13기 정규세션

ToBig's 12기 박진혁

과제 추가 설명!

과제 설명

Assignment 1

1. Assignment1은 위에서 언급 되었던 Multiclass SVM을 직접 구현하시는 것입니다. 기본적으로 사이킷 런에 있는 SVM은 멀티클래스 SVM을 지원합니다. 그러나 과제에서는 그것을 쓰면 안됩니다! 아이리스 데이터는 총 세 개의 클래스가 있으므로 이 클래스를 one-hot 인코딩 한 뒤, 각각 binary SVM을 트레이닝하고 이 결과를 조합하여 multiclass SVM을 구현하는 것입니다.

2. 위에서 말했듯 기본적으로 one vs one, one vs rest방법이 있으며 어떤 것을 구현하든 자유입니다. 만약 투표결과 동점이 나온경우(예를 들어 각각의 SVM의 결과가 A vs B C의 경우 A로 판별, B vs A C의 결과 B로 판별, C vs A B의 경우 C로 판별한 경우 투표를 통해 class를 결정할 수 없음) decision_function을 활용하시거나, 가장 개수가 많은 클래스를 사용하시거나 랜덤으로 하나를 뽑거나 하는 방법 등을 이용해 동점자인 경우를 판별해주시면 됩니다. 공식 문서를 보면 사이킷런이 어떤 방법으로 구현했는지가 글로 나와 있으므로 참조하셔도 무관합니다.

3. assignment.ipynb 파일에는 제가 iris 데이터를 로드하고 iris 데이터를 one hot인코딩 한 뒤 svm 모델로 각각 class가 1인지 아닌 지, class가 2인지 아닌지 각각 구분할 수 있는 모델 부분까지 적어놓았습니다. 또한 decision function을 호출해서 사용하는 예시도 하나 넣어 놓았으니 참고하시면 됩니다. 제가 one vs one로 구현해주세요 one vs rest로 구현해주세요라고 적어놓은 셀이 있습니다. 아무거나 구현해주세요. 개인적으로 one vs rest가 더 구현하기 쉬울것으로 생각되며, 모르는 부분은 언제든 질문해주세요! 생각보다 코드가 길지 않고 어렵지 않습니다.

One vs rest

One vs Rest 방법 예시)
데이터의 클래스가 A, B, C가 있다.

우리는 총 3개의 머신을 만든다.
1. A인지 아닌지 구분해주는 머신
2. B인지 아닌지 구분해주는 머신
3. C인지 아닌지 구분해주는 머신

자 새로운 데이터가 들어왔다. 그리고 우리는 이 데이터를 1,2,3번 머신에 돌렸다.
1번 머신은 A라고 했고, 2번머신은 B가 아니라고 했고, 3번머신은 C가 아니라고 했다.
그렇다면 이 세 개의 결과를 종합해서 우리는 새로 들어온 데이터가 A라고 판별하는 것이다.

그렇다면 머신 세 개가 1번 머신은 A가 맞다고 했고, 2번 머신은 B가 맞다고 했고, 3번 머신은 C가 맞다고 했다.
이럴 때는 어떻게 판별해야할까?

과제설명

One vs One

One vs One 방법 예시)
데이터의 클래스가 A, B, C가 있다.

우리는 총 3개의 머신을 만든다.
1. A인지 B인지 구분해주는 머신
2. B인지 C인지 구분해주는 머신
3. C인지 A인지 구분해주는 머신

자 새로운 데이터가 들어왔다. 그리고 우리는 이 데이터를 1,2,3번 머신에 돌렸다.
1번 머신은 A라고 했고, 2번머신은 C라고 했고, 3번머신은 A라고했다.
그렇다면 이 세 개의 결과를 종합해서 우리는 새로 들어온 데이터가 A라고 판별하는 것이다.

Assignment 2

1. Assignment2는 제가 anomaly-detection 데이터셋(캐글에 올라와있음)을 드립니다. 이것은 해당 결제가 사기인지 아닌 지 판별하는 데이터셋이며, 실습코드를 활용하고, 또 본인이 여태 배운 내용을 활용하여 자유롭게 데이터를 가지고 연습해주시면 됩니다.

2. 다만 이 데이터셋은 굉장히 imbalance한 데이터 셋입니다. 실제 사기를 치는 사례가 많지 않으므로 사기인 경우가 전체에 0.17프로밖에 되지 않습니다. 따라서, 그냥 데이터를 트레이닝 시키면 무조건 사기가 아니라고 판별해버릴 가능성이 높습니다. 또한, 그대로 트레이닝을 돌리게 되면 엄청나게 많은 데이터 양 때문에 트레이닝조차 힘들 것입니다. 그런데 실제로 이런 데이터에 대해 트레이닝을 하는 여러 방법들이 있으니 고민해보세요.

Assignmetn2.ipynb 파일에는 제가 데이터를 로드하는 것 까지만 구현해 놓았습니다. 이 이후 부분은 제가 실습코드로 올려드렸던 부분 등을 참조해서 구현하시면 됩니다. 제가 Colab(상당히 좋은 CPU)으로 그냥 시켜봤는데 이게 아예 못 쓸 정도는 아니지만, 그래도 트레이닝이 꽤나 오래 걸리므로, 감안해주시면 되고 답이 잘 안나오면 물어보셔도 좋습니다.

추가과제)

만약 본인이 아직, 모델을 써서 트레이닝하고 accuracy를 측정하고 Hyper parameter를 튜닝하는게 낯설다면, Assignment2 보다는 제가 실습 코드로 드린 부분을 좀 더 연습해 보시는 것을 추천드립니다.

아직 model = SVC(~), Model.predict 등의 method가 낯설고 어렵다 하시는 분들은 실습코드를 제가 새로 드린 데이터 셋(Voice정보로 남자인지 여자인지 구분하는 데이터셋)으로 연습해 보시는 것을 추천드리며, 과제를 다 하지 못하셨더라도 실습코드를 연습한 것을 같이 제출한다면 참작해드리도록 하겠습니다.