機械学習の盲点

学籍番号 21613973　田中 秀成

2018 年 11 月 23 日

1　イントロダクション

近年は機械学習という単語がメディアの注目を浴び, AI 関連の話題も至る所で目にするようになった．それは偏にディープラーニングが発展してきたおかげであるが, 果たしてそれはパワーワード化してはいないだろうか．実際, 先行論文 [1] ではニューラルネットワークの有用性を示唆しながらも, データの性質による部分もあるとしている．この論文ではそのような知見を踏まえ, 降雨予測においてロジスティック回帰・サポートベクトルマシーン (SVM)・XGBoost を用いてデータの性質を変えつつ精度比較を行い, その特徴を見ることを目的とする．

2　モデル説明

このセクションでは用いるモデルについて説明していく．今回は計量モデルとしてロジスティック回帰, 機械学習のモデルとして SVM と XGBoost を使用した．ロジスティック回帰については機械学習のモデルでもあるが, 本稿ではモデルが単純で実装が容易であることから他二つとの比較対象として位置付ける．

2.1　ロジスティック回帰

ロジスティック回帰とはリンク関数にロジットを関数を使用する一般化線形モデルの一種で, 統計学・計量経済学・機械学習の分野で幅広く分類目的で用いられる手法である．通例は 2 値分類で線形分離可能なデータに用いられるが, 多項ロジットや非線形ロジットも存在し, そのようなデータにも対応可能である．説明変数 x_i の線形結合和をシグモイド関数 $\frac{1}{1+e^{-ax}}$ で変換することにより, あるクラス j に属する確率 $\pi_{i,j}$ を算出する．導出については [2] を参照されたいが, 一般式で書くと以下のようになる．

$$\pi_{i,1} = \frac{1}{1+\Sigma_{r=2}^{J} e^{(x_i^t \beta_r)}}$$

$$\pi_{i,j} = \frac{e(x_i^t \beta_j)}{1+\Sigma_{r=2}^{J} e^{(x_i^t \beta_r)}} \quad (j = 2, 3, ...J)$$

このパラメータ β について尤度関数を計算し, 尤度が最大となるような値を求めるのだが, 解析的に不可能なため MCMC やニュートン法を用いて近似解を求める．

2.2　サポートベクトルマシーン (SVM)

サポートベクトルマシーン (SVM) とは認識性能が高く分類問題において優れているモデルの一つで, 線形でも非線形でも分類可能である．線形分離についてはマージン最大化という概念を取り入れており, 最も単純な 2 値分類問題においては $y(x) = w^t x + w_0$ の境界線が, 境界線と最も近いデータとの距離（マージン）を最大化するように分離超平面を決める．n 番目のデータに関して $u \geq 0$ ならば 1, $u < 0$ ならば -1 を取るような指示関数 $t(u)$ を用いて, $t(y(x_n))y(x_n)$ を考えると, これは必ず正の値をとる．また, $|t(y_(x_n))| = 1$ より, マージンは $\min_n(\frac{|y(x_n)|}{|w|}) = \min_n(\frac{t(y(x_n))y(x_n)}{|w|}) = \min_n(\frac{t(y(x_n))(w^t x_n + w_0)}{|w|})$ となり, この最適化問題をラグランジュ法を用いて双対問題として解く．ここで重要なのはマージン最大化の過程でほとんどのデータは無関係になっており, それにより未学習データに対しても高い精度を誇ることができる点だ．また外れ値などの影響を受け, 完全に境界線で分けることができないデータに対しては, 多少の識別誤りを許す制約を加えたソフ

トマージンを用いることで対応可能とし, その汎化性能を高めている. ただしそれでも本質的に線形分離できない問題にはソフトマージンは有用ではなく, そのようなデータに対してカーネルトリックが開発された. カーネルトリックとは, 元の特徴ベクトル x を非線形の写像 $\phi(x)$ によって変換し, その空間で線形識別を行うことを指す. 最適化問題の目的関数と識別関数 $t(y(x_n))y(x_n)$ が内積にのみ依存しているため, $\phi(x_1)^t\phi(x_2) = K(x_1, x_2)$ のように非線形に写像した空間での内積が計算できるなら, K から陰に最適な写像を構成できる. このカーネル K は計算が容易なものが実務的に用いられ,

$$\text{多項式カーネル } K(x_1, x_2) = (1 + x_1^t x_2)^p$$
$$\text{ガウスカーネル } K\ (x_1, x_2) = e^{\frac{-||x_1 - x_2||^2}{2\sigma^2}}$$
$$\text{シグモイドカーネル } K(x_1, x_2) = tanh(ax_1^t x_2 - b)$$

などがその例だ. カーネルにガウスカーネルを用いると, 従来のニューラルネットワークと同様の構造になり, シグモイドカーネルを用いると 3 層パーセプトロンと一致する. 詳細については [3] などを参照.

2.3 XGBoost

XGBoost は eXtreme Gradient Boosting の略で勾配ツリーモデルを応用させたものである. スパースデータに柔軟な対応を与え, サブサンプリングや退縮化といったシステムをモデルに組み込むことで汎化性能が高まり人気を博している. モデルの元は決定木であるため線形・非線形を問わず分類も回帰も行える. n 個のデータが m 個の特徴量を持っているとして, (すなわち説明変数 X が $n \times m$ 行列で, 被説明変数 y が $n \times 1$ 行列) $\hat{y_i} = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i)$ の下で以下の損失関数の最小化を考える.

$$L(\phi) = \sum_i l(\hat{y_i}, y_i) + \sum_k \Omega(f_k)$$
$$\Omega(f) = \gamma T + \frac{1}{2}\lambda||w||^2$$

ここで f_k は各ツリーを意味し K は総ツリー数である. ただし, 各ツリーは葉っぱの数 T と, 回帰の場合のみ各葉っぱのの重み w_i を内包する. また, l は予測と実測値の誤差を測定するもので, 例えば二乗誤差などである. Ω は罰則項で, 葉っぱの数 T が多くなりすぎることと, その葉っぱへの重みが大きくなりすぎることへのペナルティーである. これらがあることで過学習を抑制できることになり, γ と λ は使用者がチューニングして最も精度が良くなるものを選択する. この最小化問題を解くにあたり, ツリーの個数と各ツリーの葉数はクロスヴァリデーション法で決定する. K を所与とすると, 葉の返す値については t 番目のツリーを付け加え,

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y_i}^{(t-1)} + f_t(x_i)) + \Omega(f_t)$$

の最小化問題とする. ここから勾配ブースティングと呼ばれる手法を用い, $g_i = \partial_{\hat{y_i}^{(t-1)}} l(y_i, \hat{y_i}^{(t-1)})$ $h_i = \partial^2_{\hat{y_i}^{(t-1)}} l(y_i, \hat{y_i}^{(t-1)})$ とし, 目的関数を 0 近傍でのテーラー展開を行うと,

$$L^{(t)} \simeq \sum_{i=1}^{n} (l(y_i, \hat{y_i}^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_i^2(x_i)) + \Omega(f_t)$$

となる. 微分していく方針なので定数項を外して,

$$\overline{L}^{(t)} = \sum_{i=1}^{n} (g_i f_t(x_i) + \frac{1}{2} h_i f_i^2(x_i)) + \Omega(f_t)$$

またここで, I_j を T 個ある葉っぱのうち, j 番目に含まれている説明変数だとすると, その j 番目の葉っぱが返す値が w_J より,

$$\begin{aligned}\overline{L}^{(t)} &= \sum_{i=1}^{n} (g_i f_t(x_i) + \frac{1}{2} h_i f_i^2(x_i)) + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \\ &= \sum_{j=1}^{T} ((\sum_{i \in I_j} g_i) w_j + \frac{1}{2}(\sum_{i \in I_j} h_i + \lambda) w_j^2) + \gamma T\end{aligned}$$

途中の式変形については w_j を表に出したかったので, もともと i で数え上げていた n 個のデータを, 各 w_j に含まれる物毎に T 個分数え上げただけである. これを微分することで,

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}$$

を得ることができる. またこの値を再度 $\overline{L}^{(t)}$ に代入することで, スコア関数

$$\overline{L}^{(t)} = -\frac{1}{2} \sum_{j=1}^{T} \frac{(\sum_{i \in I_j} g_i)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T$$

を定める．ここまでは t 番目のツリーの j 番目の葉っぱがどのような値を返せばいいのかを求めてきたが, 肝心なデータの分割方法についてはまだ触れていない．j 番目の分割で $\overline{L}^{(t)}$ を最小にするのならば, 右辺の $\sum$ の項が大きくなれば良いことが分かる．そのため, $j-1$ 番目の葉っぱを分割する際には j 番目の分割で左右に分かれたとき, 分割前と分割後のこの項の大小関係を比べればよい．具体的にはすべての分割方法のうち,

$$L_{split} = \frac{1}{2}(\sum_{j=1}^{T} \frac{(\Sigma_{i\in I_L} g_i)^2}{\Sigma_{i\in I_L} h_i} + \sum_{j=1}^{T} \frac{(\Sigma_{i\in I_R} g_i)^2}{\Sigma_{i\in I_R} h_i} - \sum_{j=1}^{T} \frac{(\Sigma_{i\in I} g_i)^2}{\Sigma_{i\in I} h_i})$$

を最も大きくする候補を一つ選んでそこで分割を行う． その他のシステムについては元論文 [4] を参照されたい.

3　データ説明

データは気象庁のサイト [5] からダウンロードできる東京都の気象情報について 2008 年 10 月 25 日～2018 年 10 月 25 日までの 10 年分を用いる．このうち最初 9 年分を学習データとし, 残り 1 年分で検証する.

4　問題設定

説明変数としてデータのうち, 平均が 0 で分散が 1 になるように標準化した第 k 日の平均気温・最高気温・最低気温・日照時間・日光照射量・積雪量・平均風速・最大風速・平均蒸気圧・平均気圧・平均海面気圧・平均湿度・平均雲量を用いる．被説明変数は第 k+1 日の降水量とするが, 区分けを行い分類問題とする．区分けは 3 通り行い, ①[0,0～] の降雨の有無を予測する 2 区分, ②[0,0～10,10～20,20～] の 4 区分, ③[0,0～5,5～10,10～30,30～50,50～] の 6 区分とする． XGBoost のチューニングは max-depth[5,6,7], subsample[0.8,0.9], colsample-bytree[0.8,0.9], learning-rate[0.1,0.15], reg-lambda[0,1], gamma[0] の範囲に固定する．精度の比較方法は, 予測したデータ 364 個のうち区分けに的中した割合で行う.

5　検証結果

上記の条件で検証してみた結果をまとめる.

	ロジスティック回帰	SVM	XGBoost
①2 区分	0.871	0.843	0.838
②4 区分	0.780	0.777	0.742
③6 区分	0.728	0.739	0.720

表 1　検証結果の精度

2 区分においては気象庁が発表しているデータとほとんど変わらない精度が得られた．またロジスティック回帰が最も精度が良い．4 区分においてもロジスティック回帰が最も精度が良いが, SVM もそれに迫る精度となり, 2 区分の時よりロジスティック回帰と精度の差が縮まっている．6 区分になると最も精度が良いものが SVM に変わり, ロジスティック回帰が 2 番目となった．XGBoost については今回全ての区分けで最も精度が悪くなったが, 精度の分散は最も小さかった.

6　考察と展望

今回の結果からわかることは, どんなデータに対しても SVM や XGBoost などの複雑なモデルが最も良い精度を示すわけではないということである．予測データが単純な場合はロジスティック回帰のような単純なモデルが最も当てはまりよく, 複雑なデータに対しては複雑なモデルの方が良い．従って, まずはモデルをあてはめる前にしっかりとデータの性質を確認することが必要である．今後の展望としてはデータの複雑さを示す指標を考案し, それに応じて適当なモデル選択を可能にすることや回帰問題でも同様のことが言えるのかを示すことにある.

参考文献

[1] 小泉耕, 平沢正信 (2001)“降水量予測に適したニューラルネットワーク構造”, 天気.48(12), 2001-12-01, 885-892

[2] D.R.Cox(1958)“The Regression Analysis of Binary Sequences”, *Journal of the Royal Statistical Society. Series B(Methodological)*,Vol.20, No.2, 215-242

[3] 栗田多喜夫 “サポートベクターマシン入門”
産業技術総合研究所脳神経情報研究部門
[4] Tianqi Chen,Carlos Guestrin(2016) “Xgboost: A scalable tree boosting system”,
Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 785-794
[5] 国土交通省気象庁
http://www.data.jma.go.jp/obd/stats/etrn/index.php,
2018-11-18