データサイエンス・シリーズ 6

データ学習アルゴリズム

(渡辺澄夫著, 共立出版 2001)

発表日: 平成 15 年 5 月 21 日
担当者: 岩崎 唯史
担当箇所: 2.2 最適化法
2.2.1 最急降下法
2.2.2 確率項をもつ最急降下法

2.2 最適化法

損失関数 (コスト関数, 誤差, エネルギー)$L_n(w)$ を最小にするパラメータを, 直接解析的に求めるのは多くの場合困難.

$$\frac{dL_n(w)}{dw} = 0 \text{ を満たす } w \text{ を求める}$$

損失関数を最小にするパラメータを, ある操作の繰り返しによって逐次的に求めていく (最適化法).

2.2.1 最急降下法

■パラメータ w のもとでのいろいろな “誤差”

- **パラメータ:** $w \in R^d$.
- **確率変数:** $(X, Y) \in R^M \times R^N$.
- **サンプル (n 組のデータ):** $(x^n, y^n) = \{(x_1, y_1), \ldots, (x_n, y_n)\}$.
- **サンプルを発生している真の同時確率密度関数:** $q(x, y)$.
- **1 組のデータ (x, y) に対する誤差:** $\mathcal{E}(x, y, w) = -\log p(x, y|w)$.

$$\text{学習誤差: } L_n(w) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \mathcal{E}(x_i, y_i, w) = -\frac{1}{n}\sum_{i=1}^{n} \log p(x_i, y_i|w).$$

$$\text{予測誤差: } L(w) \stackrel{\text{def}}{=} \int \mathcal{E}(x, y, w) q(x, y) dxdy$$

$$= -\int q(x, y) \log p(x, y|w) dxdy.$$

$$\text{汎化誤差: } K(q||p_w) \stackrel{\text{def}}{=} \int q(x, y) \log \frac{q(x, y)}{p(x, y|w)} dxdy$$

$$= \underbrace{\int q(x, y) \log q(x, y) dxdy}_{\text{const.}} + L(w).$$

■最急降下法 (一括学習) のアルゴリズム

学習誤差 $L_n(w)$ を最小にするパラメータ w を探す場合.

離散時間 $t = 0, 1, 2, \ldots$ におけるパラメータ: $w(t) \in R^d$.

(1) 初期値 $w(0)$ を設定する (ランダム).

(2) $w(t)$ を次の漸化式により時間発展させる (η は正の実定数).

$$w(t+1) = w(t) - \eta \nabla L_n(w(t)),$$

$$\nabla L_n(w(t)) = \left(\frac{\partial L_n(w)}{\partial w_1}, \ldots, \frac{\partial L_n(w)}{\partial w_d} \right) \text{ at } w = w(t).$$

(3) 手順 (2) を繰り返し, 十分大きな時刻 T における $w(T)$ を求め, 解 (パラメータ) の候補値とする.

(4) 手順 (1)~(3) を初期値を変えて繰り返し, 得られた候補値の中で学習誤差 $L_n(w(T))$ を最小にするものを最終的な解とする.

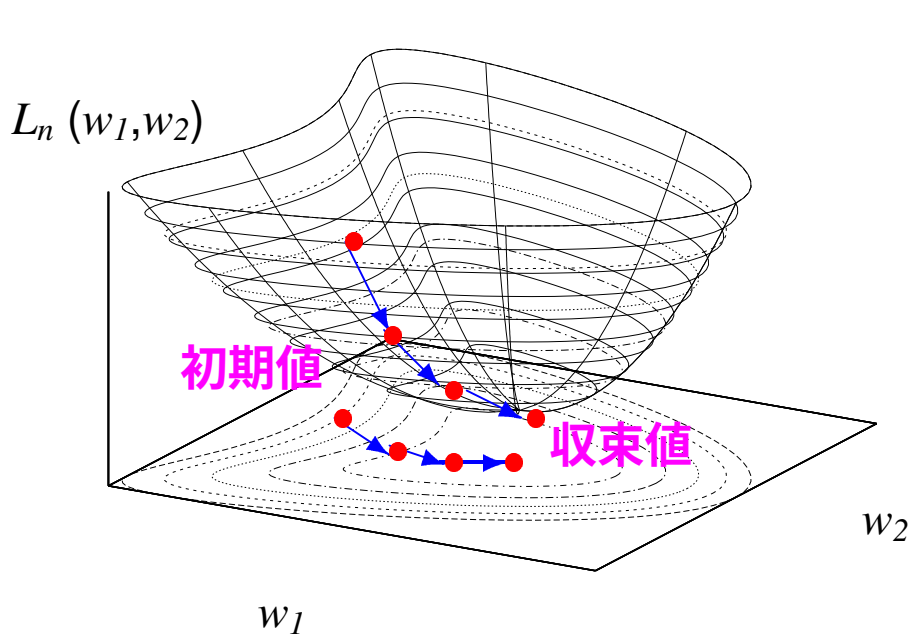
※他の損失関数に対する最適化法の場合には, $L_n(w)$ を着目する損失関数に置き換えればよい.

■最急降下法の仕組みとイメージ

$\eta > 0$ を微小な定数, w, w' を R^d 上のベクトルとしたとき, w のまわりのテーラー展開の 1 次まで

$$L_n(w + \eta w') \approx L_n(w) + \eta w' \cdot \nabla L_n(w).$$

条件 $\|w'\| = 1$ ($\|\eta w'\| =$ 一定) のもとで $L_n(w + \eta w')$ が最小となるのは, w' が $-\nabla L_n(w)$ の方向 (最急降下の方向.「$L_n(w) =$ 一定」の等高線に直交) を向いたとき.



$L_n(w)$ を局所的に最も小さくする方向へ w を逐次的に変化.

◆ [注 12] 最急降下法の問題点

- 学習誤差 $L_n(w)$ の変化 $\neq$ 予測誤差 $L(w)$ の変化

過学習: 学習誤差は減少するが, 予測誤差は増大する

- 予測誤差を最小にするには最急降下法を途中で止める方がよい

最適停止時間: 予測誤差を最小にする解を得るのに要する時間
(事前に評価するアルゴリズムなし)

■自然勾配法 (カルバック擬距離を用いた最適化法)

$$K(w\|w+\eta w')|_{\eta=0} = 0, \qquad \left.\frac{\partial K(w\|w+\eta w')}{\partial w_i}\right|_{\eta=0} = 0,$$

$$\left.\frac{\partial^2 K(w\|w+\eta w')}{\partial w_i \partial w_j}\right|_{\eta=0} = \int \log p(x,y|w) \frac{\partial \log p(x,y|w)}{\partial w_i} \frac{\partial \log p(x,y|w)}{\partial w_j} dxdy$$

$$\stackrel{\text{def}}{=} I_{ij}(w),$$

より $$K(w\|w+\eta w') \cong \frac{\eta^2}{2} w' \cdot I(w) w'.$$

「$K(w\|w+\eta w') =$ **一定」のもとで** $L_n(w+\eta w')$ **を最小にする** w' **は**

$$w' \propto I(w)^{-1} \nabla L_n(w).$$

$I(w)$ **に関してサンプルで近似 (**$p(x,y|w)$ **が真の密度関数に近い)**

$$I_{ij}(w) \cong \frac{1}{n} \sum_{k=1}^{n} \frac{\partial \log p(x_k,y_k|w)}{\partial w_i} \frac{\partial \log p(x_k,y_k|w)}{\partial w_j}.$$

◆ [注 13] 軌跡の変数変換不変性

- **最急降下法の軌跡は変数変換に関して不変でない**

  1 対 1 写像 $w = f(u)$ **による** $u \to w$ **への変数変換**

  w **上の最急降下法** $dw/dt = -\nabla L_n(w)$ **の軌跡**

  $\not\parallel$

  u **上の最急降下法** $du/dt = -\nabla L_n(f(u))$ **の軌跡を**
  $w = f(u)$ **で** w **上に射影したもの**

- **ニュートン法** $dw/dt = -(\nabla\nabla L_n(w))^{-1}\nabla L_n(w)$ **や自然勾配法** $dw/dt = I(w)^{-1}\nabla L_n(w)$ **での軌跡は変数変換に関して不変.**

◆ [注 14] 加速度項のある降下法

$w(t)$ **の時間発展** ($\eta > 0,\ \alpha > 0$)

$$w(t+1) = w(t) - \eta \nabla L_n(w) + \underbrace{\alpha(w(t) - w(t-1))}_{\textbf{加速度項}}.$$

利点:

- **損失関数の局所解 (凹凸) を乗越える.**
- $w(t)$ **の離散化によって生じる振動を抑える.**

■最急降下法 (逐次学習)

$w(t)$ の時間発展 (各 t ごとに $i(1 \leq i \leq n)$ をランダムに選出)

$$w(t+1) = w(t) - \eta(t)\nabla_w \mathcal{E}(x_i, y_i, w(t)) + \alpha(w(t) - w(t-1)).$$

$$\sum_{t'=1}^{t} \eta(t') \to \infty, \quad \sum_{t'=1}^{t} \eta(t')^2 \to 0, \quad \text{as } t \to \infty.$$

利点:

- $L_n(w(t)) = \sum_i \mathcal{E}(x_i, y_i, w(t))/n$ を $\mathcal{E}(x_i, y_i, w(t))$ で置き換えるため, 計算量少.

欠点:

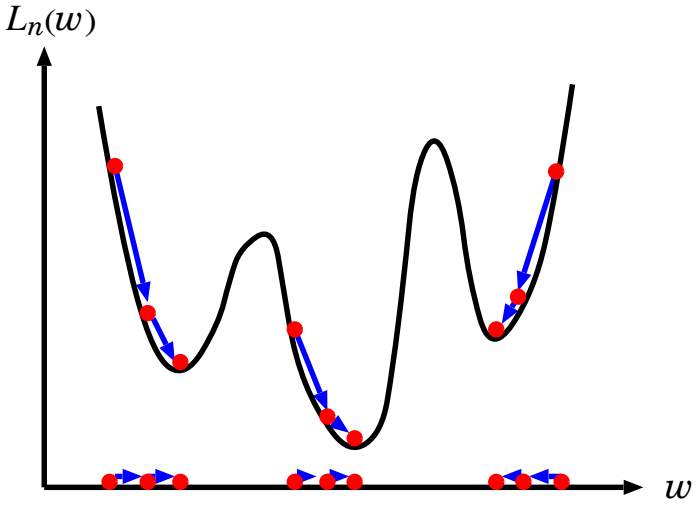
- 学習結果がサンプル i の順番に影響される.
- 解を収束させるのに $\eta(t)$ の制御が必要.

◆ [注 15]+補足: 最急降下法での注意事項

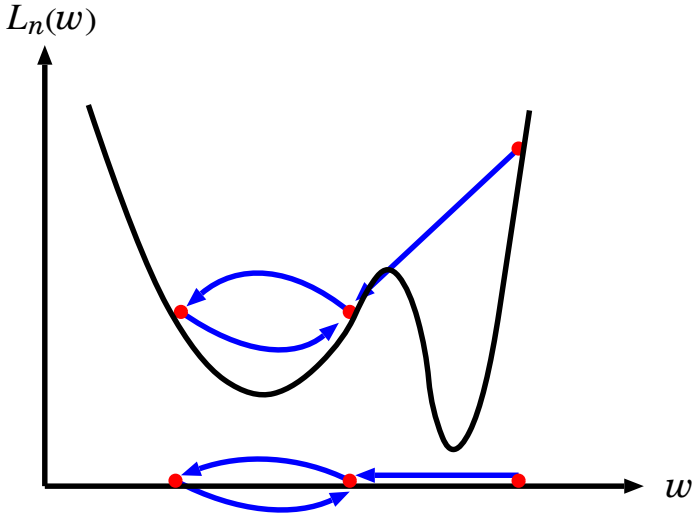
局所解近傍での損失関数のランドスケープ [単峰性/多峰性]
(損失関数の局所解近傍でのテーラー展開の 2 次以降の影響)

↓

解の初期値依存性, ジャンプ幅 η 依存性, 収束判定依存性



局所解への収束



最適解の飛越しと
局所解近傍での振動

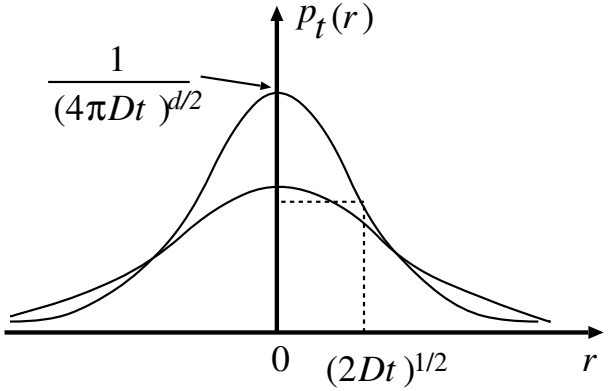
2.2.2 確率項をもつ最急降下法

■ブラウン運動

確率変数の集合 $\{R_t; t > 0, R_t \in R^d\}$ が以下の 2 条件を満たすとき, R_t を R^d 上のブラウン運動という.

(1) R_t は次の密度関数 (正規分布) をもつ (拡散係数 $D > 0$).

$$p_t(r) = \frac{1}{(4\pi Dt)^{d/2}} \exp\left(-\frac{\|r\|^2}{4Dt}\right).$$



(2) 任意の k と任意の $0 < t_1 < t_2 < \cdots < t_k$ に対して
$\{R_{t_{j+1}} - R_{t_j}; j = 1, 2, \ldots, k-1\}$ は互いに独立.

※密度関数 $p_t(r)$ は熱伝導方程式 (拡散方程式) の解

$$\frac{\partial p_t(r)}{\partial t} = D\Delta_r p_t(r).$$

■確率項をもつ方程式

$R^d \to R^1$ への関数 $V(w)$, 確率変数の集合 $\{W_k; W_k \in R^d, k = 1, 2, \ldots\}$, 単位時間 τ, ガウシアンノイズ R_τ

関数 $V(w)$ に関する最急降下法 (ノイズ含む)

$$W_{k+1} = W_k - \tau \nabla V(W_k) + R_\tau, \quad (1)$$
$$W_0 = 0. \quad (2)$$

ランジュバン方程式:
確率変数 W_k に対する微分方程式 (式 (1) で $k\tau = t$, $\tau \to 0$)

$$\frac{dW_t}{dt} = -\nabla V(W_t) + \frac{dR_t}{dt}.$$

フォッカー-プランク方程式:
確率変数 W_t の時刻 t での密度関数 $q(w,t)$ に対する偏微分方程式

$$\frac{\partial q(w,t)}{\partial t} - \nabla \cdot (\nabla V(w) q(w,t)) = D \Delta q(w,t).$$

◆ [注 18]: ノイズを入れた確率的な学習方程式

シミュレーテッドアニーリング:
ノイズ (確率項) を含む最急降下法．ノイズの大きさ (拡散係数 D) を徐々に 0 にしていく．

局所解に陥ることなく, 確率 1 で最的解を見つけることができる．
(ノイズにより局所解を抜け出す)