電子情報通信学会知識の森, no.S3-3-?-?, pp.???-???, 201?.

統計的学習

東京大学　杉山将

統計的学習 (statistical learning) とは，データの統計的な性質を利用する機械学習の枠組みである．ここでは特に，入力 x と出力 y の組からなる n 個の訓練データ $D = \{x_i, y_i\}_{i=1}^n$ を用いて，その背後に潜んでいる入出力関係を学習する教師付き学習 (supervised learning) とよばれる問題を考える．通常，訓練データは同時確率分布 $P(x, y)$ に独立同一分布 (independent and identically distributed; i.i.d.) に従って生成されていると仮定する．教師付き学習の他には，出力のない入力だけが訓練データとして与えられる場合の学習問題を議論する教師なし学習 (unsupervised learning)，教師付きと教師なしの間に位置する学習問題を議論する半教師付き学習 (semi-supervised learning)，マルコフ決定過程 (Markov decision process) とよばれる環境での学習問題を議論する強化学習 (reinforcement learning) などがある．

教師付き学習において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，学習機械は未知の状況に適応できる汎化能力 (generalization ability) をもつ．できるだけ少ない訓練データから最高の汎化能力を獲得することが教師付き学習の目標である．y が実数値をとる場合は回帰 (regression) とよばれ，y が離散値をとる場合は分類 (classification) とよばれる．また，y の相対的な大小関係を推定する順序回帰 (ordinal regression) という問題もある．

入出力関係の学習には，パラメータ (parameter) θ を含む入出力モデル (model) $y = f(x; \theta)$ を用いる．これは，入力 x を「識別」するために出力 y の事後分布 $P(y|x)$ をモデル化・学習していることに相当するため，識別モデル (discriminative model) 学習法とよばれる．また，パラメータが有限次元のときはパラメトリック法 (parametric method) とよばれ，パラメトリックモデルを用いない場合や無限次元のパラメータを含むモデルを用いる場合はノンパラメトリック法 (non-parametric method) とよばれる．教師付き学習においては，パラメータを最適に決定するための学習法 (learning method) の研究，モデルを最適に決定するためのモデル選択 (model selection) の研究，訓練データの入力 $\{x_i\}_{i=1}^n$ を最適に決定するために能動学習 (active learning) の研究，多次元の入力 x のうち出力 y を予測するのに役立つ部分集合／低次元表現を見つける特徴選択／特徴抽出 (feature selection/feature extraction) などの研究が盛んに行われている．

学習法の研究には，大きく分けて二つの流派がある．一つは頻度主義に基づくアプローチ (frequentist approach)[4, 2, 3] であり，もう一つはベイズ主義に基づくアプローチ (Bayesian approach)[5, 6] である．機械学習における頻度主義とベイズ主義は，確率論における古典統計学とベイズ統計学との関係に対応している．頻度主義的アプローチでは，適当な損失関数 (loss function) のもとでモデルのパラメータを訓練データに適合させる経験リスク最

小化原理 (empirical risk minimization principle) に基づいて学習が行われる．

$$\min_{\theta} \sum_{i=1}^{n} \ell(x_i, y_i, f(x_i; \theta))$$

ここで $\ell(x, y, y')$ は，入力 x における出力 y を y' と予測したときの損失関数である．回帰問題においては，二乗損失 (squared loss)，フーバー損失 (Huber loss)，絶対値損失 (absolute loss) などがよく用いらる．これらの損失は凸関数 (convex function) であるため，大域的最適解 (global optimal solution) を容易に求めることができる．また，フーバー損失や絶対値損失は外れ値 (outlier) に対してロバスト (robust) であるため，実用的価値が高い．一方，分類問題では誤識別率を与える 0/1 損失 (zero-one loss) を使うのが自然である．しかし，0/1 損失は非凸関数 (non-convex function) であるため，その凸近似 (convex approximation) であるヒンジ損失 (Hinge loss)，ロジスティック損失 (logistic loss)，指数損失 (exponential loss) などがよく用いられる．

経験リスク最小化原理は，統計学における最尤推定法 (maximum likelihood estimation) に対応している．従って，訓練データ数が少ない場合には，経験リスクの最小化によって学習結果がノイズの重畳した訓練データに過適合 (overfit) してしまうことがある．過適合を避けるために，正則化 (regularization) がよく用いられる．

$$\min_{\theta} \sum_{i=1}^{n} \ell(x_i, y_i, f(x_i; \theta)) + \lambda R(\theta)$$

ここで，λ (> 0) は正則化パラメータ (regularization parameter) とよばれ，正則化の強さをコントロールする．$R(\theta)$ は正則化汎関数 (regularization functional) とよばれ，ℓ_2-ノルムや ℓ_1-ノルムがよく用いられる．ℓ_1-ノルムを正則化汎関数として用いれば，解が疎 (sparse) になることが知られている．

一方，ベイズアプローチでは，パラメータの事前分布 (prior distribution) $P(\theta)$ を決めたもとで，データの尤度 (likelihood) $P(D|\theta)$ に基づいてパラメータの事後分布 (posterior distribution) $P(\theta|D)$ を計算する．

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{\int P(D|\theta)P(\theta)d\theta}$$

この式には未知の量が含まれないことから，ベイズアプローチは「学習」というよりも，むしろ「計算」であることがわかる．実際，ベイズ学習研究の主眼はいかに効率よく事後分布を計算するかにある．事後分布は事前分布の選び方に依存するため，ベイズアプローチでは学習結果を主観的にコントロールできる．これにより，訓練データ数が少ない場合でもよい学習結果が得られることがある．しかし，学習結果に主観を持ち込むことに対して否定的な意見もある．一方，最大事後確率推定法 (maximum a posteriori estimation; MAP 推定法) とよばれる，事後分布を最頻値で近似するベイズ学習法は，頻度主義における最尤推定法を正則化したものと本質的に等価である．このことから，ベイズ主義の主

観性に対する批判は，工学的にはそれほど重要ではないと考えられる．実用上は，計算に都合のよい共役事前分布 (conjugate prior) を選ぶことが多い．

事後分布 $P(y|x)$ のモデル化・推定に相当する上記の識別モデル学習法に対して，生成モデル (generative model) 学習法とよばれるアプローチもある．これは，事後分布 $P(y|x)$ がデータを「生成」している分布 $P(x,y)$ に比例することを用いて，$P(x,y)$ をモデル化・推定するアプローチである．生成モデル学習の枠組みでも，パラメトリック法・ノンパラメトリック法，および，頻度主義的学習法，ベイズ学習法が用いられる．

ところで，$P(x,y)$ が分かれば次式により $P(y|x)$ を求めることができる．

$$P(y|x) = \frac{P(x,y)}{\int P(x,y)dy}$$

しかし，逆に $P(y|x)$ が分かったとしても一般に $P(x,y)$ を求めることはできない．従って，$P(x,y)$ を推定する問題の方が $P(y|x)$ を推定する問題よりも難しいと考えられる．この考え方に基づいて，生成モデルを学習するアプローチよりも識別モデルを学習するアプローチの方がふさわしいと主張する学派がある [4]．一方，生成モデル学習の枠組みでは，データの生成分布に関する先見的知識を有効に活用することができるという利点がある．また，二つのアプローチが一致する場合もあり [1]，生成モデル学習と識別モデル学習どちらがよいかは状況に依存する．

参考文献

[1] R. O. Duda, P. E. Hart, and D. G. Stor. *Pattern Classification.* Wiley, New York, 2001.

[2] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, New York, 2001.

[3] B. Schölkopf and A. J. Smola. *Learning with Kernels.* MIT Press, Cambridge, MA, 2002.

[4] V. N. Vapnik. *Statistical Learning Theory.* Wiley, New York, 1998.

[5] 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.

[6] 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.