統計工学実験 (*Statistical Engineering*)

w.yamamoto

学習機械

判別分析も決定木分析も、統計的機械学習に分類できる手法群のひとつ。機械学習では、対象に関する属性変数 $x_1, x_2, \ldots, x_p$ と、目的変数 y との関係について、

$$y = f\left(x_1, x_2, \ldots, x_p\right) \tag{1}$$

という入出力関係を持つ学習機械 (学習すると使える機械、関数、アルゴリズム、ルールなどとも) をデータから学習する、という問題を扱う。

学習機械の種類は様々で、線形判別、線形モデル、線形ロジスティック判別とも、機械の形は

$$f\left(x_1, x_2, \ldots, x_p\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \tag{2}$$

という線形関数だが、学習に用いる最適化基準が異なる。学習される境界は直線、平面、超平面などである。

２次判別分析は、

$$\begin{aligned} f\left(x_1, x_2, \ldots, x_p\right) \quad = \quad & \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \\ & + \beta_{11} {x_1}^2 + \beta_{22} {x_2}^2 + \cdots + \beta_{pp} {x_p}^2 \\ & + 2\beta_{12} x_1 x_2 + 2\beta_{13} x_1 x_3 + \cdots \end{aligned} \tag{3}$$

という２次関数を学習する。これが学習する境界は、二次曲線、二次曲面である。

$$f\left(x_1, x_2, \ldots, x_p\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \tag{4}$$

という線形関数だが、学習に用いる最適化基準が異なる。

決定木の機械の形は

$$f\left(x_1, x_2, \ldots, x_p\right) = \begin{cases} x_1 \geq a_1 \Rightarrow c_1 \\ x_1 < a_1 \Rightarrow \begin{cases} x_5 \leq a_2 \Rightarrow c_2 \\ x_5 > a_2 \Rightarrow c_3 \end{cases} \end{cases} \tag{5}$$

と IF-THEN ルールである。これが学習する境界は、矩形領域の組み合わせである。

いずれも学習機械の出力が、変数の値に応じて変わる構造を持つため、どの範囲が y のどの値に対応するかを学習しつつ、それらの境界も学習されることになる。学習機械の性能は、誤判別率の低さ (もしくは正答率の高さ) で比較できる。

カーネル法とアンサンブル学習

判別分析も決定木も、判別の境界がさほど複雑ではない。そのため、もともと複雑な持つ問題に対して、それぞれが一定の学習限界を持ってしまう。例えば真の境界が球面の問題には、どれも良い性能を発揮しない。

この問題に対処するために、現在は

1. 線形学習機械の非線形化 (カーネル法を含む)
2. アンサンブル学習 (集団学習とも)

という２つのアプローチが主流となっている。

今回の課題

1. 決定木の学習パラメータを見直し、定期預金契約の獲得キャンペーンの戦術の提案書を作成する。提案書の部分の中身は次の通り。

1. この銀行の顧客全体がどのような構成かを、定期預金獲得への関心を念頭に、グラフや表を用いて説明する。資料を読み上げながら説明することを念頭に作成して欲しく、グラフや表は効果的に用いるべきで、多すぎても少なすぎてもいけない。(グラフや表が無駄に多いと飽きられ、少なすぎると説明が理解されない)

2. 現在の定期預金の契約状況と、契約している顧客にどのような特徴があるかを述べる。

3. 先週の課題で得た知見 (契約しているはずなのに契約していない顧客層の発見、が一番いい) を説明し、ここに重点的にキャンペーンをかけることを訴える。

4. 上のキャンペーンがどのような効果があるかを、説明する。

5. おしまい。

2. カーネル法とアンサンブル学習の方法を用いてみて、判別分析や決定木と比較し、考察せよ。それらの手法の詳細については、ウェブページのリンク先を参照すること。

`http://bit.ly/mselab2013stat` (統計工学実験)
`http://bit.ly/mselab2013` (経営情報学実験全体の情報ページ)