自動車の燃費に対する回帰

Yota Nonaka

1 Explanation of the data

1.1 Source

This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University. The dataset was used in the 1983 American Statistical Association Exposition.

1.2 Data set information

mpg(miles per gallon) とは、アメリカやイギリスで用いられている燃費の指標である。1 ガロン（約 4.55 リッター）の燃料で何マイル（1 マイル≒ 1.6km）を走ることができるのかを表したもので、この値が大きいほど燃費が良いと言える。
この実験の目標は 3 つの離散変数と 5 つの連続変数から、燃費 mpg の値を予測することである。

1.3 Attribute Information

1. mpg: continuous
2. cylinders: multi-valued discrete
3. displacement: continuous
4. horsepower: continuous
5. weight: continuous
6. acceleration: continuous
7. model year: multi-valued discrete
8. origin: multi-valued discrete
9. car name: string (unique for each instance)

The purpose of this survey is to estimate mpg using data from (2) to (9). Note that data (9) is not used for simplicity.

2 Methods

この回帰を実装する際，二つの異なる手法を用いた。一つは線形回帰であり，もう一つはガウス過程回帰である。また線形回帰の際には三種類の基底関数を利用し、それらを比較した。

2.1 Linear Regression

n 番目のデータ $\mathbf{x}_n$ の k 番目の成分はパラメータ α_k を用いて $\phi_k((\mathbf{x}_n)_k;\alpha_k)$ のように変形されるものとした。ここで ϕ が基底関数である。また $k=0$ のときのみ $\phi_0=1$ なる基底関数を用いることにした。データの予測値は $y_n=\sum_k w_k\phi_k((\mathbf{x}_n)_k;\alpha_k)$ となる。平均二乗誤差を最小化するという条件から，最適な重み $\mathbf{w}$ とパラメータ α を求めることが目標である。

2.1.1 恒等関数

基底関数として恒等関数を用いたとき，$\phi_k(\mathbf{x})=x_k$ である。よって λ というハイパーパラメータを用いると，計画行列 X に対して

$$\mathbf{w}=(X^TX+\lambda)^{-1}X^T\mathbf{y}$$

という式によって最適な重みパラメータを計算することが出来る。正則化のためのハイパーパラメータ λ は，ほかの基底関数を用いるときにも使用する。

2.1.2 シグモイド関数

基底関数としてシグモイド関数を用いると，$\phi_k(\mathbf{x})=\dfrac{1-e^{-x_k+\alpha_k}}{1+e^{-x_k+\alpha_k}}$ である。よって計画行列 Φ に対して

$$\mathbf{w}=(\Phi^T\Phi+\lambda)^{-1}\Phi^T\mathbf{y}$$

を用いると $\mathbf{w}$ を求められる。また α についても同様の計算をすると，損失関数 L の微分は

$$\frac{\partial L}{\partial\alpha_k}=2w_k\sum_n\frac{\partial\phi_k}{\partial\alpha_k}(\mathbf{x}_n)\left[-y_n+\sum_j\phi_j(\mathbf{x}_n)w_j\right]$$

という式で与えられるため，適当な初期条件を与えることによって最適化できる。シグモイド関数を基底関数として利用する場合，$\dfrac{\partial\phi_k}{\partial\alpha_k}=\dfrac{1}{2}(\phi_k^2-1)$ という関係が有用である。

2.1.3 ガウス関数

ガウス関数を基底関数として選んだとき，シグモイド関数とまったく同様の手順でパラメータを最適化することが出来る。基底関数は $\phi_k(\mathbf{x})=\exp\left[-\dfrac{(x_k-\mu_k)^2}{2\sigma_k^2}\right]$ であるから、

$$\begin{aligned}\frac{\partial\phi_k}{\partial\mu_k} &= \frac{x_k-\mu_k}{\sigma_k^2}\phi_k\\ \frac{\partial\phi_k}{\partial\sigma_k} &= \frac{(x_k-\mu_k)^2}{\sigma_k^3}\phi_k\end{aligned}$$

という関係を利用することができる。

2.2 ガウス過程回帰

前述した線形回帰の基底関数の取り方では，データの各成分同士の相関などを考慮していない。しかし基底関数に相関を取り入れると，データの成分の組み合わせの数が大きすぎるという問題が生じうる。そこで基底関数を明示的に指定しないガウス過程回帰を利用することにした。利用したカーネルはガウスカーネルであり，カーネルに含まれるハイパーパラメータは検証用データでチューニングした。
学習用データ $\mathbf{x}_n$ と正解ラベル $\mathbf{y}$ が与えられたとき，データ $\mathbf{x}^*$ に対する予測値は $y^* = \mathbf{k}_*^T K^{-1}\mathbf{y}$ となる。ここで $(\mathbf{k}_*)_n = k(\mathbf{x}^*, \mathbf{x}_n)$ と $K_(nn') = k(\mathbf{x}_n, \mathbf{x}_{n'})$ を利用した。

3 実験

全 392 個のデータのうち、20% をテストデータとして用いることにした。残りの 80% のデータに関して、5 分割交差検証を行い、5 つの場合に対する精度の平均を計算した。最後に、精度が最も良いと考えられるモデルについて、テストデータを使ってその精度を確認した。

4 結果

それぞれのモデルによる相対誤差を比較したものを表 1 に示す。この表から、ガウス過程回帰が最も精度の良いモデルであるということが読み取れる。また交差検証に対する標準偏差も小さく、安定性も保証されている。最後にガウス過程回帰を用いてテストデータの回帰を行ったところ、相対誤差は 8.69% であった。

表 1: モデルごとの精度の比較

Method	basis function or kernel	average of relative error	(STDEV)
Linear Regression	id	12.2%	1.2%
	sigmoid	18.0%	6.5%
	gaussian	17.2%	7.2%
Gaussian Process	RBF	9.5%	1.4%

テストデータに対してもよい精度を保っているため、過学習を起こしていないということも保証される。

5 結論

表 1 より、線形関数の基底関数を比較すると、シグモイドやガウシアンよりも恒等関数のような関数形が mpg の回帰の特徴をとらえているということが出来る。また、恒等関数による線形回帰の重みパラメータより、データの各成分の典型的な大きさを考慮しつつ、各成分の mpg に対する寄与を調べた。このとき最も主要な寄与として、製造年は mpg に対して大きな正の効果を与えることが分かった。次に主要な寄与として、車両の重量は mpg に対して大きな負の効果を与えていた。これらは直観的にも極めて自然である。その他の寄与はこれらに対して 10% 程度の大きさであった。
またガウス過程回帰が最も精度の良いモデルになっていることから、各成分が独立な基底関数を与えるという考えでは不十分であるということも読み取れる。つまり基底関数の選択において、各成分の相関を考

慮に入れる必要があるということである。ただ前述の通り、そのためには膨大な組み合わせを考慮する必要があり、データ数はそれほど大きくないため、ガウス過程は現実的な解決策であるといえるかもしれない。