11/30

1. GNN - 학습데이터 라벨링

However, if you are working on DOCKGROUND dataset, it should be simpler. All the correct decoys typically with large numbers like 365130, and those decoys named with r-l-1.pdb to r-l-100.pdb are typically incorrect structures.

숫자가 1 ~ 100까지는 incorrect -> Y = 0.으로 설정
숫자 100 초과는 correct -> Y = 1.으로 설정

해당 문의에 대한 답변으로 이렇게 하면 된다고 확인을 받았습니다.

For your data labeling, I believe it's correct since Dockground is the easiest dataset to label and your labeling should be correct.

1. GNN - 데이터 불균형

데이터 불균형에 관한 질문을 보냈습니다.
같이 간단한 코드를 보냈더니 기존에 하던 대로 이런식으로 하면 된다고 하셔서
그대로 진행을 해보았습니다.

단순히 correct를 더해주는 걸로 균형을 맞췄습니다.

```
train_listfiles=[]
for i in range(int(len(train_listfiles_incorrect) / len(train_listfiles_correct))):
        train_listfiles += train_listfiles_correct
train_listfiles+=train_listfiles_incorrect
```

1. GNN - 데이터 샘플링

데이터를 샘플링 한다는 코드를 구현하고자
학습 데이터 배열을 random.shuffle로 섞고 [:100]식으로
100개만 슬라이싱 해서 사용했습니다.

그런데 이렇게 진행한 것이 말이 안되는게
그럼 각 epoch마다 100개의 임의의 데이터만 학습을 하는 잘못된 방향이라서

shuffle을 통해 모든 데이터를 학습하는 방향으로 변경하였습니다.

1. GNN - 데이터 샘플러 코드 및 구현

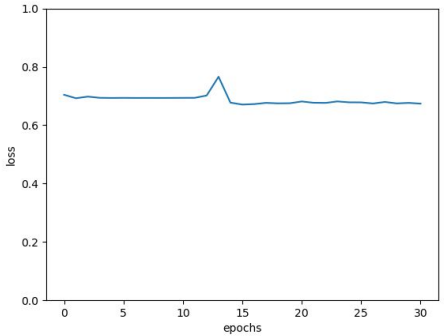
추가로 sampler에 대한 코드도 공개를 해주셨는데, 그럼에도 코드에 검색해도 나오지 않는 개인 함수들과 코드들이 여전히 많았습니다.

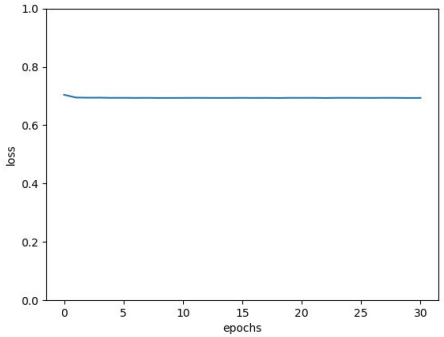
그런데 그대로하고 shuffle만 한다면 작동할 것이라고 하셔서 그대로 진행하였습니다.

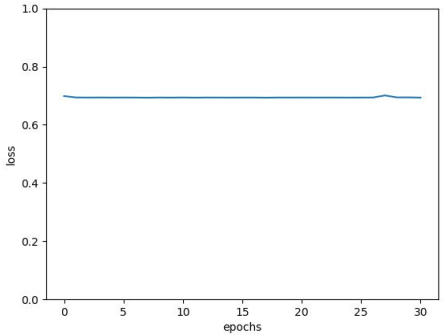
Though I implemented by data sampler, I still thought your implementation in this simple way should work if you set shuffle=True for the dataset.

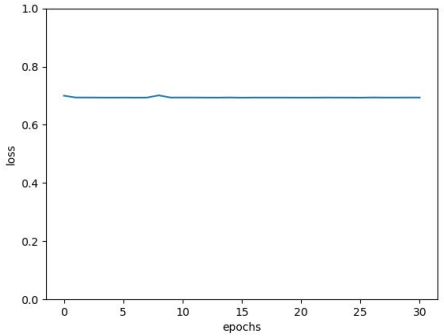
```
class Data_Sampler(Sampler):

    def __init__(self, weights, num_samples,
replacement=True):
        weights = np.array(weights) / np.sum(weights)
        self.weights = weights
        self.num_samples = num_samples
        self.replacement = replacement

    def __iter__(self):
        # return iter(torch.multinomial(self.weights,
self.num_samples, self.replacement).tolist())
        retval = np.random.choice(len(self.weights),
self.num_samples, replace=self.replacement,
p=self.weights)
        return iter(retval.tolist())

    def __len__(self):
        return self.num_samples
```

1. GNN - 학습결과_loss









0.1 0.01
0.03 0.05

batch size 2에다가
learing rate를 여러번 시도하고 있는데 loss가 처음부터 0.6부근에서 머물고 있고, 변화가 없어서 다른 learning rate로 시도중 입니다.

그래서 다양한 batch size와, learning rate 조합을 실행하여 최적의 파라미터를 뽑아야할 것 같습니다.

1. GNN - residue에 관한 문의

one-hot vector embedding으로 해서 다른 feature을 넣고 해도 된다는데,

Atom features can be extended for better representation. We have also used GNN-DOVE for other purposes. It's totally fine to add more features. But please either use one-hot vector embedding as features or normalized features.