ニューラルネットワークにおける重みの局所性と共有

氏名　佐藤 怜
所属　筑波大学 情報科学類
日付　2017.3.3
改訂　2017.3.8

1　概要

この文書ではニューラルネットワークの畳み込み層の成功の理由について考察し, 得られた知見を述べます. この文書は先行研究 [4] から多くの示唆を受けています. また, [26] の読了を前提としています.

2　畳み込みの分解

畳み込み層は全結合層によって同値な処理を表現できます. しかし実際には画像データに対して畳み込み層の方がより良い精度を達成できることから, 全結合層に対する誤差逆伝播では畳み込み層が学習するパラメータには収束し難いことがわかります. よって, 畳み込み層は全結合層により良いパラメータを学習させるための一手法であることがわかります. つまり, 畳み込み層はドロップアウトと同列の最適化の手法であると見做せます.

畳み込み層の処理は全結合層と比べて大きく飛躍しましたが, 畳み込み層は全結合層に重みの局所性と共有の 2 つの特性を加えた層です.

2.1　低次元性

畳み込み層の層間結合は全結合ではなく局所結合です. すなわち疎な法線ベクトルを持つ分離超平面を学習します. これはデータ空間に対して, 特定の次元軸群に平行な分離超平面を学習することを意味します. 個々の分離超平面が識別的である為には, 分布の分離境界が次元軸群に平行である必要があります. つまり, 新しいデータは, 幾つかの分離超平面の上側に位置していますが, 個々の分離超平面の上側であるためには幾つかの次元の活性しか関わりがないということです.

2.1.1　Ensemble

強制的に特定の次元軸に平行になった分離超平面群は, すなわちある制約のもとで協働を強いられていることになります. 学習を行った末に役に立たない (識別的でない) 分離超平面が発生しても, 他の超平面が似たような法線ベクトルを持つことがありません. 従って, 有用な分離超平面の伝播重みだけを大きくすれば良いと考えられます.

2.2 対称性

画像分布を識別する分離超平面を学習すると重みの共有は, データ分布を一定の規則で回転させながら, 低次元断面から同じ視点で見た分離超平面でデータを識別するイメージです. 言い換えれば, 違う低次元断面から見るデータ分布が類似する形になるという意味です. 先にも述べたようにこれはある回転に対して不変な分布の形ですから, 対称性があると考えられます.

2.2.1 Augmentation

Augmentation とは水増しの意です. データが対称性を持つ場合, データを回転させても同じ分離超平面である程度上手く識別できます. このようなデータ分布であるという前提がある場合は, 回転させた水増しデータで頑強なパラメータを学習していると捉えられます.

3 Experiments

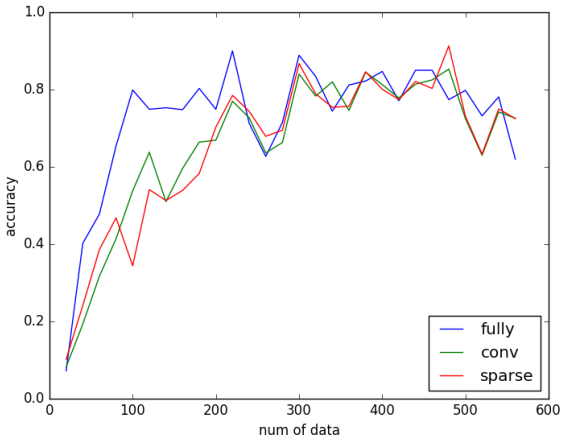

図 1　MNIST に対する学習曲線の 50 回の平均

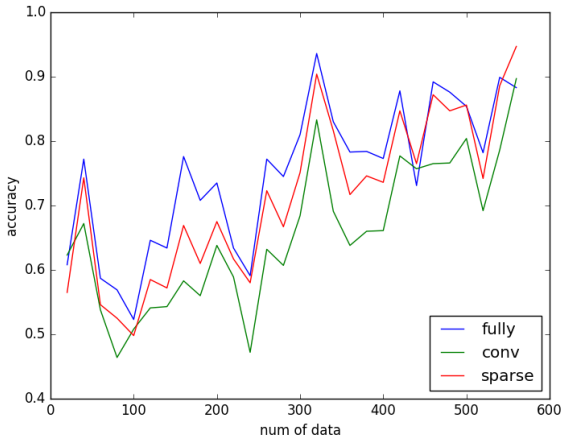

図 2　BreastCancerWisconsin に対する学習曲線の 50 回の平均

図 1は MNIST[24], 図 2は BreastCancerWisconsin[25] に対して 3 種類のニューラルネットワークをそれぞれ学習した結果です.”fully”は全結合層を, ”conv”は畳み込み層を, ”sparse”は畳み込み層と同様に疎な層間結合を持ちつつ, 重みを共有しない層を持つニューラルネットワークを表します. それぞれのモデルは中間ノードを 1 層持ち, 入力層から中間層までに上述の層を用い, 中間層から出力層への層間結合には全結合層を用います. MNIST[24] に対する実験では, conv と sparse では層間結合は 5*5 の 25 次元フィルタを用い, 画像の周囲に 2 ピクセルずつゼロパディングをします. 中間層のノード数は入力層と同じ 784 です.

BreastCancerWisconsin[25] に対する実験では, 30 次元のデータを 5*6 の画像に整形してから順伝播を行います. conv と sparse では層間結合は 3*3 の 9 次元フィルタを用い, 画像の周囲に 1 ピクセルずつゼロパディングをします.

MNIST に対しては学習可能な分離超平面群の自由度の著しい減少に対して, 精度は顕著に減少していると

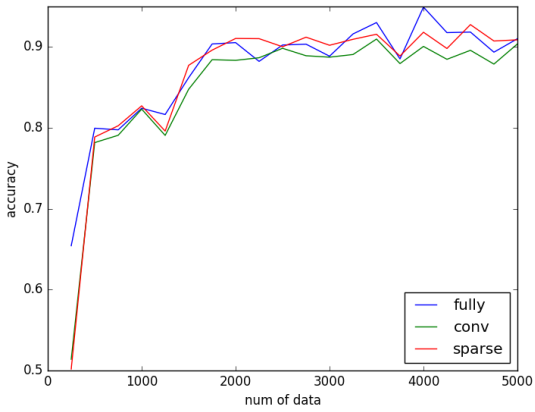

図 3　MNIST に対する学習曲線の 50 回の平均

は捉えがたい結果となりました．これはデータ分布が低次元性と対称性を備えているからだと考えられます．BreastCancerWisconcin に対しては conv と sparse の精度の減少が MNIST より大きいですから，低次元性と対称性は MNIST ほどではないことがわかります．

付録

次元の呪い

データ空間の次元の増加によって引き起こされる弊害を一般に次元の呪いと呼びます.

近傍性

一般に次元の呪いの弊害として近傍性の消失が挙げられますが, これは正しくありません. まず, 球面集中現象は一部の分布でしか発生しません. また, 発生したとしても近傍性が消失する訳ではなく, 距離の分布の分散が小さくなるだけです.

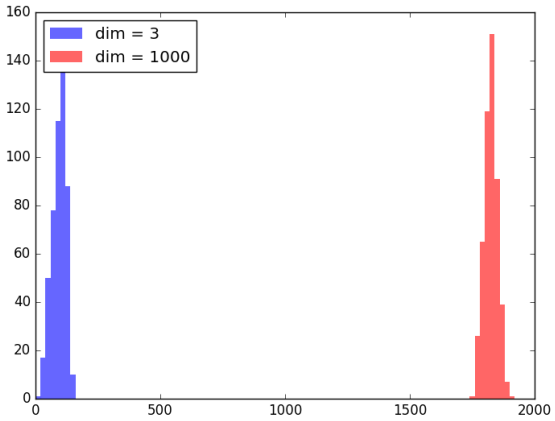

図 4　R(超) 立方体の原点からの距離の分布

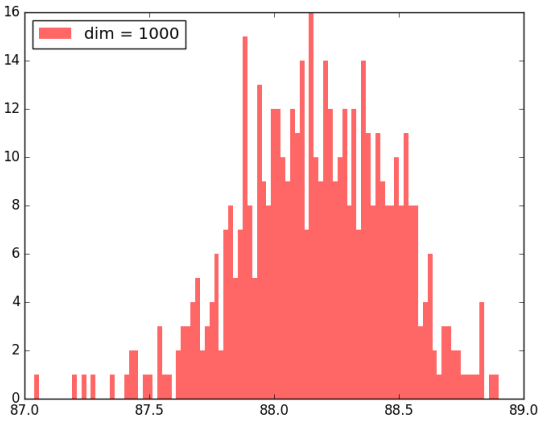

図 5　Z 超球の原点からの距離の分布

データ数

図 6は, 本質的には x の値だけが 2 クラスを分類する根拠であるようなデータに, ノイズ次元である y と z を追加したような 3 次元ベクトルの集合です.

図 6のデータ群に対して最近傍法とサポートベクトルマシンによる分類を行った結果を図 7と図 8に示します. 300 個の学習データでは, どちらも次元の増加と共に新たなデータに対する予測精度が下降していきます.

原因の考察

次元の増加と共に精度が下落するのは, 必要なデータ数が増加するからだと言われています [16] が, その考察を以下に示します.

まず図 6の分布について 2 クラスの分離境界に, 図 9のような, 厚さ ϵ の擬似的な壁を想像します. 壁にデータが属する確率は次元に依らず変わりません. しかし厚さ ϵ の壁の体積は次元の増加に伴って指数関数的に増加します. 従ってこの壁の密度は指数関数的に減少します.

すなわち, 2 クラスが接している体積に対してデータが少なすぎる状態が発生します. 壁が疎な状態では, 新

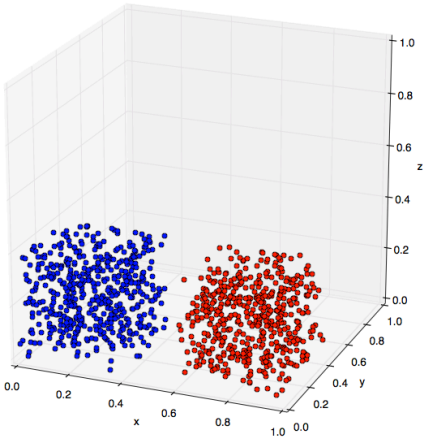

図 6　人工データ

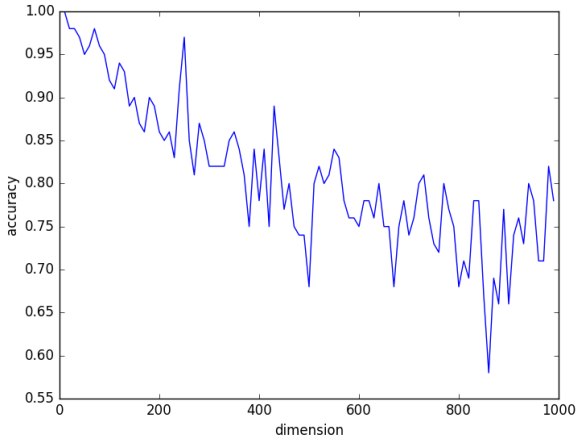

図 7　最近傍法

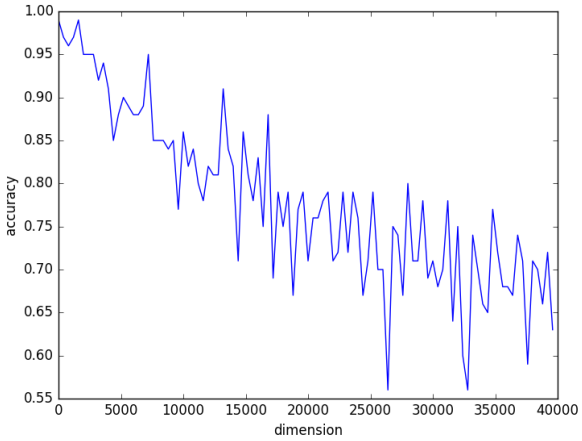

図 8　サポートベクトルマシン

たなデータが壁の近傍にプロットされた場合に誤分類し, 最近傍法の精度が低下すると考えます.

　また, サポートベクトルマシンにおいては分離超平面を支えるサポートベクトルを壁に属さない点から選んでしまうと, 分離超平面が理想的な分離境界から傾いてしまうことが想像できます. 必要なサポートベクトルの数は次元にほぼ比例すると考えられますから, 次元に比例して必要なデータ数が増加していきます.

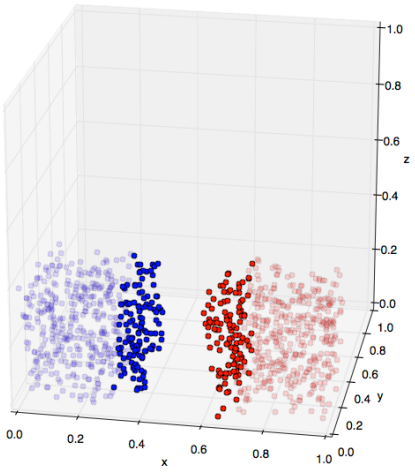

図 9　人工データ

3.1　ラグランジュの未定乗数法の幾何学的イメージ

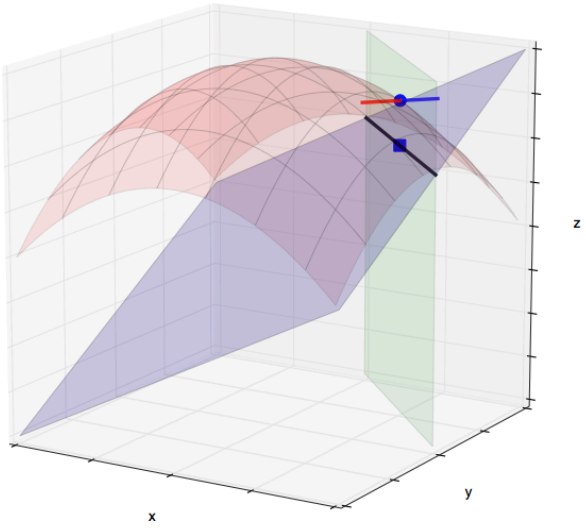

図 10　赤い曲面は目的関数 $f(x)$. 青い平面は制約関数 $g(x)$. 黒い直線は $g(x) = 0$ となる実行可能集合. 青い丸の点はラグランジュ関数 $z = L(x, y, \lambda)$ が最大となる (x, y, z). 青い四角の点は $L(x, y, \lambda)$ を最大化する最適解 (x, y). 赤い直線は最適解における $f(x)$ の勾配ベクトルを定数倍し平行移動したもの. 青い直線は最適解における $g(x)$ の勾配ベクトルを定数倍し平行移動したもの.

3.2 逐次最小問題最適化法の実装

a_n はラグランジュ乗数, x_n はデータベクトルを表す. 教師ラベルは $y_n \in \{-1, 1\}$ である. C はスラック変数のペナルティに関するハイパーパラメータである. 任意の a_n の初期値を 0 とする.K_{ij} は x_i と x_j の内積を表す.

分離超平面の法線ベクトル w は以下の式で与えられる.

$$w = \sum_{i=1}^{n} a_i y_i x_i \tag{1}$$

ここで, $a_i = 0$ であるデータについては w の値に関与しない. よって w は $0 < a_i$ を満たすデータ (サポートベクトル) についてのみ依存する.

分離超平面のバイアス項 b は以下の式で与えられる.

$$A = \{i \mid 0 < a_i < C\} \tag{2}$$

$$b = \frac{1}{|A|} \sum_{i \in A} (y_i - w x_i) \tag{3}$$

ここで, b は $0 < a_i < C$ を満たすデータ (マージン境界上) についてのみ依存する.

a_2^{new} は以下の式で与えられる.

$$a_2^{new} = a_2^{old} + \frac{y_2((wx_1 + b) - y_1) - ((wx_2 + b) - y_2)}{K_{11} + K_{22} - 2K_{12}} \tag{4}$$

KKT 条件より, 求められた a_2^{new} は以下の制約を受ける.

$$\begin{cases} max(0, a_1^{old} + a_2^{old} - C) \leq a_2^{new} \leq min(C, a_1^{old} + a_2^{old}) & (y_1 = y_2) \\ max(0, a_2^{old} - a_1^{old}) \leq a_2^{new} \leq min(C, C - a_1^{old} + a_2^{old}) & (y_1 \neq y_2) \end{cases}$$

a_1^{new} は以下の式で与えられる.

$$a_1^{new} = a_1^{old} + y_1 y_2 (a_2^{old} - a_2^{new}) \tag{5}$$

全ての a_n が以下の KKT 条件を満たすまで反復する.$\epsilon = 0$ で厳密だが収束に時間が掛かるため, [20] では 0.001 に設定している.

$$\begin{cases} y_i(wx_i + b) \geq 1 - \epsilon & (a_i = 0) \\ y_i(wx_i + b) = 1 \pm \epsilon & (0 < a_i < C) \\ y_i(wx_i + b) \leq 1 + \epsilon & (a_i = C) \end{cases}$$

反復の際, 必ずしもマージンの増加と KKT 条件に違反する a_n の減少は連動しない.

1 ステップで更新する 2 変数 a_1 と a_2 の選び方については経験的な実装が多く, 収束までの時間を大きく左右する.

参考文献

[1] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks,
http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks

[2] Min Lin, Qiang Chen, Shuicheng Yan, Network In Network,
https://arxiv.org/pdf/1312.4400v3.pdf
[3] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, Lior Wolf, DeepFace: Closing the Gap to Human-Level Performance in Face Verification,
https://www.cs.toronto.edu/~ranzato/publications/taigman_cvpr14.pdf
[4] Andrew M. Saxe, Pang Wei Koh, Zhenghao Chen, Maneesh Bhand, Bipin Suresh, Andrew Y. Ng, On Random Weights and Unsupervised Feature Learning,
http://www.robotics.stanford.edu/~ang/papers/nipsdlufl10-RandomWeights.pdf
[5] Kevin Jarrett, Koray Kavukcuoglu, Marc'Aurelio Ranzato, Yann LeCun, What is the Best Multi-Stage Architecture for Object Recognition?
http://yann.lecun.com/exdb/publis/pdf/jarrett-iccv-09.pdf
[6] Henry W. Lin, Max Tegmark, Why does deep and cheap learning work so well?,
https://arxiv.org/pdf/1608.08225v1.pdf
[7] Christopher Olah, Neural Networks, Manifolds, and Topology,
http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/
[8] 浅川伸一, ディープラーニングに用いる畳み込み演算による概念操作の表現,
http://www.jcss.gr.jp/meetings/jcss2015/proceedings/pdf/JCSS2015_P3-21.pdf
[9] 中山英樹, 画像認識分野における deep learning の発展と最新動向,
http://www.nlab.ci.i.u-tokyo.ac.jp/pdf/asj20141215.pdf
[10] 原田達也, 知能情報論 深層学習 畳み込みニューラルネットワーク,
http://www.mi.t.u-tokyo.ac.jp/harada/lectures/IIT/internal/08_deepCNN_20160615.pdf
[11] 岡谷貴之, 深層学習
[12] 中山英樹, 深層畳み込みニューラルネットワークによる画像特徴抽出と転移学習,
http://www.nlab.ci.i.u-tokyo.ac.jp/pdf/CNN_survey.pdf
[13] 奥牧人, 球面集中現象 (concentration on the sphere) と次元の呪いの関係,
http://www.sat.t.u-tokyo.ac.jp/~oku/20161025/memo.html
[14] 坂野鋭, 山田敬嗣, 怪奇!!次元の呪い-識別問題, パターン認識, データマイニングの初心者のために-(前編),
https://ipsj.ixsq.nii.ac.jp/ej/?action=pages_view_main&active_action=repository_view_main_item_de
[15] 坂野鋭, 山田敬嗣, 怪奇!!次元の呪い-識別問題, パターン認識, データマイニングの初心者のために-(後編),
https://ipsj.ixsq.nii.ac.jp/ej/?action=pages_view_main&active_action=repository_view_main_item_de
[16] Ela Pekalska, Bob Duin, The peaking paradox,
http://37steps.com/2279/the-peaking-paradox/
[17] 佐野宏喜, はじめよう多変量解析〜主成分分析編〜,
http://www.slideshare.net/sanoche16/tokyor31-22291701
[18] Christopher M. Bishop, パターン認識と機械学習 上
[19] Christopher M. Bishop, パターン認識と機械学習 下
[20] John C. Platt, Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines,
https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/tr-98-14.pdf
[21] 中川裕志, 5. サポートベクターマシン,

http://www.r.dl.itc.u-tokyo.ac.jp/%7enakagawa/SML1/kernel1.pdf
[22] 福水健次, 4. サポートベクターマシン,
http://www.ism.ac.jp/%7efukumizu/ISM_lecture_2010/Kernel_4_SVM.pdf
[23] 数原良彦, SMO 徹底入門 - SVM をちゃんと実装する,
http://www.slideshare.net/sleepy_yoshi/smo-svm
[24] Yann LeCun, Corinna Cortes, Chris Burges, MNIST handwritten digit database,
http://yann.lecun.com/exdb/mnist/
[25] Dr. William H. Wolberg, W. Nick Street, Olvi L. Mangasarian, Breast Cancer Wisconsin (Diagnostic) Data Set,
https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)
[26] 佐藤怜, 全結合ニューラルネットワークの解釈