3.5多クラス分類の例 ニュース配信の分類

17t4015s
内間けんじ

目次

3.5.0 多クラス分類

前節では２つのクラスに分類する方法の例を示した。

今節では、Reutersのニュース配信を46種類の相互排他的なクラスに分類するネットワークの構築を行う。

→多クラス単一ラベル問題

3.5.1 Reutersデータセット

Reutersデータセットは1986年にReutersによって配信された短いニュース記事とそれらのトピック46種類を集めたもの。

今回は訓練データでは、出現頻度が最も高い10000個の単語に制限し、訓練サンプル8982個、テストサンプル2246個とする。

3.5.2 データの準備

データのベクトル化を行い、リストをを0と1のベクトルに変換する。

例えば、シーケンス[3,5]の場合はインデックス3と5が1でありそれ以外が0である10000次元のベクトルに変換される。

ラベルのベクトル化でも**one-hotエンコーディング**使い、ラベルのインデックスの位置に1、それ以外に0が含まれるようなベクトルに変換する。

補足：もう一つのラベルベクトル化の方法

ラベルをエンコードするもう一つの方法として、そのラベルを整数のテンソルとしてキャストする方法がある。

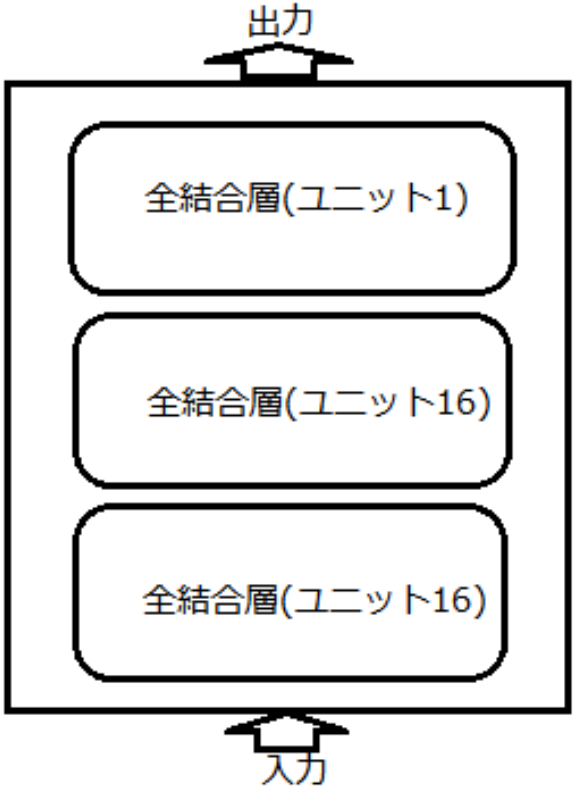
```
y_train = np.array(train_labels)
y_test = np.array(test_labels)
```

このアプローチで変わるのは損失関数の選択だけである。

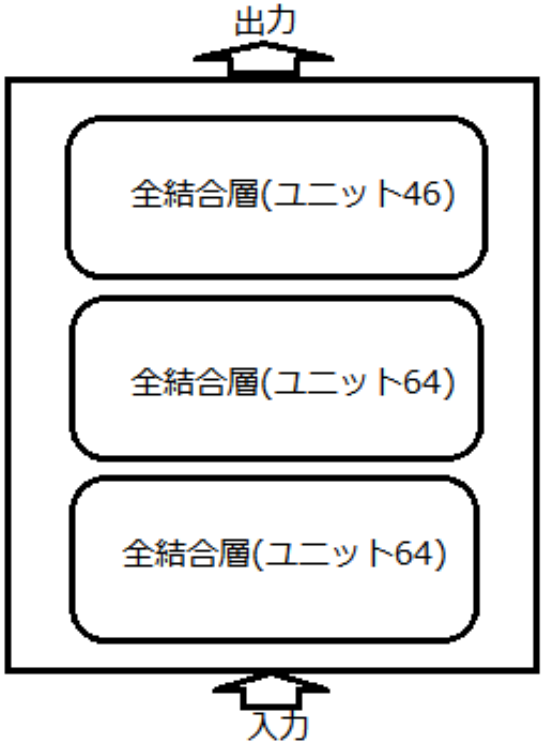
・one-hotエンコーディング→categorical_clossentropy

・ラベルが整数の場合→sparse_categorical_crossentropy

3.5.3 ニューラルネットワークの構築



3.4：2クラスの分類

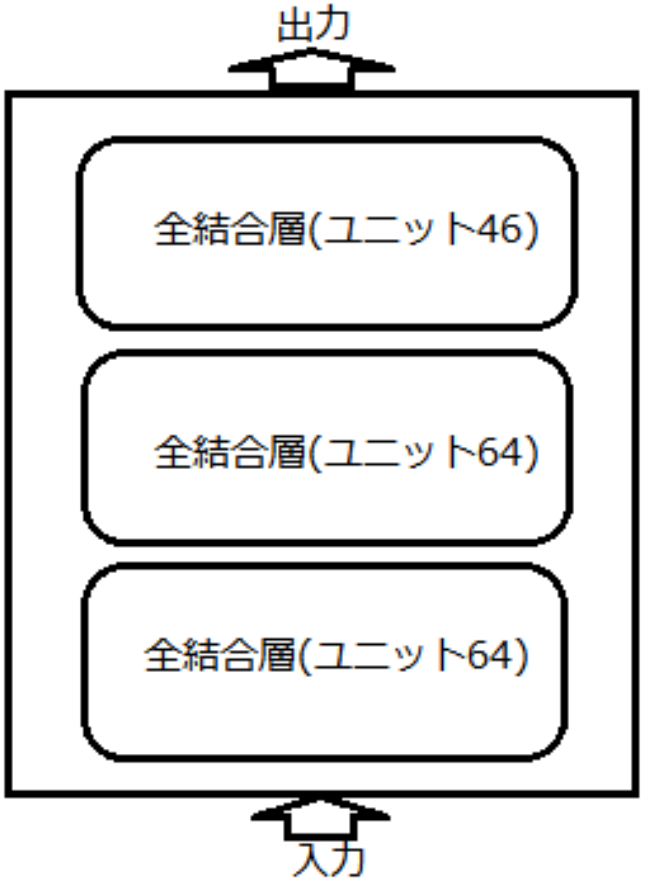


3.5節:46クラスの分類

3.5.3 ニューラルネットワークの構築

ネットワーク最後の層はサイズが46のdense層となる

活性化関数としてソフトマックスを使用する。この場合最適な損失関数はcategorical_crossentropyとなる



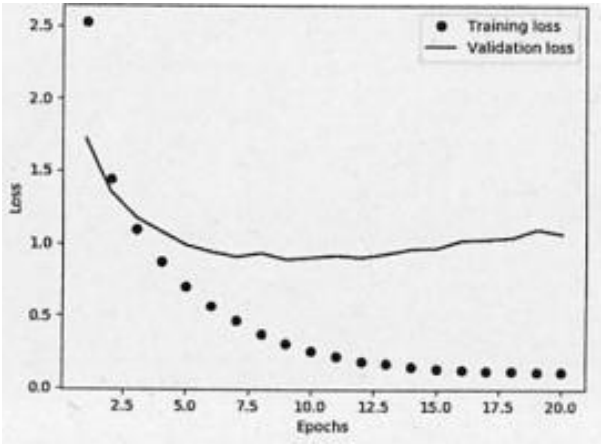
3.5.4 アプローチの検証

１．訓練データのうち1000サンプルを検証データセットとして使用するために分けておく。

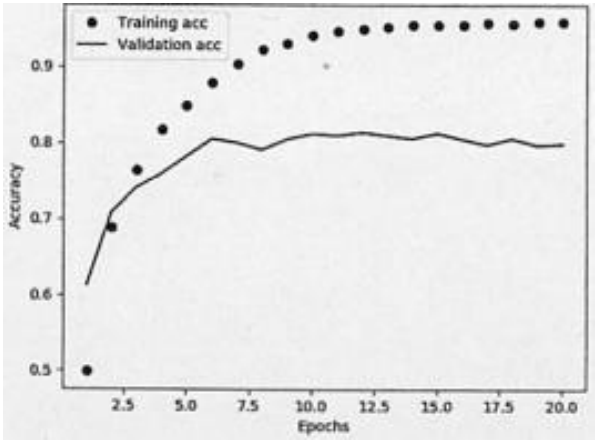
２．512サンプルのミニバッチで20エポックの訓練を行う

３．損失地と正解率をプロットする

3.5.4 アプローチの検証



訓練データでの損失値(ドット)と
検証データでの損失値(折れ線)



訓練データでの正解率と(ドット)
検証データでの正解率(折れ線)

8エポック後に過学習に陥ってる→新しいネットワークで8エポック訓練。

→テストデータで評価した結果78％の正解率

まとめ

・データ点N個のクラスに分類するとき、ネットワーク最後の層はサイズNの層でなければならない。

・多クラス単一ラベル分類問題では、活性化関数としてソフトマックスを使う(出力はNの出力クラスに対する確率分布)

・損失関数は交差エントロピーを使用すべきである。（ネットワークによって出力される確率分布とラベルの真の分布との距離を最小化する）

・ラベルのベクトル化はone-hotエンコーディングとラベルを整数としてエンコードする方法がある。