**9.1 Generalized Additive Models**

線形回帰モデルはシンプルであり広く用いられているが、現実の世界の現象は線形で表現できるわけでなく、適切でないこともある。非線形な関係を説明するために用いられるモデルが”Generalized Additive Models (GAM)”である。

E (Y |X1, X2, . . . , Xp) = α + f1(X1) + f2(X2) + · · · + fp(Xp).

で表されるモデルであり、Generalized Linear Models (GLM)のようにロジット関数などのリンク関数に当てはめることができる。GAMは様々な設定で線形モデルから置き換わる事ができる。

9.1.1 Fitting Additive Model

GAMをあてはめるためには、backfitting algorithmを用い、そのスムーシング演算子にcubic smoothing splineを利用する。他にも局所多項式回帰、カーネル平滑化などの演算子も利用できる。線型方程式を解くガウス＝サイデル法と等価である。 GAMは最大対数尤度を求める事になるが、重み付け線形回帰が重み付けbackfitting algorithmに置き換わるものである。

9.1.2 Example： Additive Logistic Regression

generalized additive logistic modelは下記のようになる。

logPr(Y =1|X) =α+f1(X1)+···+fp(Xp).

このモデルにweighted scatter smootherを用いたアルゴリズムをあてはめた (Algorithm 9.2)。これを用いて、4601通のメールからスパムメールを分類するためにGAMを適応した。スパムメール：1、正常メール：0と二値化し、57項目の予測因子を設定しランダムに抽出した1536通をテストセットとし、3065通をトレーニングセットとした。overall error rateは5.3%であった（線形ロジスティックは7.6%）。Figure 9.1を見ると多くは非線形である事がわかる。このことから線形ロジスティックより、GAMの精度が上がった理由である。モデルのあてはまりを良くするために、交互作用を検証する必要がある。交互作用項はマニュアルで入れることもできるし、MARSを用いて自動で入れることもできる。

9.1.3 Summary

GAMはGLMの「解釈のしやすさ」を残しながら、より汎用性の高いモデルである。backfitting法も簡素でありそれぞれの説明変数に適切なあてはめを選んでくれる。以上よりGAMは統計の政界で広く使われてきた。ただしGAMは大きなデータマイニングに応用するときに欠点がある、backfitting法はすべての説明変数に対し適応するが、多くの説明変数がある時は適切でない。大きなデータマイニングにはブースティングがより有効とされ、モデルに多くの交互作用も含める事ができる。

**9.2 Tree-Based Methods**

9.2.1 Background

Tree-Based Methodsは特徴のあるもの同士を囲い、それぞれをシンプルなモデルに合わせるもので、シンプルな概念でありかつ強力な方法である。Figure9.2上左図は分類が複雑であるが、上右図のように、再帰的な2分割法を用いることでシンプルに分類できる。下図左はそれを対応させたもので、下図右は立体的に表現している。視覚的で解釈しやすい。

9.2.2 Regression Trees

回帰木を成長させる方法は、貪欲法と呼ばれるアルゴリズムを用いる。すべてのデータが含まれる点から開始する。予測変数j、分割点sを考えた時、予測空間は

R1(j, s) = {X|Xj ≤ s} and R2(j, s) = {X|Xj > s}.

この時のj、sは9.13式で求められる。

最適な木の大きさはデータに依存する。最適な木を探すためにcost-complexity pruningという方法がある。T0の刈り込みによりできる部分木をT ⊂ T0 とし、それぞれのαに対するCα(T)を最小にするTα ⊆ T0 を求める。



αが正であれば、モデルのあてはまりと木の大きさはトレードオフの関係にある。このTαを求める方法がweakest link pruningである。

9.2.3 Classification Trees

目的が分類になった場合、分割点と木の刈り込みの基準をアルゴリズムに追加するだけで良い。ただし回帰木の場合残差平方和を使ったが、分類木では適切でない。そこで誤分類率、ジニ指数、クロスエントロピーを用いる。後者二つは誤分類率に比べ、ノードの純度に関してより感度が良い。

9.2.4 Other Issues

Categorical Predictors

q個の2分割点があった場合、2q-1-1通りに分けられ、qが大きな数であれば計算が難しい。ただしアウトカムが2値であればシンプルになる。アウトカムが多値にあれば複雑になる。

The Loss Matrix

Missing Predictor Values

欠測値がありそれを解析外にするとトレーニングセットに重大な問題が生じる。解決策は二つあり、”missing”として分類することと、代理の値と分割点を設けることである。

Why binary splits?

分割を多分割にするとデータがすぐにばらけてしまい、次のステップで不十分なデータしか残らない。多分割はできるだけ行わない方が良い。

Other Tree Procedures

Linear Combination Spirits

木の予測能は向上するが、解釈が難しくなる。

Instability of Trees

木は分散が大きいという欠点がある。階層的な構造のため、前の分割でのエラーが後の分割時点で増幅されてしまう。バギングはこの問題を軽減してくれる。

Lack of smoothness

もう一つの欠点は、予測に平滑さがないことである。MARSはCARTを修正したもので予測に平滑さを与える。

Difficulty in Capturing Adding Structures

木は加法モデルを構成できないことも問題である。これもMARSで解決できる。

Spam Example

Figure 9.4の10回交差検証法で決定木の精度を見ると、木のサイズでなくαに依存していることがわかる。

Figure 9.5がspamメールの決定木である。overall error rateはGAMに比べ50%高くなっている。感度特異度を算出し、ROC曲線を描いてみると、GAMの方が決定木よりAUCが高かった。