2010 年 11 月 2 日
情報システム工学科 塩貝亮宇

1 Abstract

多数の隠れ層を持つボルツマン機械の新しい学習アルゴリズムについて紹介する.

データに依存する期待値はシングルモードに焦点を当てた傾向の変動する近似 (variational approximation) により推定される. そして, データに依存しない期待値は (持続する—我慢する—不屈の—不変の) マルコフ連鎖によって推定される.

2 つの全く異なる

対数尤度の勾配法は

推定について予期されるテクニックの使用は

ボルツマン機械の学習, そして共に複数の隠れ層と 100 万のパラメータを .

学習はより効果的に層と層の間で使われる.

層と層の間で ('single bottom up pass') により初期化される変動的な影響をを許す事前学習 ('pre-training') のフェーズでの学習はより効果的である. (層と層の間の学習にシングルボトムアップからの初期化「事前学習」)

The learning can be made more efficient by using layer-by-layer 'pre-training' phase that allows variational inference to be initialized with a single bottomup pass.

変動による影響を許す.

に使われるためシングルモードに集中して,

持続するマルコフ連鎖.

その学習は

私たちは, MNIST と NORB のデータセット

ディープボルツマンマシン学習が良い生成モデルであることを手書き数字と可視オブジェクトの認識タスクでのパフォーマンスの良い実験結果を提示する.

'pre-training' フェーズでは変化を許す inference efficient estimate 推量 approximate おおよその variational 変化の、変動の single mode 単一モード gradient 勾配 log-likelihood 対数尤度

persistent 存在し続ける, 持続する MNIST NORB ディープボルツマン機械の学習は良い生成モデルかつ手書きのデジタル文字と可視オブジェクトの理解タスク

2 Introduction

ボルツマン機械のオリジナルの学習アルゴリズム (Hinton and Sejnowski,1983) はランダムに初期化されたマルコフ連鎖を必要とした. それら (ランダムに初期化されたマルコフ連鎖) はデータ依存とデータ非依存の期待値を推定する.

データ依存とデータ非依存の期待値を推定する. バイナリ変数のペアの結合 (平衡分布に近づくためにバイナリ値のペアの結合) 平衡分布を推定するために必要とするアプローチだ.

データ非依存の期待値バイナリ値の結合が互いに

ランダムなマルコフ連鎖の初期化によるアプローチとデータ依存とデータ非依存から予測される 2 値のペアによって接続される. 予期される勾配法による学習では, 最大尤度学習による勾配が必要とされている.

2 つの期待値の違いは最大尤度学習について必要とされる勾配だ. シミュレーテッドアニーリングの助けと同等で, この学習の手続きは実際には非常に遅い.

学習方法は 'RBM'ではより一層効果的にはたらくが, それは隠れ素子間に結合を持たない.

多層の隠れ層は, ひとつの RBM の隠れた活動として扱うことで学習可能できる. 高階 RBM の訓練データとして. (Hinton et al.,2006;Hinton and Salakhutdinov,2006)

しかしながら, 多層がこのように貪欲に学習させられ, 層と層の方法, 入れ子モデルの結果は多層のボルツマン機械と同様の結果にならない. (Hinton et al.,2006)

これは ‘deep brief net‘と呼ばれる (ハイブリッド—合成—掛け合せ) 生成モデルで最上位の 2 つの層に無方向の結合, それの下層は全て下方向の結合を持つ.

この論文では, よりいっそう効果的な全結合の ‘general Boltzmann machine‘ の学習手続きを提案する.

わたしたちはもし, 隠れ素子の形成する多層がわずかに変化した RBM の持つスタックをディープボルツマン機械の重みを初期化して

隠れ素子間の結合が制限されて,

示す. わたしたちの (提案する) 新しい学習手続きについて説明する前に,

3 Boltzmann Machine BM's

ボルツマン機械は確率的なバイナリ ((0, 1)) 素子の結合による対称結合網である.

これは可視素子 $\boldsymbol{v} \in \{0,1\}^D$ と隠れ素子 $\boldsymbol{h} \in \{0,1\}^P$ の組を含む.(図 (1) 参照)

状態 $\{\boldsymbol{v}, \boldsymbol{h}\}$ のエネルギーは以下のように定義され,

$$E(\boldsymbol{v}, \boldsymbol{h}; \theta) = -\frac{1}{2}\boldsymbol{v}^T \boldsymbol{L}\boldsymbol{v} - \frac{1}{2}\boldsymbol{h}^T \boldsymbol{J}\boldsymbol{h} - \boldsymbol{v}^T \boldsymbol{W}\boldsymbol{h}, \tag{1}$$

ここで,$\theta = \{\boldsymbol{W}, \boldsymbol{L}, \boldsymbol{J}\}$ はモデルのパラメータ:(表記を明確にするため, バイアス項を除く)$\boldsymbol{W}, \boldsymbol{L}, \boldsymbol{J}$ として表される, $\boldsymbol{W}$ を可視層から隠れ層, $\boldsymbol{L}$ を可視層から可視層, $\boldsymbol{J}$ を隠れ層から隠れ層を表現する相互対称項 (?) とする.

$\boldsymbol{L}$ と $\boldsymbol{J}$ の対角成分には 0 が代入される.

このモデルに代入される可視ベクトル $\boldsymbol{v}$ の確率は

$$p(\boldsymbol{v}; \theta) = \frac{p^*(\boldsymbol{v}; \theta)}{Z(\theta)} \tag{2}$$

$$= \frac{1}{Z(\theta)} \sum_h \exp(-E(\boldsymbol{v}, \boldsymbol{h}; \theta)), \tag{3}$$

$$Z(\theta) = \sum_{\boldsymbol{v}} \sum_{\boldsymbol{h}} exp(-E(\boldsymbol{v}, \boldsymbol{h}; \theta)), \tag{4}$$

ここで p^* は非正規確率で $Z(\theta)$ は分配関数を表わす.

条件付き分布を越える隠れ素子と可視素子は

$$p(h_j = 1 | \boldsymbol{v}, \boldsymbol{h}_{-j}) = \sigma(\sum_{i=1}^{D} W_{ij} v_i + \sum_{m=1}^{P} \tag{5}$$

$$jJ_{jm} h_j), \tag{6}$$

$$p(v_i = 1 | \boldsymbol{h}, \boldsymbol{v}_{-i}) = \sigma(\sum_{j=1}^{P} W_{ij} h_j + \sum_{k=1}^{D} \tag{7}$$

$$iL_{ik} v_j), \tag{8}$$

ここで,$\sigma(x) = 1/(1 + exp(-x))$ はロジスティック関数である.

変数の更新は,Hinton と Sejnowski(1983) の提案で, それらは式 (2) から導出できる対数尤度の勾配を必要とし:

$$\Delta \boldsymbol{W} = \alpha(E_{P_{data}}\left[\boldsymbol{v}\boldsymbol{h}^T\right] - E_{P_{model}}\left[\boldsymbol{v}\boldsymbol{h}^T\right]), \tag{9}$$

$$\Delta \boldsymbol{L} = \alpha(E_{P_{data}}\left[\boldsymbol{v}\boldsymbol{v}^T\right] - E_{P_{model}}\left[\boldsymbol{v}\boldsymbol{v}^T\right]), \tag{10}$$

$$\Delta \boldsymbol{J} = \alpha(E_{P_{data}}\left[\boldsymbol{h}\boldsymbol{h}^T\right] - E_{P_{model}}\left[\boldsymbol{h}\boldsymbol{h}^T\right]), \tag{11}$$

ここで,α は学習率をあらわし,
$E_{P_{data}}[\cdot]$ は全てのデータの分布

$$P_{data}(\boldsymbol{h},\boldsymbol{v};\theta) = p(\boldsymbol{h}|\boldsymbol{v};\theta)P_{data}(\boldsymbol{v}), \tag{12}$$

$$P_{data}(\boldsymbol{v}) = \frac{1}{N}\sum_{n}\delta(\boldsymbol{v}-\boldsymbol{v}_n) \tag{13}$$

$$\tag{14}$$

の関係を示しており, 経験分布 $E_{P_{dmodel}}\cdot$ データ依存期待値 $E_{P_data}\cdot$ とモデルの期待値 $E_{P_model}\cdot$.

そのうちデータに依存する期待値として $E_{P_{data}}[\cdot]$, モデルの期待値として $E_{P_{model}}[\cdot]$ について言及する.

このモデルでは, 正確な最尤推定学習 (‘maximum likelihood learning‘) は扱いにくい. なぜならば, データ依存の期待値もモデルの期待値も正確な計算には隠れ層の素子の数の累乗倍の時間がかかるからだ.

Hinton と Sejnowski(1983) は (データ依存とモデルの) 期待値の両方を近似させるためのアルゴリズムにギブスサンプリングを使うことを提案した.

それぞれの学習の反復について, 離散マルコフ連鎖 (‘separate Markov chain‘) は全ての訓練データベクトルを近似するために $E_{P_{data}}\cdot$ を実行し, そして, 追加の鎖は $E_{P_{model}}\cdot$ を近似する.

a separate Markov chain is run for every training data vector to approximate $E_{Pdata}[\cdot]$,and an additional chain is run to approximate $E_{Pmodel}[\cdot]$.

この学習アルゴリズムにおける主要な問題は定常分布に至るのに時間がかかる. とくに, モデルの期待値を推定するとき, ギブス鎖は高い多数モデルのエネルギーの谷を探索する必要があるだろう.

これは現実世界のモデリングの分布例えば

ほとんど全部の画像が極端に低い確率の画像データセット, しかし, かなり似た確率で発生する多くのとても異なったイメージがある.

$\boldsymbol{J}=0,\boldsymbol{L}=0$ と設定し, よく知られている Restricted Boltzmann machine(RBM) モデルに直す.(Smolensky,1986)(図 1, 右)

一般的なボルツマン機械と比較して RBM's の推定は正確だ.

RBM の正確な最大尤度推定学習 (‘maximum likelihood learning‘) とはいえ, まだ扱いにくく, 学習は Constrastive Divergence を実行することで, 効果的に実行できる.(Hingon,2002)

更なる観測として (Welling と Hinton,2002;Hinton,2002) Constrastive Divergence のパフォーマンスが良いことが観測された. 条件付分布 $p(\boldsymbol{h}|\boldsymbol{v};\theta)$ から正確な標本を獲得することが重要で, 完全結合のボルツマンマシンの学習のときは扱いにくいものだった.

4 Using Persistent Markov Chains to Estimate the Model's Expectations

CD learning の代替として, モデルの期待値を近似するために stochastic approximation procedure(SAP) を使用することができる. (Tieleman,2008;Neal,1992).

SAP は Robbins-Monro 型の統計近似アルゴリズムに属する. (Robbins and Monro,1951; Younes,1989,2000). そのアイディアの背景は直接的だ.

θ_t を現在のパラメータ, X^t を現在の状態としよう. このとき,X^t と θ^t は以下のように順番に更新される.

- X^t が与えられ, 新しい状態 X^{t+1} は不変な p_θ をもつ遷移オペレータ $T_{\theta_t}(X^{t+1};X^t)$ から抽出される.

- 新しいパラメータ θ_{t+1} は扱いにくいモデルの期待値を X^{t+1} と関係する期待値によって置き換えられる. A new parameter θ_{t+1} is then obtained by replacing the intractable model's expectation by the expectation with respect to X^{t+1}

十分に正確な量の状態ほとんど漸近的な点に収束する点は保障する漸近的に安定な点は与えられる. Precise sufficient condtion that gurantee almost sure convergence to an asymptotically stable point are given in.

ひとつの必要とされる状態は時間による学習率の低下で,

$$\sum_{t=0}^{\infty} \alpha_t = \infty \tag{15}$$

$$\sum_{t=0}^{\infty} \alpha_t^2 < \infty \tag{16}$$

$$\tag{17}$$

この状態は,

$$\alpha_t = \frac{1}{t} \tag{18}$$

と設定することを満足させる.

一般的に, 実際にやってみると, 数列 $\|\theta_t\|$ は束縛され, そして, マルコフ鎖, 遷移核 ('transition kernel')T_θ によって得られ, エルゴート的だ.

学習率の状態と一緒で, ほぼ確実に収束する.

直感的に何故この次のようなのだろう: 学習率は

5 A Variational Approach to Estimationg the Data-Dependent Expectations

変分学習 (*variationallearning*)(Hinton and Zemel, 1994; Neal and Hinton, 1998), 各訓練ベクトルについて潜在的な変数からの真の事後分布 $p(\boldsymbol{h}|\boldsymbol{v};\theta)$ は近似事後 $q(\boldsymbol{h}|\boldsymbol{v};\mu)$ に置き換えられ, そしてパラメータは次式の対数尤度上の下限の勾配により更新される.

$$\ln p(\boldsymbol{v};\theta) \leq \sum_{\boldsymbol{h}} q(\boldsymbol{h}|\boldsymbol{v};\mu) \ln p(\boldsymbol{v},\boldsymbol{h};\theta) + \mathcal{H}(q) \tag{19}$$

$$= \ln p(\boldsymbol{v};\theta) - KL\left[q(\boldsymbol{h}|\boldsymbol{v};\mu)\|p(\boldsymbol{h}|\boldsymbol{v};\theta)\right], \tag{20}$$

ここで,$\mathcal{H}(\cdot)$ はエントロピー関数をあらわす. 変分学習 (*variationallearning*) は訓練データの対数尤度を最大化しようとする良い手続きで, これは近似値と真の事後確率の間の Kullback-Leibler divergences を最小化するパラメータを見つける.

平均場 ('naive mean field') のアプローチを用いることで, 必要とする近似事後確率を全て因数分解し

$$q(\boldsymbol{h};\mu) = \Pi_{j=1}^{P} q(h_i), with q(h_i = 1) = \mu_i \tag{21}$$

ここで P は隠れ素子の数である.

対数確率の下限は次式の形体をとり,

$$\ln p(\boldsymbol{v};\theta) \geq \frac{1}{2}\sum_{i,k} L_{ik} v_i v_k + \frac{1}{2}\sum_{j,m} J_{jm}\mu_j\mu_m \tag{22}$$

$$+ \sum_{i,j} W_{ij} v_i \mu_j - \ln Z(\theta) \tag{23}$$

$$+ \sum_{j} \left[\mu_j \ln \mu_j + (1-\mu_j)\ln(1-\mu_j)\right]. \tag{24}$$

θ について修正した変動パラメータ μ に関係する下限を最大化しながら学習は進んでいき, 平均場の修正点の方程式は

$$\mu_j \leftarrow \sigma(\sum_i W_{ij} v_i + \sum_m \tag{25}$$

$$jJ_{mj}\mu_m) \tag{26}$$

これは SAP をモデルパラメータ θ を適応による更新 (Salakhutdinov, 2008).

変動近似 ('variational learning') がボルツマンマシンの学習則でモデルの近似した期待値と一緒近似した期待値を使用することができないことを強調する. なぜならば, 負の符号 (式 6) は近似値と真の分布の間のダイバージェンスを最大化するためにパラメータを変化させるための変動学習を引き起こすだろう.

もし, しかしながら, モデルのもつ期待値の推定に持続する鎖 ('persistent chain') が使われ, データ依存の期待値を推定するために変動学習は適応される.

naive mean-field の選択はよく考えられている. まず, 収束が非常に速く, 学習をとても容易にする.

2 つ目に, イメージやスピーチなどの補間に適応され, シングルモードをとるための隠れ状態によって与えられる.

実際は, 真の 'posterior unimodel'を

それの活動を事後確率を使うシステムについて.

同じセンサーの良い表現は入力を対数尤度を増加させるが, より

そのセンサー入力の適当な行動は.

6 Deep Boltzmann Machine(DBM's)

一般に, 複雑な学習に興味を持つことは稀で, 完全に結合されたボルツマンマシン.

代わりに, 図 (2) の左に示した Deep Multilayer Boltzmann machine を考える. この Deep Multilayer Boltzmann machine は, それぞれの層はとらえることが困難で, その層より下位の隠れた特徴との間を高階に相関をもつ.

Deep Boltzmann machine は様々な理由から興味深い. まず, 'deep blief network'に似ており, 複雑さを増加させる中間層の表現での学習の可能性が物体や音声認識の問題を解く見込みある方法として考えられる.

つぎに, 高階表現は分類されていないセンサーからの入力の供給と非常に制限された分類されたデータを手からの指定されたタスクについてモデルをわずかに微調整する.

最後に,deep Belief network とはちがい, 近似を推定する手続き, 加えて最初のボトムアップパス, トップダウンフィードバックを組み込むことができ, ディープボルツマンマシンは不確かな増殖について, そして, より一層にロバスト不明瞭な入力.

図 (2 右) の内側の層に結合のない 2 層のボルツマンマシンについて考える. 状態 $\boldsymbol{v}, \boldsymbol{h}^1, \boldsymbol{h}^2$ のエネルギーは以下のように定義し,

$$E(\boldsymbol{v}, \boldsymbol{h}^1, \boldsymbol{h}^2; \theta) \quad = \quad -\boldsymbol{v}^T \boldsymbol{W}^1 \boldsymbol{h}^1 - \boldsymbol{h}^{1T} \boldsymbol{W}^2 \boldsymbol{h}^2, \tag{27}$$

ここで,$\theta = \boldsymbol{W}^1, \boldsymbol{W}^2$ はモデルのパラメータ, 可視層から隠れ層, 隠れ層から隠れ層の対称的な項を示す.

$$p(\boldsymbol{v};\theta) = \frac{1}{Z(\theta)}\sum_{\boldsymbol{h}^1,\boldsymbol{h}^2}\exp(-E(\boldsymbol{v},\boldsymbol{h}^1,\boldsymbol{h}^2;\theta)). \tag{28}$$

$$\tag{29}$$

可視と隠れ素子の 2 つからロジスティック関数によって得られる条件付分布は

$$p(h_j^1=1|\boldsymbol{v},\boldsymbol{h}^2) = \sigma(\sum W_{ij}^1 v_i + \sum_m W_{jm}^2 h_j^2) \tag{30}$$

$$p(h_m^2=1|\boldsymbol{h}^1) = \sigma(\sum_j W_{im}^2 h_i^1) \tag{31}$$

$$p(v_i=1|\boldsymbol{h}^1) = \sigma(\sum W_{ij}^1 h_j). \tag{32}$$

$$\tag{33}$$

最尤推定学習による近似について, 我々は General Boltzmann machines についての学習手続きを, まだ, 適応することができる. しかし, それらはむしろ遅く, とくに隠れ素子が形成する層が可視層から離れて増加する.

しかしながら, モデルパラメータを速く初期化する方法がある.

次の章では利にかなったモデルパラメータの初期化について説明する. sensible 利にかなった

7 Greedy Layerwise Pretraining of DBM's

Hinton et al.(2006) は層と層は教師無し学習で RBM のもつある層のスタックを同時に学習することができる貪欲な学習アルゴリズム紹介した.

RBM の持つスタックが学習した後で, 全てのスタックは 'deep belief network'と呼ばれる 1 つの確率モデルを観測することができる.

驚くべきことに, このモデルはディープボルツマンマシンではない.

トップの 2 つの層は無方向グラフの restricted Boltzmann machine であるが, 下層は有向グラフの生成モデルを形成する (図 2 参照).

このスタック内での最初の RBM の学習後, 生成モデルは次のように書くことができる.

$$p(\boldsymbol{v};\theta) = \sum_{\boldsymbol{h}^1} p(\boldsymbol{h}^1;\boldsymbol{W}^1)p(\boldsymbol{v}|\boldsymbol{h}^1;\boldsymbol{W}^1), \tag{34}$$

$$\tag{35}$$

ここで,$p(\boldsymbol{h}^1;\boldsymbol{W}^1)=\sum_{\boldsymbol{v}} p(\boldsymbol{h}^1,\boldsymbol{v};\boldsymbol{W}^1)$ は暗黙的に事前に $\boldsymbol{h}^1$ について定義されたパラメータである.

スタック内での 2 番目の RBM は $p(\boldsymbol{h}^1;\boldsymbol{W}^1)$ を $p(\boldsymbol{h}^1;\boldsymbol{W}^2)=\sum^{\boldsymbol{h}^2} p(\boldsymbol{h}^1,\boldsymbol{h}^2;\boldsymbol{W}^2)$ に置き換える.

もし,2 番目の RBM が正しく初期化できていれば (Hinton at el.,2006), $p(\boldsymbol{h}^1;\boldsymbol{W}^2)$ は $\boldsymbol{h}^1$ による事後確率の集合よりも良いモデルとなるだろう. ここで事後確率の集合は全てのトレーニングの場合において因数分解した事後確率の単に混った場合である. よって,$1/N\sum_n p(\boldsymbol{h}^1|\boldsymbol{v}_n;\boldsymbol{W}^1)$.

2 番目の RBM は $p(\boldsymbol{h}^1;\boldsymbol{W}^1)$ からより良いモデルによって置き換えられ, これは

ボトムアップの向きに $\boldsymbol{W}^1$ をトップダウンの向きに $\boldsymbol{W}^2$ を使うことは, $\boldsymbol{v}$ に依存する $\boldsymbol{h}^2$ の証言 ('evidence') をダブルカウントすることになるだろう.

DBM のモデルパラメータを初期化するために, 貪欲な,RBM の持つスタックの学習による層と層の事前学習を提案する. しかし, (層と層を) 結合した後のとき, トップダウンとボトムアップからのダブルカウントによる小さな変更の影響は除かれる.

下位レベルの RBM について, 2 倍の入力と可視層から隠れ層への重み, 図 2 右に示す. 変更した RBM のパラメータの繋りは隠れと可視の状態について次のような事後確率が定義される.

$$p(h_j^i = 1|\boldsymbol{v}) = \sigma(\sum_i W_{ij}^1 v_i + \sum_i W_{ij}^1 v_i), \tag{36}$$

$$p(v_i = 1|\boldsymbol{h}^1) = \sigma(\sum_j W_{ij}^1 h_j). \tag{37}$$

Contrastive divergence learning はうまく働き, そして, 変更した RBM は訓練データを再構築するのに良い.

反対に,RBM のトップレベルについては隠れ素子の数を 2 倍にする.

このモデルについての条件付分布は次の形をとる.

$$p(h_j^1 = 1|\boldsymbol{h}^2) = \sigma(\sum_m W_{jm}^2 h_m^2 + \sum_m W_{jm}^2 h_m^2) \tag{38}$$

$$p(h_m^2 = 1|\boldsymbol{h}^1) = \sigma(\sum_j W_{jm}^2 h_j^1). \tag{39}$$

これら 2 つのモジュールが 1 つのシステムとして落ち付いたとき, 最初の隠れ層への総合的な入力は次式の $\boldsymbol{h}^1$ についての条件付分布の半分になる.

$$p(h_j^i = 1|\boldsymbol{v}, \boldsymbol{h}^2) = \sigma(\sum_i W_{ij}^i v_i + \sum_m W_{jm}^2 h_m^2) \tag{40}$$

$$\tag{41}$$

v と h^2 についての条件付分布は式 (16),(18) に定義したままである.

composed model によって定義された条件付き分布は DBM によって定義された式 (11,12,13) の条件付分布と同じものとなる.

従って, 意欲的な事前訓練は 2 つを変更した RBM 対称の重みをもつ無方向グラフモデルを導く-deep boltzmann machine.

2 つ以上の RBM のスタックを貪欲に事前学習させたとき, スタックの最初と最後の RBM は変更のためだけに必要となる.

RBM の全ての仲介人はどちら向きの重みであってもディープボルツマンマシンの形体をとる入れ子になったとき, それらの重みを半分にする.

こうした手法による貪欲な事前訓練による DBM の重みは 2 つの役目を果たす.

まず, 累乗的な面, 気の効く値でそれを初期化. 2 つ目, RBM のスタックのパスを上向きに通ることによって近似の影響が非常にはやくなることが保障される.

可視素子上でデータベクトルを与えると, 各層における隠れ素子は. (トップダウンからの入力がない一番上の層を除く)

この高速な推定の近似は平均場法として初期化され, ランダムに初期化された場合よりも高速に収束する.

8 Evaluting DBM's(DBM の評価)

近年,Salakhutdinov と Murray(2008) はモンテカルロに基づいた, Annealed Importance Sampling(AIS), RBM の機能の一部として効果的な推定を用いた.

この章では, AIS が deep boltzmann machine の機能の一部を推測するのに効果的に用いることができることを示す.

変動推測 (‘variational inference‘) と一緒で, これはテストデータの対数確率上で下限の良い推定が獲得できることを許す.

2 つの分布を仮定する確率密度関数：$p_A(x) = p_A^*/Z_A$ と $p_B(x) = p_B^*/Z_B$ より, いくつかの空間 から 2 つの分布を定義する.

一般的に $p_A(\boldsymbol{x}) = p_A^*(\boldsymbol{x})/Z_A$ はいくつかの単純な既知の分布 Z_A と容易に i.i.d な標本を描くことができる.

AIS は Z_B/Z_A を仲介となる分布の数列を用いて $p_0, \ldots, p_k$ と $p_0 = p_A$ と $p_K = p_B$. の比より推定される.

それぞれの仲介となる分布は, 非正規確率

容易に評価することができる. そして, マルコフ連鎖の遷移オペレータ $T_k(\boldsymbol{x}'; \boldsymbol{x})$ を $p_k(\boldsymbol{x})$ を不変なまま用いることで $\boldsymbol{x}'$ から $\boldsymbol{x}$ の標本を得ることができる.

‘deep Boltzmann machine‘の層と層の特定の構造を用いることは, モデルの持つ機能の一部を推定するために, より一層効果的に AIS scheme で導き出せる.

$$p_k(\boldsymbol{h}^1) = \sum_{\boldsymbol{v}, \boldsymbol{h}^2} p(\boldsymbol{v}, \boldsymbol{h}^1, \boldsymbol{h}^2) =$$
$$\frac{1}{Z_k} \Pi_i (1 + \exp(\beta_k \sum_j h_j^1 W_{ij}^1)) \Pi_k (1 + \exp(\beta_k \sum_j h_j^1 W_{jk}^2)) (42)$$

このアプローチはシミュレーティッドアニーリングに近い. β_k を 0 から 1 へと徐々に変化させ (または, 逆温度 ‘inverse tempature‘) 単純な ‘uniform model‘から最終的な複合モデルに変化させていく. 式 (11,12,13) を用いることは, ギブス遷移のオペレータを不変的な $p_k(\boldsymbol{h}^1)$ のまま直接的に導き出すことができる.

かつて, 大域分配関数 (‘global partion function‘)$\hat{Z}$ の推定から獲得し, テストケース $\boldsymbol{v}^*$ を与えることによって式 (7) の下限を推定できた.

$$lnp(\boldsymbol{v}^*; \theta) \geq -\sum_{\boldsymbol{h}} q(\boldsymbol{h}; \mu) E(\boldsymbol{v}^*, \boldsymbol{h}; \theta) + \mathcal{H}(q) - \ln Z(\theta) \quad (43)$$
$$\approx -\sum_{\boldsymbol{h}} q(\boldsymbol{h}; \mu) E(\boldsymbol{v}^*, \boldsymbol{h}; \theta) + \mathcal{H}(q) - ln, \quad (44)$$
$$(45)$$

ここで,$\boldsymbol{h} = \boldsymbol{h}^1, \boldsymbol{h}^2$ を定義する. それぞれのテストベクトルについて, この下限は平均場の更新方程式を使うことで変動パラメータ μ の値を最大化することができる.

さらに, 明示的に隠れ素子 $\boldsymbol{h}^2$ の状態を足し合わせることで, 我々はテストデータの対数確率の下限を獲得することができる.

もちろん, $\sum_{\boldsymbol{h}^1, \boldsymbol{h}^2} p^*(\boldsymbol{v}, \boldsymbol{h}^1, \boldsymbol{h}^2)$ を推定するために AIS を適応することができる. そして, 実際に真のテストデータの対数確率から大分配関数を推定することができる.

これはしかしながら, 計算的に高価で. それぞれのテストデータに離散 AIS(‘separate AIS‘) を実行することで, 良いパフォーマンスを得ることができる.

2 つ以上のディープボルツマンマシンの層を学習させたとき, 奇数か偶数の層を明確に足し合わせることができる.

この結果はモデルの分配関数とテストデータの対数確率のしっかりとした下限の良い推定結果を与えるだろう.

9 Discriminative Fine-tuning of DBM's

学習させた後に, 各層におけるバイナリの統計的活動の特徴は (決定した—決定された—確立した) ものに置き換えることができ, 実数の確率, (そして—と), ディープボルツマンマシンは以下の方法を用いることで (決定された) 多層のニューラルネットワークを初期化する.

それぞれの入力ベクトル $\boldsymbol{v}$, 平均場の影響 (?mean field inference) は, 事後確率 $q(\boldsymbol{h}|\boldsymbol{h})$ の近似を用いて獲得されていた. 事後確率 $q(h_j^2|\boldsymbol{v})$ の近似の周辺確率, このデータと一緒に,「増大された」('augmented') 図 (3) に示す深層多層ニューラルネットワーク (deep multilayer NN) への入力を作成していた. 標準的な誤差逆伝播法は識別的な微調整するモデルに (使われるだろう—使うことができる—用いられる—使われていた).

その入力の (独自—特徴的) な表現は識別的なニューラルネットワークから 'DBM'に変換できることだ. 一般的に勾配法を元にした微修正 (gradient-based fine-tuning) は $q(\boldsymbol{h}^2|\boldsymbol{v})$ の選択を気にしないかもしれない. この事後確率の近似の周辺分布 $q(h_j^2 = 1|\boldsymbol{v})$, データと一緒に,

したがって, $\boldsymbol{W}^2$ の最初の層の進行 (drive) はゼロに向かい, 標準的なニューラルネットワークの結果を得るだろう. 反対に, ネットワークは最初の層の $\boldsymbol{W}^1$ からの進行はゼロに向かうだろう.

これらの実験から, しかしながら, ネットワークは全体の作成された予測について増大された入力を用いる.

10 Experimental Results

実験では,MNIST と NORB のデータセットを用いる.

学習を早めるために, データセットをそれぞれが 100 個の場合を含んでいる 'mini-batche'に再分割し, そして, 各 'mini-batch'の重みを更新する. 'fantasy particles'の数は 100^2. (確率 ‖ 統計) 的近似アルゴリズムについて,'$fantasyparticle$'の数は常に '$5Gibbsupdate$'を用いる.初期の学習率は0.005にセットされ, 徐々に 0 に向って減少していく.DBM のもつ識別的な微調整 ('$discriminativefine-tuning$') には, 私たちはそれぞれの世代の 5000 より広い '$mini-batch$'について 3 行ずつ探索をおこない,

11 MNIST

MNIST デジタル文字セットは 60,000 の訓練データと 10,000 のテスト画像 28×28 画素の 0　9 の手書き文字から出来ている. 我々の最初の実験では, 2 つのディープボルツマン機械：ひとつは 2 つの隠れ層 (500 個と 1000 個の素子) を持つものと, もうひとつは 3 層の隠れ層を持つ (500,500,1000 個の隠れ層) を持つ (図 (4)).

モデルの持つ部分関数の推定に, β_k のぼんやりした ('spaced') 一様な 0　1.0.

テーブル 1 はテストの平均上での対数確率の下限 ('2-と 3-層の ') 個々のボルツマン機械の推定値を示している.

この結果は比較してわずかに優れている.

2 層の 'Deep Belief network'.(Salakhutdinov and Murray,2008).

2 つの 'DBM'を観測すると,0.9 以上 1.15 百万のパラメータを含み, 表われない 'overfitting'. 訓練データの推定値とテストの対数確率の違いは, 図 (4) の示すサンプルは 2 つの全てのバイナリ状態がランダムに初期化され, そして, 100,000 ステップ実行したギブスサンプリングから生成された. 確かに, 全ての標本は本物の手書き文字のように見える. また, 貪欲な事前学習 ('greedy pretraining') なしで, MNIST デジタル手書き文字の 'DBM'モデルの学習を成功させることはできなかったことに留意する.

失なわれた変動束縛 ('variational bound') が推定することが

最終的に識別的な微調整 (‘discriminative fine-tuning‘) は全ての MNIST テストセットでは, エラー率は BM の 0.95% に達成する. これは, わたしたちの知るところで,MNIST タスクの順列の不変性の結果となる.

3 層 BM は 1.01% のエラー率よりわずかに悪い.

これは, 比較すると SMV’s の 1.4% に至り (Decoste and Scholkopf,2002), ランダムに初期化した誤差逆伝播の 1.6% に至り, ‘deep blief network‘の 1.2% に至る (described in Hinton et al.(2006)).

12 NORB

MNIST の結果を示すと多くの他のモデルの方がわずかに良い結果を示すが, 比較的に手書き文字認識の単純なタスクは,.

この章では, NORB での結果を示し, MNIST のより異なる結果を示す.

NORB(LeCun et al.,2004) は

車, トラック, 飛行機, 動物, 人間の 5 つの物体ごとに.

各オブジェクトはそれぞれ異なる視点から撮影され, 様々な光源の位置もとで変更する.

それぞれの訓練データは 25 オブジェクトのステレオ画像のペアと 5 つのクラス, 24, 300 ステレオペアが,25 個の異なるオブジェクト. ゴールは前もって観測されない一般的なオブジェクト.

それぞれの画像は $[0, 255]$ の範囲の整数のグレースケール値 96×96 画素.

実験を早めるために, 画像の 9216 から 4488

ascent 上がること

diagonal 斜めの

stochastic 確率の

13 Conclusions

多層ボルツマンマシンの訓練アルゴリズムについて提示し, それは良い生成モデルの学習であることを示した. この手続きは ‘real-value‘,‘count‘,‘tabular data‘とともにボルツマンマシン学習の拡張をおこない累乗系 (‘exponential family‘) の分布を提供した.(Welling et al.,2005) AIS シミュレータが, 変動推測に沿う, テストデータを与えられる多数の隠れ層をもつボルツマンマシンの対数確率の下限の推定値を使用することができる.

最終的に特徴的な (‘discriminatively‘) はよく調整された DBM’s のパフォーマンスは MNIST と NORB 3D object の認識タスクで良い結果を示した.