忙しい人のための情報幾何入門

平成 30 年 8 月 4 日

1 微分幾何学

S を多様体とし, 関数, 曲線はすべて C^∞ 級とする. S 上の関数全体の集合を $C^\infty(S)$ とする.
$X : C^\infty(S) \to C^\infty(S)$ が**微分作用素**であるとは,

線形性 $X(a\varphi + b\psi) = aX(\varphi) + bX(\psi)$

Leipniz 則 $X(\varphi\psi) = \varphi X(\psi) + X(\varphi)\psi$

をみたすことをいう. $C^\infty(S)$ 上の微分作用素を S 上の**ベクトル場**という. S 上のベクトル場全体を $\mathcal{X}(S)$ とする.
各 $p \in S$ で, **接ベクトル** $X_p : C^\infty(S) \to \mathbb{R}$ を

$$X_p(\varphi) = (X\varphi)(p)$$

で定義する. p における接ベクトル全体の集合を T_pS と書き, p における S の**接空間**という. T_pS は線形空間である.
$g : \mathcal{X}(S) \times \mathcal{X}(S) \to C^\infty(S)$ が **Riemann 計量**とは, 各 $p \in S$ で,

$$g_p(X_p, Y_p) = g(X, Y)(p)$$

によって定まる $g_p : T_pS \times T_pS \to \mathbb{R}$ が T_pS 上の内積となっていることをいう.
Riemann 計量を定めると, S 上の曲線に対して, **直交**が定義される. S 上の曲線 $\alpha : (-\varepsilon, \varepsilon) \to S$ に対し, $t \in (-\varepsilon, \varepsilon)$ における接ベクトル $\dot{\alpha}_t \in T_{\alpha_t}S$ を

$$\dot{\alpha}_t(\varphi) = \frac{\mathrm{d}(\varphi \circ \alpha)}{\mathrm{d}t}(t)$$

で定める. $\alpha_0 = \beta_0 = p$ なる 2 曲線 α と β が p で g に関して直交するとは,

$$g_p(\dot{\alpha}_0, \dot{\beta}_0) = 0$$

となっていることをいう.
$\nabla : \mathcal{X}(S) \times \mathcal{X}(S) \to \mathcal{X}(S)$ が**アフィン接続**であるとは,

$$\begin{aligned}
&\nabla_X(Y + Z) = \nabla_X Y + \nabla_X Z \\
&\nabla_{X+Y} Z = \nabla_X Z + \nabla_Y Z \\
&\nabla_X(\varphi Y) = (X\varphi)Y + \varphi\nabla_X Y \\
&\nabla_{\varphi X} Y = \varphi\nabla_X Y
\end{aligned}$$

をみたすことをいう. アフィン接続を定めると, S 上のベクトル場に対して, **平行**が定義される. 曲線 α が与えられたとき, α 上の各 α_t に $\dot{\alpha}_t$ を対応させることで, 曲線に沿ったベクトル場 $\dot{\alpha}$ が定義できる. ベクトル場 $X \in \mathcal{X}(S)$ がアフィン接続 ∇ に関して, 曲線 α に沿って平行とは,

$$\nabla_{\dot{\alpha}} X = 0$$

となっていることをいう. また,

$$\nabla_{\dot{\alpha}} \dot{\alpha} = 0$$

をみたす曲線 α を ∇ に関する**測地線**という.
∇ の曲率 $R : \mathcal{X}(S) \times \mathcal{X}(S) \times \mathcal{X}(S) \to \mathcal{X}(S)$ と捩率 $T : \mathcal{X}(S) \times \mathcal{X}(S) \to \mathcal{X}(S)$ を

$$\begin{aligned} R(X, Y, Z) &= \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z \\ T(X, Y) &= \nabla_X Y - \nabla_Y X - [X, Y] \end{aligned}$$

で定義し, 曲率と捩率がともに恒等的に 0 のとき, S は ∇ に関して**平坦**であるという. S の部分多様体 M に対して, M が ∇ に関して**自己平行**であるとは, 任意のベクトル場 $X, Y \in \mathcal{X}(M)$ に対して, $\nabla_X Y \in \mathcal{X}(M)$ となることをいう.
∇^0 が Riemann 多様体 (S, g) の Levi-Civita 接続であるとは,

$$\begin{aligned} \nabla^0_X Y - \nabla^0_Y X - [X, Y] &= 0 \\ X g(Y, Z) &= g(\nabla^0_X Y, Z) + g(Y, \nabla^0_X Z) \end{aligned}$$

をみたすことをいう. Riemann 幾何学でアフィン接続といえば, 通常は Levi-Civita 接続のことを指す.

2 情報幾何学

ここでは, 確率分布は有限集合上の分布のみを扱う. $\mathcal{X}$ を有限集合として, $\mathcal{X}$ 上の確率分布全体の集合を,

$$S := \left\{ p : \mathcal{X} \to (0, 1) \;\middle|\; \sum_{x \in \mathcal{X}} p(x) = 1 \right\}$$

とする. S はユークリッド空間の開部分集合なので, 多様体である.
E_p は p に関する期待値を表すとする.

$$g_p(X_p, Y_p) := E_p[(X \log p)(Y \log p)]$$

で Riemann 計量を定める. これを **Fisher 計量**という.
∇^0 を Levi-Civita 接続として,

$$\begin{aligned} g_p(\nabla^e_X Y, Z) &= g_p(\nabla^0_X Y, Z) - \frac{1}{2} E_p[(X \log p)(Y \log p)(Z \log p)] \\ g_p(\nabla^m_X Y, Z) &= g_p(\nabla^0_X Y, Z) + \frac{1}{2} E_p[(X \log p)(Y \log p)(Z \log p)] \end{aligned}$$

によって定まるアフィン接続 ∇^e と ∇^e をそれぞれ, **指数型接続**と**混合型接続**という.
$K < \sharp\mathcal{X}$ とする. 関数 $C : \mathcal{X} \to \mathbb{R}$ と $F : \mathcal{X} \to \mathbb{R}^K$ が与えられたとき, パラメータ $\theta \in \Theta \subset \mathbb{R}^K$ を用いて,

$$p_\theta(x) = \exp(C(x) + \theta^T F(x) - \psi(\theta))$$

で表される分布族 $\{p_\theta \mid \theta \in \Theta\}$ を, **指数型分布族**という. ここで,

$$\psi(\theta) = \log \left(\sum_{x \in \mathcal{X}} \exp(C(x) + \theta^T F(x)) \right)$$

S の部分多様体 M が ∇^e-自己平行であることと, M が指数型分布族であることは同値である.
$q_0, q_1, \ldots, q_K \in S$ が与えられたとき, パラメータ $\eta \in \mathcal{H} \subset \mathbb{R}^K$ を用いて,

$$p_\eta(x) = \left(1 - \sum_{k=1}^{K} \eta_k \right) q_0(x) + \sum_{k=1}^{K} \eta_k q_k(x)$$

で表される分布族 $\{p_\eta \mid \eta \in \mathcal{H}\}$ を, **混合型分布族**という. S の部分多様体 M が ∇^m-自己平行であることと, M が混合型分布族であることは同値である.
$p, q \in S$ に対して,

$$\mathrm{KL}[p \| q] := \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$$

を p から q への **Kullback-Leibler ダイバージェンス**という.
Kullback-Leibler ダイバージェンスは対称ではないが, 三角不当式

$$\mathrm{KL}[p\|q] + \mathrm{KL}[q\|r] \geq \mathrm{KL}[p\|r]$$

が成り立ち, p と q を結ぶ ∇^m-測地線と q と r を結ぶ ∇^e-測地線が q で直交しているときに限り, Pythagoras の定理

$$\mathrm{KL}[p\|q] + \mathrm{KL}[q\|r] = \mathrm{KL}[p\|r]$$

が成り立つ.
確率モデル $M = \{p_\theta \mid \theta \in \Theta\}$ と, 未知の分布に従う独立な確率変数列の実現値 $x_1, \dots, x_N \in \mathcal{X}$ が与えられたとき,

$$\hat{\theta}_N := \arg\max_\theta \prod_{n=1}^N p_\theta(x_n)$$

を最尤推定量という.
経験分布 $\hat{p}_N$ を,

$$\hat{p}_N(x) = \frac{1}{N}\sum_{n=1}^N \delta(x - x_n)$$

で定義すると,

$$\hat{\theta}_N = \arg\min_\theta \mathrm{KL}[\hat{p}_N\|p_\theta]$$

が成り立つ.
$\hat{p}_N$ を通る ∇^m-測地線が p_{θ^*} で M と直交しているとする. このとき, 任意の θ に対し, Pythagoras の定理

$$\mathrm{KL}[\hat{p}_N\|p_{\theta^*}] + \mathrm{KL}[p_{\theta^*}\|p_\theta] = \mathrm{KL}[\hat{p}_N\|p_\theta]$$

が成り立つので,

$$\mathrm{KL}[\hat{p}_N\|p_{\theta^*}] \leq \mathrm{KL}[\hat{p}_N\|p_\theta]$$

よって,

$$\theta^* = \arg\min_\theta \mathrm{KL}[\hat{p}_N\|p_\theta] = \hat{\theta}_N$$

であるから, 最尤推定は経験分布から確率モデルへの ∇^m-直交射影に他ならない. 特に, 確率モデルが指数型ならば, 最尤推定は一意的である.

3 隠れ変数を含むモデル

$\mathcal{X} = \mathcal{Y} \times \mathcal{Z}$ として, $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ 上の確率分布全体をそれぞれ $S^{\mathcal{X}}, S^{\mathcal{Y}}, S^{\mathcal{Z}}$ とする. 指数型分布族 $M = \{p_\theta \in S^{\mathcal{X}} \mid \theta \in \Theta\}$ と, 未知の分布に従う独立な確率変数列 $\{x_n = (y_n, z_n)\}_{n=1}^N$ の一部 $\{y_n\}_{n=1}^N$ が与えられたとする. $\{z_n\}_{n=1}^N$ を**隠れ変数**という. 経験分布を,

$$\hat{q}_N(x) = \frac{1}{N}\sum_{n=1}^N \delta(y - y_n)$$

として,

$$D := \left\{ p \in S^{\mathcal{X}} \mid \sum_{z \in \mathcal{Z}} p(y, z) = \hat{q}_N(y) \right\}$$

とする. D は ∇^m-自己平行で, D の各点は $\hat{q}_N(y)r(z|y)$ と表さる.

$$\theta_{t+1} = \arg\min_\theta \mathrm{KL}[\hat{q}_N(y)r_{\theta_t}(z|y)\|p_\theta(y, z)]$$

に従ってパラメータを更新するアルゴリズムを **EM アルゴリズム**という.

$$\hat{q}_N(y)r_{\theta_t}(z|y) = \arg\min_{p \in D} \mathrm{KL}[p\|p_{\theta_t}]$$

より, EM アルゴリズムは M から D への ∇^e-射影と D から M への ∇^m-射影を交互に繰り返していることがわかる.

$$D = \{\hat{q}_N(y) r_\eta(z|y) \mid \eta \in \mathcal{H}\}$$

として,

$$\eta_{t+1} = \arg\min_{\eta} \mathrm{KL}[\hat{q}_N(y) r_\eta(z|y) \| p_{\theta_t}(y,z)]$$

$$\theta_{t+1} = \arg\min_{\theta} \mathrm{KL}[\hat{q}_N(y) r_{\eta_{t+1}}(z|y) \| p_{\theta}(y,z)]$$

でパラメータを更新するアルゴリズムを**一般化 EM アルゴリズム**という. 各 η に対して, $r_\eta(z|y)$ が条件付き独立となるようにパラメータを決める事が多い.
複雑な分布を独立分布で近似することを**平均場近似**という. 平均場近似は ∇^e-自己平行部分多様体への ∇^e-射影なので, 一般に一意に定まらない.