数理工学実験
テーマ:ニューラルネットワークによる機械学習

2回生　田中風帆 (1029321151)
実施場所:自宅

実施:2022年1月18日
提出:2022年1月27日

目 次

図 目 次

表 目 次

第 I 部
概要

このレポートではニューラルネットワークによる機械学習に関連する課題を解いた結果を報告した。誤差逆伝播法や勾配消失問題に関する課題を解いたのち、損失関数やノード数、エポック数、最適化手法の違いにより結果にどのような差異が生まれるかを観察した。さらにオートエンコーダを用いて MNIST の画像を 2 次元まで圧縮し散布図にした。最後に VAE に関する不等式を証明し、画像の生成を行った。

第 II 部

課題 22

ここでは、スカラー変数 x, w_2, w_3 に対して

$$u^{(l)} = x, h^{(l)} = f(u^{(l)}), u^{(l)} = w^{(l)} h^{(l-1)}, \hat{y} = f^{(3)}(u^{(3)}) \tag{0.1}$$

を定義し、$f(u) = u^2, L(\hat{y}) = \hat{y}^2$ とする。この時、$u^{(l)}, h^{(l)}, \hat{y}^{(l)}, L$ を x, w_2, w_3 の関数として表し、$\frac{\partial L}{\partial w_l}, l = 2, 3$ を求めた。また、この例において

$$\delta^(3) = (2\hat{y})(2u^(3)), \delta^(2) = \delta^(3) w^(3)(2u^(2)) \tag{0.2}$$

を用いて誤差逆伝播法の構造を説明した。

1　回答

$u^{(l)} = x$ より、$h^{(1)} = f(x) = x^2$ となる、よって $u^{(2)} = w_2 h^{(1)} = w_2 x^2$ である。これより、$h^{(2)} = f(u^{(2)}) = f(w_2 x^2) = w_2^2 x^4$ で、$u^{(3)} = w_3 h^{(2)} = w_3 w_2^2 x^4$ である。$\hat{y} = f(u^{(3)}) = w_3^2 w_2^4 x^8$ となるので、$\delta^{(3)} = \frac{\partial \hat{y}^2}{\partial \hat{y}}.2u^{(3)} = 4\hat{y}u^{(3)} = 4w_3^3 w_2^6 x^{12}$、$\delta^{(2)} = \delta^{(3)} w_3 (2u^{(2)}) = 4w_3^3 w_2^6 x^{12} w_3.2.w_2 x^2 = 8w_3^4 w_2^7 x^{14}$ となる。これより、$\mathfrak{L}(\hat{y}) = w_3^4 w_2^8 x^{16}$ となり、$\frac{\partial L}{\partial w_2} = 8w_3^4 w_2^7 x^{16}$、$\frac{\partial L}{\partial w_3} = 4w_3^3 w_2^8 x^{16}$ が導かれる。以上より解は以下のようになる。

$$\begin{aligned} h^{(1)} &= x^2 \\ u^{(2)} &= w_2 x^2 \\ h^{(2)} &= w_3^2 w_2^4 x^8 \\ u^{(3)} &= w_3 w_2^2 x^4 \\ \hat{y} &= w_3^2 w_2^4 x^8 \\ \frac{\partial L}{\partial w_2} &= 8w_3^4 w_2^7 x^{16} \\ \frac{\partial L}{\partial w_3} &= 4w_3^3 w_2^8 x^{16} \end{aligned} \tag{1.1}$$

誤差逆伝播法の構造について解説する。
上の導出から、$\frac{\partial L}{\partial w_3} = \delta^{(3)} h^{(2)} = 4w_3^3 w_2^6 x^{12} w_2^2 x^4 = 4w_3^3 w_2^8 x^{16}, \frac{\partial L}{\partial w_2} = \delta^{(2)} h^{(1)} = 8w_3^4 w_2^7 x^{16}$ であることがわかる。これは確かに、上で求めた $\frac{\partial L}{\partial w_3}$ と $\frac{\partial L}{\partial w_3}$ の値に一致する。$\frac{\partial L}{\partial w_i} (i = 2, 3)$ はその回での誤差関数の値にパラメータ w がどれだけ寄与しているかを表す。よって例えば最急降下法などでは、学習率を α とした時、w_i を $w_i - \alpha \frac{\partial L}{\partial w_i}$ と更新すれば良い。この例だと、$w_2 = w_2 - \alpha(8w_3^4 w_2^7 x^{16})$ と更新し、$w_3 = w_3 - \alpha(4w_3^3 w_2^8 x^{16})$ と更新すれば良い。これが誤差逆伝播法によるパラメータの更新の構造である。

第 III 部

課題 23

ここでは、シグモイド関数は勾配消失問題を起こす一方、ReLU は勾配消失を起こす可能性が低いという現象について説明を与える。

2　回答

活性化関数を f と表記する。誤差逆伝播法において、$\delta_i^{(l)} = \sum_{k=1}^{d_{l+1}} \delta_k^{(l+1)} w_{k_i}^{(l+1)} f'(u_i^{(l)})$ と表されることと、$\frac{\partial L}{\partial w_{ij}^{(l)}} = \delta_i^{(l)} h_j^{(l-1)}$ であることから、層が深くなれば深くなるほど $\frac{\partial L}{\partial w_{ij}^{(l)}}$ を求める過程で活性化関数の微分が多く乗算されることになる。シグモイド関数は $f(x) = \frac{1}{1+e^{-x}}$ で表され、その微分は $f'(x) = 0 < \frac{e^{-x}}{(1+e^{-x})^2} < 1$ である。よって $\lim_{n\to\infty} f'(x)^n = 0$ となり、勾配消失が起こる。一方 ReLU は $1 < x$ において微分の値は 1 となるため、乗算しても勾配消失は起こりづらい。以上が与えられた現象に対する説明である。

第 IV 部

課題 24

この課題では MNIST を用いて一層の NN を学習させた。その際、損失関数を MSE にしてクロスエントロピーの結果と比較した。ただし、最適化関数は Adam、バッチのサイズは 128、エポック数は 10 とした。以下の課題において、特に断りがない限り、「課題 24 の NN」はこの設定の NN を指す (ただし損失関数はクロスエントロピー)。

3　回答

最終エポックにおいて、テストデータに対する正解率は以下のようになった。

表 1: 上から MSE を用いた場合のテストデータに対する正解率、クロスエントロピーを用いた場合の正解率

MSE	0.9246
クロスエントロピー	0.9305

4　考察

両者とも精度に大きな差は見られなかったが、クロスエントロピー誤差の方に若干の優位性が見られた。MSE の微分は引数 x の一次関数で表されるのに対し、クロスエントロピーの微分は $-\frac{1}{x}$ の形で表されることから、x が 0 に近い時、クロスエントロピーを用いた場合の損失関数の変動が大きくなり、効率よく学習が進むのが原因ではないかと推察する。

第 V 部
課題 25

ここでは

$$f(x_1, x_2) = 10 - 10exp(-0.2(x_1^2 + x_2^2)) - exp((cos(x_1) + cos(x_2))/2) \quad (4.1)$$

に対して回帰問題を考えた。n 個のデータ $(x_{i1}, x_{i2}), i = 1, ..., n$ からこの関数上の値 $y_i = f(x_{i1}, x_{i2}) + \epsilon$ を生成し、データセット $\{x_{i1}, x_{i2}, y_i\}$ を基に NN による関数 $f_{NN}(x_1, x_2)$ を学習し、$(x_1, x_2, f_{NN}(x_1, x_2))$ を 3 次元プロットとして図示した。ただし、今回 ϵ は [-0.01,0.01] の一様分布から生成し、$n = 1000$ とした。また、x_{i1}, x_{i2} はどちらも [-10.0,10.0] の一様分布から生成した。batch の数は 100 で、1000 エポックの学習を行った。隠れ層の数は 2 層で、それぞれ入力数は 32,16 とした。最適化アルゴリズムとして Adam を用い、誤差関数は MSE、活性化関数は両層とも ReLU とした。

5　回答

学習の結果得たモデルのプロットは以下の赤線で表される。

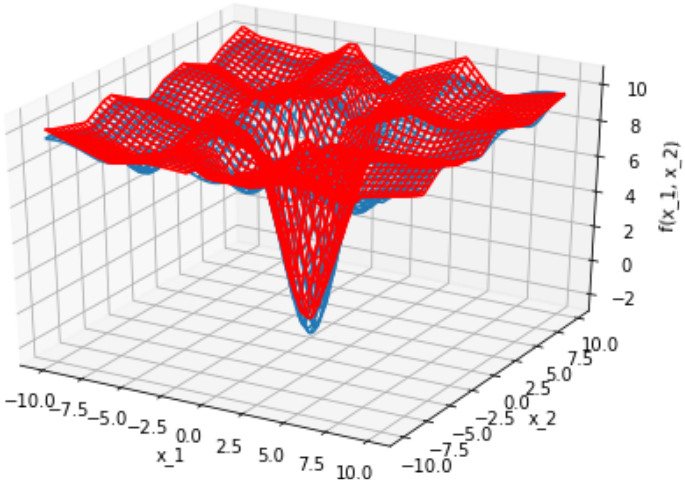

図 1: 学習の結果得たモデル. 赤線がモデルで、青線が本来の関数値.

6　考察

隠れ層の数を 2 枚にしたことによって、複雑な非線形関数に対する回帰が行えるようになったと推察する。また、学習ははじめの 10 回程度で収束していることから、1000 エポックも回す必要はなかったと結論づけられる。

第 VI 部

課題 27

課題 24 で用いた NN(ただし損失関数はクロスエントロピー誤差) に隠れ層を 1 枚追加した。その際、隠れ層のノードの数を 10,100,1000,10000,100000 とした時の結果を比較した。

7　回答

結果は以下のようになった。上から順に、隠れ層のノード数が 10,100,1000,10000,100000 の時のテストデータに対する最終的な正答率を示している。

表 2: 上から順に、隠れ層のノード数が 10,100,1000,10000,100000 の時のテストデータに対する最終的な正答率を表す.

隠れ層のノード数	正答率
10	0.9321
100	0.9769
1000	0.9819
10000	0.9820
100000	0.9821

8　考察

隠れ層のノード数が増えれば増えるほど、テストデータに対する正答率が上がることがわかった。これはノード数の多い NN の方がより多くのパラメータを調節することができることに起因すると考察する。ただしノード数が 1000 を超えてからはそこまで大きな違いは見られなかった。これは認識の難しいデータが一定数含まれていることが原因であると思われる。

第 VII 部
課題 28

ここでは多層 NN において、エポック 100 としてエポックを増やした時の MNIST 画像の学習の様子を報告した。ただし隠れ層の数は 3 で活性化関数は全て ReLU(最終層のみ softmax)、入力数はそれぞれ 1024,256,256、最適化アルゴリズムはクロスエントロピー、バッチサイズは 128 とした。以下の課題において、特に断りがない限り「課題 28 の NN」はこの設定の NN を指す (ただしエポックは 10)。

9　回答

結果の学習曲線は以下のようになった。

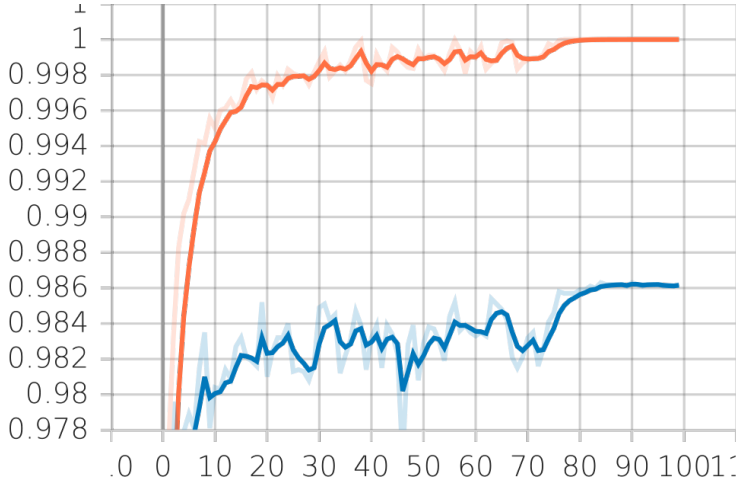

図 2: 縦軸がテストデータに対する正答率、横軸がエポック数. 青線がテストデータに対する結果で、橙線が訓練データに対する結果.

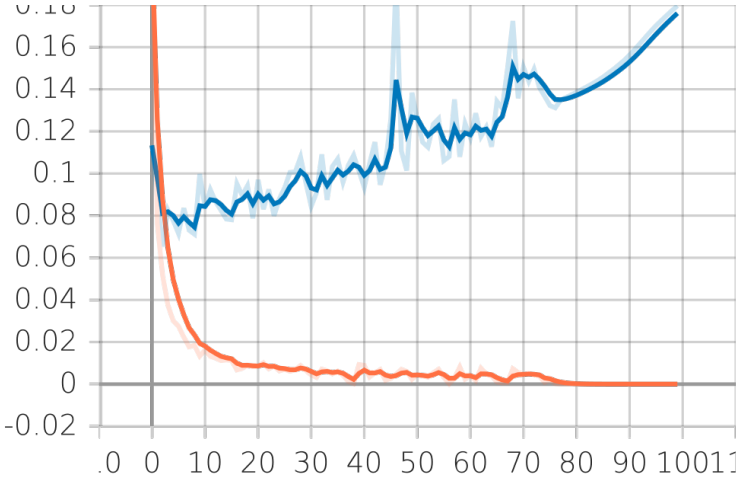

図 3: 縦軸が誤差関数の値、横軸がエポック数. 青線がテストデータに対する結果で、橙線が訓練データに対する結果.

10 考察

初めのうちはエポックを増やすごとにテストデータに対する正答率が上がり誤差関数の値が小さくなっていくのを見て取れたが、さらに増やしていくと正答率は横這いとなり、また誤差関数の値はむしろ大きくなっていくこと

がわかった。一方、訓練データに対してはエポックを増やせば増やすほど正答率は上がり、誤差関数の値は小さくなっていく傾向にあった。エポックを増やしすぎると過学習が起こるのが原因であると推察する。

第 VIII 部
課題 29

ここでは、課題 28 の NN における最適化アルゴリズムとして SGD, AdaDelta, AdaGrad, Adam, FTRL, RMSProp を用いた場合の結果を比較した。

11　回答

結果は以下のようになった。上から順に、SGD, AdaDelta, AdaGrad, Adam, FTRL, RMSProp を用いた場合のテストデータに対する最終的な正答率を示している。

表 3: 上から順に、SGD, AdaDelta, AdaGrad, Adam, FTRL, RMSProp を用いた場合のテストデータに対する最終的な正答率を表す.

アルゴリズム	正答率
SGD	0.9478
AdaDelta	0.9506
AdaGrad	0.9561
Adam	0.9823
FTRL	0.1135
RMSProp	0.1135

12　考察

SGD, AdaDelta, AdaGrad, Adam は全て高い精度で画像を認識することができたが、FTRL, RMSProp では全く画像を正確に認識することができなかった。これは最後の二つのアルゴリズムが多クラス分類または画像認識のタスクに不向きであることを示しており、適切な最適化アルゴリズムを選択することが重要であるといえる。

第 IX 部
課題 30

ここでは、MNIST と CIFAR-10 それぞれで畳み込み層を入れたかどうかでどれほど精度に違いがあるかを確認した。

13　回答

結果は以下のようになった。上の表は MNIST で、畳み込み層を入れた場合と入れなかった場合、下の表は CIFAR-10 で、畳み込み層を入れた場合と入れなかった場合のテストデータに対する正答率を表している。

表 4: MNIST に関する表. 上から順に、畳み込み層を入れた場合と入れなかった場合のテストデータに対する正答率を示す.

畳み込み層	正答率
あり	0.9941
なし	0.9855

表 5: CIFAR-10 に関する表. 上から順に、畳み込み層を入れた場合と入れなかった場合のテストデータに対する正答率を示す.

畳み込み層	正答率
あり	0.7730
なし	0.3575

14　考察

MNIST では畳み込み層の有無にかかわらず高精度のモデルが構築できたが、畳み込み層のあるモデルの方に若干の優位性が見られた。一方で CIFAR-10 の場合は畳み込み層のないモデルは正答率が 3 割台であり、畳み込み層のあるモデルであっても正答率が 8 割を切るなどあまり高精度とは言えなかった。CIFAR-10 においては畳み込み層を入れた場合とそうでない場合の精度の差が大きかったことから、畳み込み層の効果は白黒画像よりもカラー画像に対して大きく発揮されることがわかった。カラー画像には R,G,B の三層があるため学習が難しく、またモデルの改善の余地が大きいことに起因すると推察する。

第 X 部

課題 31

ここではオートエンコーダを用いて MNIST の 0,1,2 の画像 100 枚ずつを 2 次元に圧縮し、その散布図を表示した。

15　回答

結果は以下のようになった。

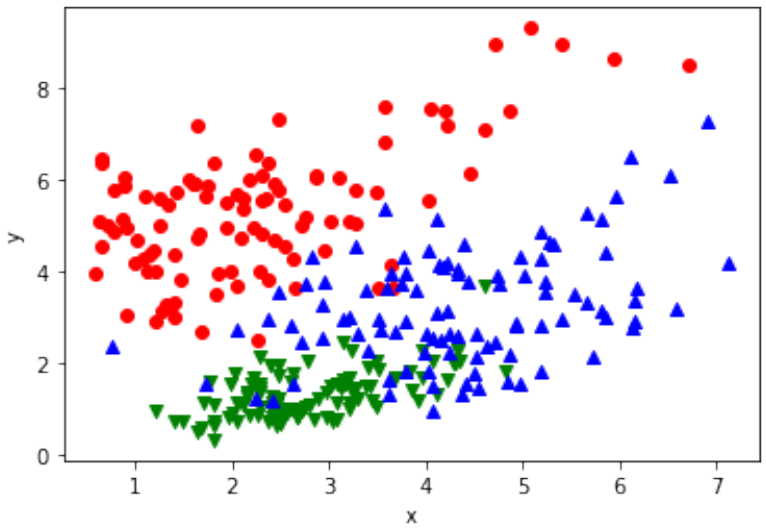

図 4: 0,1,2 の散布図. 赤が 0, 緑が 1, 青が 2 である.

16　考察

得られたプロットより、確かに数字ごとに特徴が圧縮され 2 次元で適切に表現されていることがわかった。数字がそれぞれ固まった位置にプロットされていることも見て取れる。

第 XI 部
課題 32

ここでは

$$\begin{aligned}\log p_\theta(\boldsymbol{x}) \geq -D_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})) \\ + \int \log q_\phi(\boldsymbol{z}|\boldsymbol{x}) \log p_\theta(\boldsymbol{x}|\boldsymbol{z}) d\boldsymbol{z}\end{aligned} \tag{16.1}$$

であることを示した。

17　回答

上の式の成立を示す。この際、イェンセンの不等式を用いた。イェンセンの不等式は以下で表される。
$p(x)$ を $\int p(x)dx = 1$ を満たす実数上の可積分関数とする。また、$y(x)$ を実数上の可積分関数とする。このとき次が成り立つ。

$$\int_{-\infty}^{\infty} f(y(x))p(x)dx \geq f(\int_{-\infty}^{\infty} y(x)p(x)dx) \tag{17.1}$$

この定理は、f の $f(\int_{-\infty}^{\infty} y(x)p(x)dx)$ における接線を g とおいて、常に $f(x) < g(x)$ が成り立つことから導かれる [1]。
まず、式

$$\begin{aligned}-D_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})) = \\ -\int q_\phi(\boldsymbol{z}|\boldsymbol{x}) \log \frac{q_\phi(\boldsymbol{z}|\boldsymbol{x})}{p(\boldsymbol{z})} d\boldsymbol{z} = \\ \int q_\phi(\boldsymbol{z}|\boldsymbol{x})(\log p(\boldsymbol{z}) - \log q_\phi(\boldsymbol{z}|\boldsymbol{x}))d\boldsymbol{z}\end{aligned} \tag{17.2}$$

が成立する。よって、

$$\begin{aligned}\log p_\theta(x) = \log \int p(\boldsymbol{z}) p_\theta(\boldsymbol{x}|\boldsymbol{z}) d\boldsymbol{z} \\ = \log \int q_\phi(\boldsymbol{z}|\boldsymbol{x}) \frac{p(\boldsymbol{z}) p_\theta(\boldsymbol{x}|\boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})} d\boldsymbol{z} \geq \\ \int q_\phi(\boldsymbol{z}|\boldsymbol{x}) \log \frac{p(\boldsymbol{z}) p_\theta(\boldsymbol{x}|\boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})} d\boldsymbol{z} = \\ \int q_\phi(\boldsymbol{z}|\boldsymbol{x})(\log p(\boldsymbol{z}) + \log p_\theta(\boldsymbol{x}|\boldsymbol{z}) - \log q_\phi(\boldsymbol{z}|\boldsymbol{x}))d\boldsymbol{z} = \\ \int q_\phi(\boldsymbol{z}|\boldsymbol{x})(\log p(\boldsymbol{z}) - \log q_\phi(\boldsymbol{z}|\boldsymbol{x}))d\boldsymbol{z} + \int q_\phi(\boldsymbol{z}|\boldsymbol{x}) \log p_\theta(\boldsymbol{x}|\boldsymbol{z}) d\boldsymbol{z} = \\ -D_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})) + \int q_\phi(\boldsymbol{z}|\boldsymbol{x}) \log p_\theta(\boldsymbol{x}|\boldsymbol{z}) d\boldsymbol{z}\end{aligned} \tag{17.3}$$

となり、題意の成立が示された。この際、途中の不等号の部分でイェンセンの不等式を用いている。

18　課題 33

ここでは、CIFAR-10 に対して VAE を用いて画像を生成した。エポック数は 30 とした。

19　回答

生成した画像は以下である。

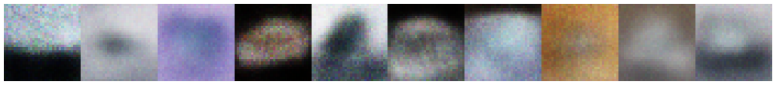

図 5: VAE で生成した画像

20　考察

畳み込み層を使用し、エポックも 30 にしたものの、あまりうまく画像を生成することができなかった。これと同じモデルでエポック 10 の時、MNIST ではうまく画像生成できていたことから、カラー画像になると画像の生成が難しくなるということが予測された。これは白黒画像に対し、カラー画像には R,G,B の 3 要素が加わることに起因すると推察する。

第 XII 部
まとめ

このレポートではニューラルネットワークによる機械学習に関連する課題を解いた結果を報告した。誤差逆伝播法や勾配消失問題に関する課題を解いたのち、損失関数やノード数、エポック数、最適化手法により結果にどのような差異が生まれるかを観察した。結果、エポック数を増やしすぎると過学習が起こることなどを確認することができた。さらにオートエンコーダを用いて MNIST の画像を 2 次元まで圧縮し散布図にした。最後に VAE に関する不等式を証明し、画像の生成を行った。

参考文献

[1] 「凸関数と Jensen の不等式」http://www.mathlion.jp/article/ar128.html