情報システム工学科
2010 年 11 月 2 日
塩貝亮宇

1 Notes on Contrastive Divergence

Contrastive Divergence について記述するが, これは Geoffrey Hinton より提案された近似最尤学習アルゴリズムだ.

1.1 What is CD, and why do we need it?

点データの確率のモデル
x を用いて $f(x;\Theta)$ の形式をとるここで Θ はモデルパラメータのベクトルだ. 仮定しよう.
x の確率,$p(x;\Theta)$ はすべての x について 1 に向かって積分する必要があり,

$$p(x;\Theta) = \frac{1}{Z(\Theta)} f(x;\Theta) \tag{1}$$

ここで,$Z(\Theta)$, 分配関数として知られ, 以下のように定義される.

$$Z(\Theta) = \int f(x;\Theta) dx \tag{2}$$

モデルパラメータ Θ を学習させ, データの訓練セットの確率を最大化し, $\boldsymbol{X} = x_1, \ldots, K$ は

$$p(\boldsymbol{X};\Theta) = \Pi_{k=1}^{K} \frac{1}{Z(\Theta)} f(x_k;\Theta) \tag{3}$$

または, 同等に,$p(\boldsymbol{X};\Theta)$ の負の対数を最小化し, $E(\boldsymbol{X};\Theta)$ と示し, 当然, エネルギー

$$E(\boldsymbol{X};\Theta) = logZ(\Theta) - \frac{1}{K} sum_{k=1}^{K} logf(x_k;\Theta) \tag{4}$$

まず, 確率のモデル関数を選択し,$f(x;\Theta)$, 正規分布の pdf(probability density function),$N(x;\mu,\sigma)$ となるだろう. だから,$\Theta = \{\mu,\sigma\}$. pdf の積分は 1(標準的な結果, したがって些細ではない), だから,$logZ(\Theta) = 0$.
式 4 を微分すると訓練データの平均,$\boldsymbol{X}$, と最適な μ と関連する
そして, 同様に σ に最適な σ は訓練データの分散の平方根
たまに, この場合では, 個々のエネルギー関数を正確に最小化する手続きが存在する.
特定のパラメータ空間のエネルギー関数の.
明るく日が出る, 見ること最小点と直線に歩く.
今, 確率モデル関数 $f(x;\Theta)$ を選択すると,
N の正規分布の和, だから,$\Theta = \{\mu_i, \ldots, N, \sigma_1, \ldots, N\}$ と

$$f(x;\Theta) = sum_{i=1}^{N} N(x;\mu_i,\sigma_i) \tag{5}$$

これは,sum-of-experts や混合モデルと同等のので, 全てのエキスパート上の重みと均等で; 別の重みを持つことはモデルへの些細な拡張だ.
再び, 正規分布の 1 へ向けた積分の事実を使って, 式 2 から $logZ(\Theta) = logN$ を見ることができる.
しかしながら, (我々の) モデルパラメータの各々に関して式 4 の微分は他のモデルパラメータに依存する方程式を生み出し, だから, 最適モデルのパラメータをすぐに計算することはできない.
かわりに, 方程式の偏微分と line search によって見つけ出したパラメータ空間でのエネルギーの局所最適解による勾配降下法を用いることができる.

フィールドの比喩に話を戻すと, line search による勾配降下法はトーチと共にする夜の荒野と等価なものだ.

私たちは, 自身の立っている点をフィールドの勾配として感じることができるかどうか, さもなければ, 私たちと同じ方向の短い距離の高さの相対をトーチを使って推定する. (有限な差分を使った数値微分)

そのとき, 旅の方向でのトーチの光, その方向のフィールドのなかで最も低い点.

私たちは, その点へ向って歩くことができ, 新しい方向と距離まで歩くことを反復.

最終的に, 確率モデル関数,$f(x;\Theta)$ を選択し, N 正規分布の生産は

$$f(x;\Theta) = \Pi_{i=1}^{N} N(x;\mu_i,\sigma_i) \tag{6}$$

これは,product-of-experts model と等価なものだ.

分配関数 $Z(\Theta)$ は, もはや一定ではない.

わたしたちは, 二重正規分布からなるモデルだと見なすことができる, ともに $\sigma = 1$.

もし,$\mu_1 = -\infty$ と $\mu_2 = \infty$ ならば,$Z(\Theta) = 0$, と同時に (しながら), もし $\mu_1 = \mu_2 = 0$ ならば $Z(\Theta) = 1/2\sqrt{\pi}$.

これが可能なとき, この場合,Θ として与えられる分配関数は正確で, 式 2 の積分を仮定すると代数的にとり扱うことができない. (この場合, 他の確率モデル関数)

この場合, 式 4 を評価して数値積分を使う必要があり, そして, 局所最適解を見つけるために勾配降下法を使う.

高次データ空間の積分時間 (回数) は切り抜けられ, そして, 高次パラメータ空間はこの問題と合成する.

これは, 我々が評価できないエネルギー関数の最小を試みる.

これは CD は私たちは手助けする. エネルギー関数を評価できないとは言うものの, エネルギー関数の勾配を推定する方法を CD は提供する. 荒野の比喩に戻すとするならば, 今度は何の明りもない荒野で自分自身を発見し (i.e. エネルギー関数は計算できない),.

CD の影響は私たちにバランス感覚を与えて, 足元の荒野の勾配を感じることを許す.

方向のとても小さいステップの最急降下をとることによって, 局所最適解を見つけることができる.

2 How does CD work?

説明すると,CD はエネルギー関数の勾配を推定し, モデルパラメータ Θ, 訓練データ $\boldsymbol{X}$ の集合を与える.

勾配方程式は導出できる. 式 4 の偏微分をまず書き下せる

$$\frac{\partial E(\boldsymbol{X};\Theta)}{\partial\Theta} = \frac{\partial logZ(\Theta)}{\partial\Theta} - \frac{1}{K}\sum_{i=1}^{K}\frac{\partial logf(x_i;\Theta)}{\partial\Theta} \tag{7}$$

$$= \frac{\partial logZ(\Theta)}{\partial\Theta} - \Big\langle \frac{\partial logf(x;\Theta)}{\partial\Theta} \Big\rangle_{\boldsymbol{X}} \tag{8}$$

ここで,$\big\langle\cdot\big\rangle_{\boldsymbol{X}}$ は $\cdot$ の期待値データ分散 $\boldsymbol{X}$ を与える.

最初の項の右側は分配関数 $Z(\Theta)$ に由来するが, 式 2 に表すように,x の積分を含む.

これを代入すると,

$$\frac{\partial logZ(\Theta)}{\partial\Theta} = \frac{1}{Z(\Theta)}\frac{\partial Z(\Theta)}{\partial\Theta} \tag{9}$$

$$= \frac{1}{Z(\Theta)}\frac{\partial}{\partial\Theta}\int f(x;\Theta)dx \tag{10}$$

$$= \frac{1}{Z(\Theta)} \int f(x;\Theta) \frac{\partial log f(x;\Theta)}{\partial \Theta} dx \tag{11}$$

$$= \int p(x;\Theta) \frac{\partial log f(x;\Theta)}{\partial \Theta} dx \tag{12}$$

$$= \Big\langle \frac{\partial log f(x;\Theta)}{\partial \Theta}_{p(x;\Theta)} \Big\rangle \tag{13}$$

を得る.

議論として, この積分は一般的に代数的には取り扱うことができない.

しかしながら, 式 (14) の形式は, 提案分布 $p(x;\Theta)$ のサンプルを描く数値的に近似することができる.

サンプルは $p(x;\Theta)$ は分配関数の値を知らないために直接的に描かれない.

しかし, 提案分布から描かれるデータへ訓練データに変形するため, MCMC sampling の多くのサイクルを使うことができる.

これは 2 つの確率の比 $p(x';\Theta)/p(x;\Theta)$ の計算をともなうだけの変形可能だ. だから, 分配関数は消去できる.

$\boldsymbol{X}^n$ は MCMC の n サイクルを使って変形された訓練データを表わし, だから,$\boldsymbol{X}^0 \equiv \boldsymbol{X}$.

これを式 (8) に戻すと,

$$\frac{\partial E(\boldsymbol{X};\Theta)}{\partial \Theta} = \Big\langle \frac{\partial log f(x;\Theta)}{\partial \Theta} \Big\rangle_{\boldsymbol{X}^\infty} - \Big\langle \frac{\partial log f(x;\Theta)}{\partial} \Big\rangle_{\boldsymbol{X}^0} \tag{14}$$

が得られる.

まだ計算によるハードルをのりこえるために- 多くの MCMC サイクル正確な勾配を求めるためには長すぎる計算が必要になる.

Hinton は勾配の近似は少しの MCMC サイクルだけが必要となる. と主張する.

直感の裏付け (intuition behind) は提案分布へ向う目標分布からわずかな反復、だからアイディアの方向は訓練データの方向をよりよいモデルへ向う必要がある.

実験的にヒントンは最尤推定の解答へ向かう収束するためのアルゴリズムについて 1 回の MCMC サイクルで十分だと発見している.

だから, エネルギー関数を最小化するために下り坂へ行きたいと望むことを考慮して, パラメータの更新する方程式は次のように書け

$$\Theta_{t+1} = \Theta_t \eta \left(\Big\langle \frac{\partial log f(x;\Theta)}{\partial \Theta} \Big\rangle_{\boldsymbol{X}^0} - \Big\langle \frac{\partial log f(x;\Theta)}{\partial \Theta} \Big\rangle_{\boldsymbol{X}^1} \right) \tag{15}$$

ここで,η はステップサイズの因子であるが, 収束時間と安定性を元となり実践的に選択されるだろう.