正則モデルにおける能動学習

渡辺 一帆
東京工業大学大学院 総合理工学研究科 電子機能システム専攻

平成 14 年 10 月 28 日

1 能動学習

システムの入力 x と出力 y の間の関数関係を統計的学習データから推定する問題 (回帰問題) を考える。このとき、学習データを採取する点を選択することで、より高精度の推定を実現する学習方式を能動学習という。学習データの選択には様々な基準が考えられ、研究がなされている。以下では、まず、最尤推定を用いるときに学習データを発生させる分布を最適化する能動学習について、その後、ベイズ推定における逐次的な能動学習について見ていく。

2 最尤推定における能動学習

2.1 回帰モデル

回帰モデルは、入力 $x \in R$ から出力 $y \in R$ への確率的関係を、

$$y = f(x, \theta) + (\text{雑音})$$

($\theta \in R^M$ はパラメータ) と仮定し、学習データ $\{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$ から入出力間の関数を推定する手法である。雑音が平均 0、分散 σ^2 の正規分布に従うとすると、入力 x が与えられたときの出力 y の条件つき確率密度関数は、

$$p(y|x, \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{1}{2\sigma^2}(y - f(x, \theta)^2)\}$$

である。以下では、真の関数が $f(x, \theta_0)$ で表されるとし、パラメータ θ を最尤推定法により推定する場合を考える。このとき、 2 乗誤差の和

$$\sum_{i=1}^{N}(y_i - f(x_i, \theta))^2$$

を最小にする θ が最尤推定量であり、$\hat{\theta}$ で表す。

$f(x, \hat{\theta})$ を用いて, テスト入力 x に対する予測を行うとき、テスト入力 x がある分布 $q(x)$ に従って発生すると仮定し、予測誤差として、平均 2 乗誤差

$$\int\int (y - f(x, \hat{\theta}))^2 q(x) p(y|x, \theta_0) dxdy$$

を考える。この予測誤差がより小さくなるように学習データの入力点 $\{x_i\}_{i=1}^N$ を選ぶことで能動学習が行える。これは、$\{x_i\}_{i=1}^N$ を分布 $r(x)$ から発生させるとすると、予測誤差の学習データの出方に関する期待値

$$E \equiv E[\int\int (y - f(x, \hat{\theta}))^2 q(x) p(y|x, \theta_0) dxdy]$$

を最小にする分布 $r(x)$ を求めることで行われる。

2.2 最適設計の基準

予測誤差の期待値 E は以下のように近似できる。

$$E = \sigma^2 + \frac{\sigma^2}{N}\mathrm{Tr}[IJ^{-1}(r)] + o(N^{-1}) \tag{1}$$

ここで、I,$J(r)$ は $M \times M$ 行列で、

$$I_{ij} = \int \frac{\partial f(x,\theta_0)}{\partial \theta_i}\frac{\partial f(x,\theta_0)}{\partial \theta_j} q(x)dx$$

$$J_{ij}(r) = \int \frac{\partial f(x,\theta_0)}{\partial \theta_i}\frac{\partial f(x,\theta_0)}{\partial \theta_j} r(x)dx$$

である。

(式 (1) の証明)

$$(y - f(x,\hat{\theta}))^2 = (y - f(x,\theta_0))^2 + (f(x,\hat{\theta}) - f(x,\theta_0))^2 + 2(f(x,\hat{\theta}) - f(x,\theta_0))(y - f(x,\theta_0))$$

より、

$$\int\int (y - f(x,\hat{\theta}))^2 q(x)p(y|x,\theta_0)dxdy = \sigma^2 + \int (f(x,\hat{\theta}) - f(x,\theta_0))^2 q(x)dx$$

$f(x,\hat{\theta})$ を θ_0 のまわりで Taylor 展開すると、

$$f(x,\hat{\theta}) - f(x,\theta_0) = \sum_{i,j} I_{ij} E[(\hat{\theta_i} - \theta_{i,0})(\hat{\theta_j} - \theta_{j,0})] + E[O(\|\hat{\theta} - \theta_0\|^3)]$$

となる。統計的正則モデルの漸近理論を用いると、分布 $r(x)$ を用いて学習したとき、$\sqrt{N}(\hat{\theta} - \theta_0)$ の分布は $N \to \infty$ のとき、平均 0、分散共分散行列 $\sigma^2 J^{-1}$ の正規分布に漸近し、

$$E[(\hat{\theta_i} - \theta_{i,0})(\hat{\theta_j} - \theta_{j,0})] = \frac{\sigma^2 J^{ij}}{N} + o(N^{-1}) \quad (J^{ij} \text{ は } J^{-1} \text{ の成分}) \tag{2}$$

$$E[O(\|\hat{\theta} - \theta_0\|^3)] = O(N^{-\frac{3}{2}}) \tag{3}$$

が成立する。以上より式 (1) が得られる。**(式 (1) の証明終)**

学習データ数 N が大きいとき、式 (1) において学習データの設計によって変化するのは、$\mathrm{Tr}[IJ^{-1}(r)]$ の部分だけなので、これを最小化する r を見つければよい。一般にはこれは真のパラメータ θ_0 を含んだ式なので、最適な r を学習前に構成することはできない。しかし、線形モデルの場合には θ_0 に依存しないために事前に最適設計を求めることができる。$r = q$ のときは $\mathrm{Tr}[IJ^{-1}(r)] = M$ であり、パラメータ数に一致する。r を最適化し、能動的な学習を行うと予測誤差の値を小さくすることができる。実際多項式近似モデル

$$f(x,\theta) = \sum_{j=0}^{M-1} \theta_j x^j \qquad x \in (-\infty, \infty)$$

においては $\mathrm{Tr}[IJ^{-1}(r)]$ の下限が 1 であり、能動学習により予測誤差がパラメータ数に依存しなくなることが示されている。

3 Bayes 推定における能動学習

3.1 Bayes 決定理論による能動学習

上と同様の回帰モデルを考え、学習データ全体 $\{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$ を D で表す。パラメータの事前分布が $p(\theta)$ とすると、データを観測した後のパラメータ θ の事後分布は

$$p(\theta|D) = \frac{1}{Z}\prod_{i=1}^{N} p(y_i|x_i, \theta)p(\theta)$$

である (Z は正規化定数)。この事後分布を用いて、新しい入力 x に対する出力 y の予測分布を

$$p(y|x, D) = \int p(y|x, \theta)p(\theta|D)d\theta$$

のように計算する。ここで、予測の期待損失を最小にするようなデータの選択基準を考える。この基準は Bayes 決定理論の枠組みで考えることができる。個々の予測の良さは予測の目的に依存し、入力 x に基づいて、何らかの行動 $a \in A$ を行い、その結果が出力 y の値に依存するという状況を考える。出力が y であるときに行動 a を行った場合の損失が損失関数 $L(y, a)$ の値で計れるとする。Bayes 決定理論では、学習データ D と入力 x に基づいて、損失の期待値 (期待損失) を最小化する行動

$$\hat{a}(x, D) = \arg\min_{a \in A} \int L(y, a)p(y|x, D)dy$$

を選択する ($\hat{a}(x, D)$ を決定関数と呼ぶ)。

入力 x に対する出力 y の値の予測を目的とする場合には、予測値 $\hat{y}(x)$ そのものが行動となり、損失関数は $L(y, \hat{y}) = (y - \hat{y})^2$ を用いることが考えられる。このとき、最適な決定関数は y の条件つき期待値 $E(y|x, D)$ となり、条件つき分散が対応する期待損失となる。

能動学習における新しい観測点 $\tilde{x}$ の選択も Bayes 決定理論の 1 つと考えることができる。新しいデータ $(\tilde{x}, \tilde{y})$ が与えられたとき、それを加えたデータ $\tilde{D}$ に基づいて予測を行った場合の損失の期待値 $L(\tilde{x})$ を損失関数として用いることができる。このためには、次の 3 つの量を考慮する必要がある。

1. $\tilde{x}$ に対する出力 $\tilde{y}$ の確率分布
   (真の分布は未知なので現在までのデータに基づく予測分布 $p(y|x, D)$ を用いる)
2. 学習結果の評価に用いられるテスト入力の分布 $q(x)$
3. $\tilde{D}$ に基づく最適行動関数 $\tilde{a}(x, \tilde{D})$(ここでは y の条件つき期待値 $E(y|x, \tilde{D})$)

Bayes 決定理論の原理に従って，$\tilde{x}$ を以下の $L(\tilde{x})$ を最小にするように選ぶ。

$$L(\tilde{x}) = \int L(\tilde{x}, \tilde{y}, x)q(x)p(\tilde{y}|\tilde{x}, D)dxd\tilde{y}$$

ここで、

$$L(\tilde{x}, \tilde{y}, x) = \int L(y, \hat{a}(x, \tilde{D}))p(y|x, \tilde{D})dy \tag{4}$$

3.2 漸近解析

2.1 と同様の回帰モデルを考え，予測の損失は 2 乗誤差の場合を考える。(4) の損失関数は

$$L(\tilde{x}, \tilde{y}, x) = \sigma^2 + \int (f(x, \theta) - E(y|x, \tilde{D}))^2 p(\theta|\tilde{D})d\theta$$

である。第 2 項は最適な予測値の分散 (予測分散) である。これは、損失関数が 2 乗誤差のときには予測分散が、次の観測点の選択に関して重要であることを示している。

以下では学習データ数が大きいときの予測分散について考察する。適切な正則条件の下で、事後分布はガウス関数に収束する。事後分布 $p(\theta|D)$ をその極大値 $\hat{\theta}$ のまわりで以下のように展開する。

$$\log p(\theta|D) \simeq \log p(\hat{\theta}|D) - \frac{1}{2}(\theta-\hat{\theta})^t \Sigma_n^{-1}(\theta-\hat{\theta})$$

(ここで $\Sigma_n^{-1} = -\frac{\partial^2 \log p(\hat{\theta}|D)}{\partial\theta\partial\theta^t}$ は経験フィッシャー情報行列と呼ばれる。)θ の事前分布として正規分布 $p(\theta) = N(\theta_a, C_0^{-1})$ をとると、$F'(\hat{\theta})$ を $f'(x_i,\theta) = \frac{\partial f(x_i,\hat{\theta})}{\partial\theta}$ を列とする行列とし、$M \times M$ 行列 V_n を

$$V_n = (C_0 + \beta F'(\hat{\theta})F'(\hat{\theta})^t)^{-1} \tag{5}$$

$(\beta = 1/\sigma^2)$ とするとき、θ の事後分布が $\hat{\theta}$ の近傍で正規分布 $N(\hat{\theta}, V_n)$ で近似できることが示せる。この漸近正規性は、ゆるやかな正則条件のもとで、多くの確率分布モデルについて成り立つ。

$f(x,\theta)$ を $\hat{\theta}$ のまわりで展開すると、

$$f(x,\theta) \simeq f(x,\hat{\theta}) + f'(x,\hat{\theta})^t(\theta-\hat{\theta})$$

であり、これを用いると予測分散を

$$\mathrm{var}(f(x,\theta)) \simeq f'(x,\hat{\theta})^t V_n f'(x,\hat{\theta}) \tag{6}$$

と近似することができる。

新しい観測点 $\tilde{x}$ の効果を評価する。(5) 式より

$$V_{n+1}^{-1} \simeq V_n^{-1} + \beta f'(\tilde{x},\hat{\theta})f'(\tilde{x},\hat{\theta})^t \tag{7}$$

を得る[1] 。さらに逆行列の補題を使って

$$V_{n+1} \simeq V_n - \frac{V_n f'(\tilde{x},\hat{\theta})f'(\tilde{x},\hat{\theta})^t V_n \beta}{1 + f'(\tilde{x},\hat{\theta})^t V_n f'(\tilde{x},\hat{\theta})\beta}$$

これと、(6) から、$(\tilde{x},\tilde{y})$ を観測した後の損失関数の減少量 $\Delta_L(x,\tilde{x})$ は、

$$f'(x,\hat{\theta})^t \frac{V_n f'(\tilde{x},\hat{\theta})f'(\tilde{x},\hat{\theta})^t V_n \beta}{1 + f'(\tilde{x},\hat{\theta})^t V_n f'(\tilde{x},\hat{\theta})\beta} f'(x,\hat{\theta}) \tag{8}$$

となる。この漸近展開に基づく能動学習では、上式の損失の減少の期待値をテスト入力の分布 $q(x)$ で平均する。$f'(x,\hat{\theta})$ の平均を $E(f')$、分散共分散行列を $V(f')$ とする。式 (8) の分数部分を $U_n(\tilde{x})$ と書くと、損失減少分の平均 $\Delta_L(\tilde{x}) = E_q[\Delta_L(x,\tilde{x})]$ は、

$$\mathrm{Tr}[U_n(\tilde{x})V(f')] + E(f')^t U_n(\tilde{x}) E(f')$$

で近似できる。最適な観測点は $\Delta_L(\tilde{x})$ を最大化することによって得られる。

参考文献

[1] 福水健次, 渡辺澄夫 “多項式近似における学習データの最適設計と予測誤差” 信学論 A, Vol.J79-A, No.5, pp.1100-1108, 1996.

[2] Gerhard PAASS(訳：麻生英樹) “予測とモデル選択のための質問選択” 情報処理, Vol.38, No.7, pp.562-568, 1997.

[1] (7) 式の行列式をとると、

$$|V_{n+1}^{-1}| = |V_n^{-1}|(1 + \beta f'(\tilde{x},\hat{\theta})^t V_n f'(\tilde{x},\hat{\theta}))$$

したがって、事後分布の分散共分散行列の行列式 $|V_{n+1}| = 1/|V_{n+1}^{-1}|$ を最小化するためには、現在の予測分散 $f'(\tilde{x},\hat{\theta})^t V_n f'(\tilde{x},\hat{\theta})$ を最大にする $\tilde{x}$ を選べばよい。この基準は D-optimality とも呼ばれる。