情報システム工学科

2009 年 4 月 12 日

塩貝亮宇

Reducing the Dimensionality of Data with Neural Networks

T1076006 塩貝亮宇

1 paragraph 1

(dimensionality reducing facilitates, 次元減少を手助け)、可視化、コミュニケーション、そして、高次元データの格納。

シンプルで広く使われている手法は主成分分析 (PCA) で、これはデータセットの最良の分散の方向を発見し、そして、それらの方向のそれぞれに沿った (次元, 軸,coordinate) によって表現される。

高次元データを低次元データに変換するための多層'encoder' 網と同様の'decoder' 網をコードから再生するための非線形 PCA の一般化について説明していく。

2 層のランダム重みから開始する、

それらはオリジナルデータとそれの復元の間の相違を最小化することによって

同時に訓練することができる。

その必要とされる勾配は最初から復号網と最初から符号化網から派生する誤差逆伝播はチェインルールを用いることによって容易に獲得される。

全てのシステムは'autoencoder' と呼ばれ、そして、図 (1) に描かれる。

2 paragraph 2

多層の隠れ層を持つ非線形'autoencoder' の重みを最適することは難しい。

広範の初期の重みについて、'autoencoder' は一般的に (貧しい, 貧弱な,poor) 局所最適解を発見し; 小さな初期化重みを伴ない、最初の層はの勾配はとても小さく、多くの隠れ層を伴なって'autoencoder' を訓練することは実行不可能だ。

もし、初期の重みが良い解に近ければ (勾配降下,gradient descent) は良く働き、

しかし、初期の重みは同時にひとつの特徴のある層を学習するとても異なるタイプアルゴリズムを発見する。

我々はバイナリデータについて' 事前訓練' 手続きを紹介し、実数データに向けて一般化し、データセットの変化についてよく働くことを示す。

3 paragraph 3

バイナリベクトルの集合体 (例えば、画像) は'restricted Boltzmann machine' と呼ばれる 2 層のネットワークを使うことでモデル化され、バイナリピクセルは統計的に結合されており、バイナリ特徴は対称結合の重みが使われる。

画素は RBM の可視素子と一致する。なぜならば、それらの状態は観測され; その特徴検出器は'hidden' 素子と一致する。

可視素子と隠れ素子の間の $(\boldsymbol{v}, \boldsymbol{h})$ 関節配置が持つエネルギは

$$E(\boldsymbol{v}, \boldsymbol{h}) = -\sum_{i \in pixels} b_i v_i - \sum_{j \in features} b_j h_j - \sum_{i,j} v_i h_j w_{ij} \tag{1}$$

によって与えられ、ここで、v_i と h_j は画素 i と特徴 j のバイナリ状態、そして b_j はそれらのバイアス、そして w_{ij} はそれらの間の重みである。

ネットワークには
このエネルギー関数を通して全ての取りうる画像について代入され、(8) で説明される。
訓練画像の確率は
に順応する重み
によって引き起こされる。
与えられた訓練画像、各特徴検出器 j のバイナリ状態 h_j には
確率 $\sigma(b_j + \sum_i v_i w_{ij})$ に伴ない 1 が代入され、
ここで $\sigma(x)$ はロジスティック関数 $1/[1 + \exp(-x)]$、b_j は j のバイアス、v_i は画素 i の状態、そして w_{ij} は i と j の間の重み。
かつて、バイナリ状態は隠れ素子について選択された、'confabulation(談笑)' は各 v_i が確率 $\sigma(b_j + \sum_j h_j w_{ij})$ にともない 1 がセットされ、ここで b_j は i のバイアスだ。
隠れ素子の状態は、そのとき更新される。
もう一度繰り返して隠れ素子が'confabulation' の特徴を表現することができる。
その重みは

$$\Delta w_{ij} \quad = \quad \epsilon(\langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{recon}) \tag{2}$$

ここで ϵ は学習率、$\langle v_i h_j \rangle_{data}$ は
データによって特徴検出器が走らされているとき画素 i と特徴検出器 j が同時に起きた回数のごく一部、そして $\langle v_i h_j \rangle_{recon}$ は極一部の'confabulation' について一致している。
同様の学習則の単純な版は、そのバイアスについて使われる。
その学習は訓練データの対数確率の勾配に続いて正確ではないけれども良く働く。(6)
バイナリ特徴の単層は画像のセット内の構築物をモデル化するための最良な方法ではない。
特徴検出器のある層を学習後、
私たちはそれらの活動を取り扱うことができる
-データによってそれらを走らせたとき- 特徴の 2 番目の層の学習についてのデータとして。
特徴検出器の最初の層は、そのとき次の RBM 学習について可視素子になる。
この (層による層の,layer-by-layer) 学習は求められるだけ繰り返すことが出来るだろう。
モデルに代入される訓練データの対数確率の下限の改善は常に余分な層に加えられることを示せるだろう、
提供される特徴検出器毎層の数は減少できず、そして、それらの重みは正しく初期化される (9)。
この境界は高層がより少ない特徴検出器を持つときには適応されないが、
層による層の学習アルゴリズムは、'deep autoencoder' の重みの事前訓練法にも関わない。
特徴を捉える各層は強い、
その層より下になる素子の活動間の高階の相互関係。
データセットの広範の様々について、この低次元を前進的に公開 (progressive reveal) する方法、非線形構造。

4　4

特徴検出器の多層を事前訓練した後、そのモデルは同じ重みを使う初期化するための encoder と decoder のネットワークを生成するために'unfolded'(Fig.1)。
global fine-tuning ステージは、決定的に統計活動が置き変わり、真値の確率
そして
最適 reconstruction についての重みの fine-tune するための全ての'autoencoder' を通る誤差逆伝播を使う。

2009 年 4 月 12 日
情報システム工学科
塩貝亮宇

5　5

連続的なデータについて、その最初のレベルの RBM の隠れ素子はバイナリのままである、しかし可視素子はガウシアンノイズをともなう線形素子によって置き換えられる (10)。

もし、このノイズが素子の分散を持つならば、

その特徴の隠れ素子についての統計的な更新則

と

素子の分散と平均 $b_i + \sum_j h_j w_{ij}$ のガウス分布から抽出されるために可視素子 i について更新則。

6　6

我々の全ての実験において、全ての RBM の可視素子は真値の活動を持ち、ロジスティック素子について $[0, 1]$ の範囲内である。

高階の RBM を訓練している間中、

前の RBM で隠れ素子の活動の振舞いがセットされる、

しかし、全ての RBM の隠れ素子。

そのトップの RBM の隠れ素子は

RBM の持つロジスティック可視素子から入力によるガウシアンの分散から描かれる統計的な真値状態を持つ。

これは連続的な値と PCA との促進を比較する。

事前訓練の詳細と fine-tuning は (8) で見つけることができる。

7　7

事前訓練のアルゴリズムは deep network を fine-tune することを許し、

私たちはランダムに 2 次元で選択される 3 つから生成される'curves' の画像を含んでいる合成データのセット上でとても深い autoencoder を訓練した (8)。

この訓練セットについて、固有の次元は既知であり、そして、

画素の明度と 6 つの数は高階の非線形であるそれらを生成するために使われる。

0 から 1 の範囲にある画素の明度、そしてそれは非ガウシアン、

だから autoencoder でロジスティック出力素子を使った、

そして、学習の fine-tuning ステージを最小化した

クロスエントロピー誤差

$$\left[-\sum_i p_i \log \hat{p}_i - \sum_i (1 - p_i) \log(1 - \hat{p}_i) \right] \tag{3}$$

ここで、p_i は画素 i の明度、$\hat{p}_i$ は reconstruction の明度である。

8　8

autoencoder は 28x28 サイズ 400-200-100-50-25-6 の層を伴なう encoder を構成し、対称のデコーダだ。

コード層の 6 素子は線形で、そして他の多くの素子はロジスティックだ。

そのネットワークは 20,000 画像で訓練され、そして 10,000 の新しい画像でテストされる。

autoencoder は各 784 画素の最も完全に復元する真値 6 つ数をどのように変換するか発見する (図 2A).

PCA は、より悪い復元を与える。

事前訓練無しで、とても深い autoencoder は常に訓練データの平均を復元し、なおかつ、曳索された fine-tuning(8)。

データとコード間の単層の隠れ層を伴なう浅い autoencoder は、しかし事前訓練は非常によくそれら全ての訓練時間を減少する (8)。

パラメータの数が同様であるとき、deep autoencoder は他の shallow なものよりもより低い復元エラーで生成することができ、しかし、この長所はパラメータの増加に伴ない消失する (8)。

9 9

次に、我々は MNIST 訓練セット上の全ての手書き文字に向けた余分なコードへ向けて 784-1000-500-250-30 autoencoder を使う (11)。

事前訓練と fine-tuning に使用した Matlab コードは (8) に在る。

再び、全ての素子はコード層で 30 線形素子について除かれるロジスティックがあった。

全ての 60,000 訓練画像で fine-tuning した後、その autoencoder は新しい画像 10,000 でテストされ、そして、PCA した復元でよりよいが生成される。(Fig.2B)

2 次元 autoencoder は最初の主要な構成の 2 層より良いデータの可視化を生成する。(Fig.3)

10 10

私たちは

Olivetti 顔画像データセットからグレイスケール画像パッチについて 30 次元を発見するための線形素子を伴なう 625-2000-1000-500-30 autoencoder 使った (12)。

その autoencoder は明らかに PCA よりパフォーマンスが優れている。

11 11

ドキュメントで訓練したとき、autoencoder は速い回復を許すコードを生成する。

我々は newswire story 804,414 のそれぞれを 2000 の commonest word stem の確率のドキュメント指定ベクトルとして提示し (13)、そして、fine-tuning についてマルチクラスのクロスエントロピー関数

$$\sum_i p_i \log \hat{p}_i \tag{4}$$

の使用を伴なうストーリーの半分で 2000-500-250-125-10 autoencoder を訓練した。その 10 コード素子は線形で、そして隠れ素子はロジスティックのままである。

2 つの測度を使ったコード間の角度の余弦が同様の測度として使われたとき、その autoencoder は latent semantic analysis(LSA) より明らかにパフォーマンスが優れてい (14)、よく知られているドキュメント回復手法は PCA をベースにしている。(Fig.4)

autoencoder(8) もまた local linear embedding よりパフォーマンスが優れており、最近の非線形次元はアルゴリズムで減少する。(15)

12　12

層による層の事前訓練もまた分類と回帰に使うことができる。

MNIST 手書き数字認識処理の幅広く使われている版で、

その最良なエラー率の報告はランダムに初期化したバックプロパゲーションで 1.6%、サポートベクターマシンで 1.4% だ。

784-500-500-2000-10 ネットワークで層による層の事前訓練後、緩やかな勾配と小さな学習率を用いることは 1.2% に至る。

事前訓練は一般化を助ける、なぜならばモデリングイメージに由来する重みの情報の全体。

そのラベルでとても制限されたの情報は事前訓練によって発見される重みをちょうど、わずかにだけ使用される。

13　13

これは深い autoencoder を通るバックプロパゲーションは非線形次元の減少についてとても効果的であることは 1980 年代から明らかで、

計算機は十分に高速に動作することが提供され、訓練セットは十分に大きく、そして初期重みは良い解答に十分に近い。全ての 3 つの状態は、今、十分である。

ノンパラメトリックな手法とは異なり (15,16)、autoencoder はデータとコード空間の間もそれらが提供される広範のデータセットもマッピングが与えられ、

そして、fine-tuning のスケールは訓練ケースの数を伴なう時間と空間で線形だ。