電子情報通信学会知識の森, no.S3-4-1-5, pp.21-23, 2012.

回帰（関数近似）

東京工業大学　杉山将

教師付き学習 (supervised learning) とは，入力 x と出力 y の組からなる n 個の訓練データ $\{x_i, y_i\}_{i=1}^n$ を用いて，その背後に潜んでいる入出力関係を学習する問題である [1, 2]．入出力関係をうまく学習することができれば，学習していない入力 x に対する出力 y を予測できるようになる．すなわち，未知の状況に適応する汎化能力 (generalization ability) が獲得できる．与えられた訓練データからできるだけ高い汎化能力を獲得することが教師付き学習の目標である．ここでは，訓練データ $\{x_i, y_i\}_{i=1}^n$ が同時確率密度 $p(x, y)$ に独立同一分布 (independent and identically distributed; i.i.d.) に従うと仮定し，出力 y の条件付き期待値 $\mathrm{E}_{p(y|x)}[y]$ を推定する問題を考える．出力 y が実数値を取るとき回帰 (regression) 問題と呼び，y がカテゴリ値を取るとき分類 (classification) 問題と呼ぶ．

回帰問題における最も基礎的な学習法は，線形モデル (linear model) を用いた最小二乗法 (least-squares) であろう．線形モデルは，基底関数 $\{\varphi_j(x)\}_{j=1}^t$ の線形和によって関数を近似するモデルである．

$$f_{\mathrm{linear}}(x) = \sum_{j=1}^{t} \theta_j \varphi_j(x)$$

最小二乗法は，二乗誤差基準のもとでパラメータ $\{\theta_j\}_{j=1}^t$ を訓練データに適合させる方法である．

$$\min_{\{\theta_j\}_{j=1}^t} \sum_{i=1}^{n} \left(y_i - f_{\mathrm{linear}}(x_i)\right)^2$$

これは，以下のガウスモデルの最尤推定法 (maximum likelihood estimation) に対応している．

$$q_{\mathrm{Gauss}}(y|x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - f_{\mathrm{linear}}(x))^2}{2\sigma^2}\right)$$

線形モデルに対する最小二乗法の最適化問題は凸であり，大域的最適解が解析的に求まるという長所がある．しかし，あらかじめ基底関数を固定しておく必要があるため，線形モデルは柔軟性に欠ける．

基底関数にもパラメータを含めることによって，より柔軟な関数近似を行うことができる．そのような非線形モデルの代表例は，動径基底関数 (radial basis function; RBF) モデルである．RBF モデルの最小二乗解は次式で求められる．

$$\min_{\{\theta_j, \mu_j, \Sigma_j\}_{j=1}^t} \sum_{i=1}^{n} \left(y_i - f_{\mathrm{RBF}}(x_i)\right)^2$$

$$f_{\mathrm{RBF}}(x) = \sum_{j=1}^{t} \theta_j \exp\left(-\frac{1}{2}(x - \mu_j)^\top \Sigma_j^{-1} (x - \mu_j)\right)$$

このように RBF モデルでは，線形結合係数 $\{\theta_j\}_{j=1}^t$ だけでなくガウス基底関数の中心 $\{\mu_j\}_{j=1}^t$ と共分散行列 $\{\Sigma_j\}_{j=1}^t$ も訓練データを使って適応的に決める．従って，非常に柔軟なモデリングが可能である．しかし，最適化問題が非凸であるため大域的最適解を求めることは困難であり，勾配降下法などにより局所的最適解を求めるのが一般的である．

線形モデルと RBF モデルの中間に位置づけられるのが，カーネルモデル (kernel model) である．ガウスカーネルモデルの最小二乗解は次式で求められる．

$$\min_{\{\theta_j\}_{j=1}^n} \sum_{i=1}^n (y_i - f_{\text{kernel}}(x_i))^2$$
$$f_{\text{kernel}}(x) = \sum_{j=1}^n \theta_j \exp\left(-\frac{(x-x_j)^\top (x-x_j)}{2\sigma^2}\right)$$

これは凸最適化問題であり，線形モデルのときと同様にして大域的最適解を解析的に求めることができる．更に，パラメータ数が訓練データ数と共に増加することから，線形モデルよりも柔軟に関数近似を行うことができる．

分類問題でも最小二乗法を利用することはできるが，ロジスティック回帰 (logistic regression) を用いれば確率的な出力が得られ便利である．ここでは，出力が $y = \pm 1$ の二値分類問題を考える．ロジスティック回帰では，出力 y の条件付き確率 $p(y|x)$ を次のようにモデル化する．

$$q_{\text{logistic}}(y|x) = \frac{1}{1 + \exp\left(-y f_{\text{linear}}(x)\right)}$$

線形モデル $f_{\text{linear}}(x)$ のパラメータ $\{\theta_j\}_{j=1}^t$ の最尤推定量は次式で求められる．

$$\min_{\{\theta_j\}_{j=1}^t} \sum_{i=1}^n \log\left(1 + \exp\left(-y_i f_{\text{linear}}(x_i)\right)\right)$$

これは凸最適化問題であり，勾配降下法や準ニュートン法によって大域的最適解を求めることができる．カーネルモデル $f_{\text{kernel}}(x)$ に対するロジスティック回帰も同様に定義することができ，やはり凸最適化問題として定式化される．

最尤推定法の近似性能は，基底関数の選び方に依存する．学習結果の汎化性能は，カルバック・ライブラー (KL) 情報量を使って測るのが一般的である．真の分布 $p(x,y)$ から，学習結果 $q(y|x)p(x)$ への KL 情報量は次式で与えられる．

$$\text{KL}[p(x,y)||q(y|x)p(x)] = \int p(x,y) \log \frac{p(x,y)}{q(y|x)p(x)} dxdy$$

KL 情報量がゼロになることと学習結果 $q(y|x)$ が真の条件付き分布 $p(y|x)$ と一致することは等価である．情報量規準 (information criterion) とは KL 情報量の推定量の事を指し，例えば赤池情報量規準 (Akaike information criterion; AIC)[3] は次式で定義される．

$$\text{AIC} = -2 \sum_{i=1}^n \log q(y_i|x_i) + 2t$$

AICは，適当な条件の下でKL情報量のよい推定量になっている．従って，AICを最小にするようにモデルを決定すれば，高い汎化能力が得られると期待される．しかし，カーネルモデルのようにパラメータ数が訓練データ数と共に増加するモデルや，RBFモデルのように特異性を持つモデルに対してはAICの近似精度は良くないことが知られている [4].

参考文献

[1] 元田 浩, 栗田 多喜夫, 樋口 知之, 松本 裕治, 村田 昇（編）. **パターン認識と機械学習（上）: ベイズ理論による統計的予測**, シュプリンガー・ジャパン, 東京, 2007.

[2] 元田 浩, 栗田 多喜夫, 樋口 知之, 松本 裕治, 村田 昇（編）. **パターン認識と機械学習（下）: ベイズ理論による統計的予測**, シュプリンガー・ジャパン, 東京, 2008.

[3] H. Akaike. A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, AC-19, 6, 716–723, 1974.

[4] S. Watanabe, Algebraic analysis for nonidentifiable learning machines, *Neural Computation*, 13, 4, 899–933, 2001.