

for i in range(POINT\_N):

 plt.scatter(data[i:,0], data[i:,1], c=colors[res[i]], marker='o')

plt.scatter(Center[:,0], Center[:,1], c='cyan', marker="\*", s=100)

plt.scatter(cen[:,0], cen[:,1], c='red', marker="\*", s=100)

plt.show()



- результат для 300 точек



- результат для 1000 точек