データサイエンスとその活用

2024年11月16日

早稲田大学客員教授
岩沢 宏和

Iwahiro

本講演の内容

1. はじめに
   (1) 自己紹介・立場
   (2) 数理の楽しみ
   (3) 本講演で伝えたいこと
2. データサイエンスとは
   (1) データサイエンス，予測モデリング，機械学習
   (2) 機械学習とは何か
3. 機械学習手法の発想
   (1) GLM
   (2) 決定木
   (3) ランダムフォレスト
   (4) 開発例
4. むすび

Iwahiro

はじめに

Iwahiro

（本講演に即した）自己紹介

- OLIS主催，「保険」フォーラムとの関連でいえば，
  - 保険数理の専門家であるアクチュアリーに関連する分野（特にデータサイエンスと損保数理）の教育者．
    - 日本アクチュアリー会で年間100時間以上講義をしている唯一の人物
    - 東京大学（3＋α科目），早稲田大学（3科目），東北大学（集中講義+セミナー），信州大学（4人で1科目）
  - 特に日本のアクチュアリー界のデータサイエンス普及の中心人物（研修・講座・WG・研究会…）．
  - また，研究者ではないが，近年はデータサイエンス分野の論文（主に共著）を毎年複数書いている．
- フォーラムのテーマにある「数理科学人材」との関連でいえば，
  - アクチュアリー分野をはじめ，数理の素養を活かしていろいろなことをしている．特に，数理パズルは…
  - 著書多数．「数学セミナー」誌ではこれまで多数の記事（連載を含む）を書いており，特に「エレガントな解答をもとむ」欄でよく出題している．
- 実務に関していえば，
  - 法人に属してアクチュアリー実務（年金アクチュアリー）をしていたのは20世紀．

Iwahiro

講演者の立場

- データサイエンスセンターを有している本学にて「データサイエンスとは何か」を高い壇上から語るのにふさわしいほど私はデータサイエンスを代表する者ではまったくない.
- それでも，（規模だけでいえば，きわめて巨大な）産業で活躍する専門家集団の中でデータサイエンスの普及の中心にはいるので，実社会でのデータサイエンスについて，学生のみなさんに参考となる話はできるかと思う.
- 特に，最先端「技術」の話をするというよりは，「数理科学」人材が，地に足の着いた形で活躍できる分野としてのデータサイエンスの一側面の話をしたい.

Iwahiro

数理の楽しみ

問題1

2人のプレーヤーが歪みのないコインをそれぞれ5枚ずつもっている．コイン5枚には計10個の面があるので，各プレーヤーは，そこに好きなように1から10の数を一つずつ（各数は1回だけ使って）書く．準備ができたら，2人はそれぞれ自分の5枚のコインを投げ，積が大きいほうを勝ちとする（積が同じときは，勝負がつくまで繰り返す）．勝つ確率を最大化するには，コインの面にどのように数を書いたらよいだろうか.

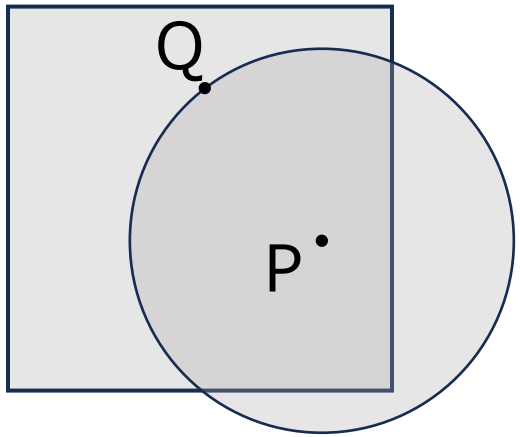
もし両方が最適な戦略をとれば，当然，勝負は五分五分になるが，相手がどんな戦略をとっても勝つ確率が五分五分以上となる方法はあるか．あるとすればそれはどのようなものか．同等のものが複数あるとしたらそのすべてを挙げよ.

Iwahiro

数理の楽しみ

問題2

1辺の長さが1の正方形内に2点P,Qをランダムにとる*とき，点Pを中心とし線分PQを半径とする円と，正方形との共通部分の面積の期待値を求めよ.



* ある点を「正方形内にランダムにとる」というのは，正方形内の任意の領域について，その点がその領域内に入る確率が，その領域の面積に比例するようになっている，ということ.

Iwahiro

本講演で伝えたいこと

- 世の中でデータサイエンスの人気は高く，講演者が関係している業界でも同様である.
- データサイエンスのスキルはいろいろとあるが，数理に強い人が，この分野のさまざまな手法や技術を数理的に理解することを通して習得しようとするのは，かなり有効なアプローチだと思う.
- また，専門的な研究者の道に進まなくても，数理的な発想をもとに実用的な研究開発をすることは十分に可能であるし，社会的価値が高いことだと思う.
- 数理が好きならば，これらのことに携わるのは，有利であったり価値が高かったりするのに加え，きっと楽しいことだと思う.

Iwahiro

（本講演でいう）
データサイエンスとは

Iwahiro

いくつかの用語

- データサイエンス，予測モデリング，機械学習
  - 今日の話の限りでは，対象となる手法はほとんど同じ．
- データサイエンスと予測モデリング
  - アクチュアリー分野で主に使われる（べき）データサイエンスは，予測モデリングとよばれるもの．
  - 予測モデリングとは，「真のモデルでなく有用なモデルを求め，予測の精確さによってモデル選択を行う」という文化のもとでの統計モデリング．
- 予測モデリングと機械学習
  - 予測モデリングで使われる手法には機械学習以外のものもあるが，以下では主に機械学習のことを念頭に置く．
- 機械学習とAI
  - 両者がほぼ同義で使われるときもあるが，今回は区別し，機械学習のみ念頭に置く．
  - AIは，人間の知的な行動を模倣するシステムや技術の総称．
  - 機械学習は，AI分野で生まれたものも含まれるが，利用上は模倣に主眼はなく，データを使ってモデルを訓練し，パターンを見つけ出して予測や判断を行う技術．

以下では機械学習の話をする．

Iwahiro

機械学習とは

データ

機械学習手法

・計算集約的（ただし，計算量が爆発しない巧妙なアルゴリズム）

・自動化（ただし，念入りなチューニングも可能）

・柔軟な表現力（行儀のよいあらゆる関数をいくらでも近似できる）

予測
モデル
（機械学習モデル）

Iwahiro

機械学習手法の例

- 線形回帰を発展させたもの
  - GLM　←あまり機械学習とはよばれないが…
  - RidgeやLassoなどの正則化線形回帰
  - 正則化GLM
  - サポートベクトル回帰
- 決定木系
  - 決定木
  - ランダムフォレスト
  - 勾配ブースティング木（XGBoost，LightGBMなど）
- ニューラルネットワーク系

  …多層でよく作り込まれたものは深層学習（ディープラーニング）とよばれる
  - パーセプトロン
  - 3層ニューラルネットワーク
  - 畳み込みニューラルネットワーク
  - 再帰型ニューラルネットワーク

Iwahiro

機械学習手法の発想

Iwahiro

GLMとは何か

$x:=(x_1,\cdots,x_p)^T$，$\beta:=(\beta_1,\cdots,\beta_p)^T$として，線形モデルを

$Y\sim N(\mu,\sigma^2)$（正規分布）

$\mu=\beta_0+x_1\beta_1+\cdots+x_p\beta_p=\beta_0+x^T\beta$　（線形表現）

と捉えたうえで，

Yが従う分布：　正規分布　→　正準指数型分布族の分布

線形表現で表されるもの：　μ　→　$g(\mu)$

と一般化したのがGLM.

Iwahiro

GLMのねらい

- GLMの解は$-\ell(X, \beta_0, \beta)$を最小化するβ_0,βによって定まる回帰式.
- これは，最尤法を根拠として，統計学的に正当化される．また，その解は，コンピュータにより高速に精確に一意に決定される.
- …というよりも，こうして，最尤法に基づいてコンピュータにより高速に精確に一意に解が求まるモデルとして考案されたのがGLM！

Iwahiro

GLMの特徴

- GLMは「真のモデル」の候補として得られるものではない.
- GLMは，最尤法をコンピュータで行おうとしたときに効率よく実行できる範囲で線形モデルを拡張したとしたらどこまで広げられるかという問題を追求することから得られるもの.
- GLMは，AIC等の情報量規準の計算とも相性がよい.
- こういうモデルを採用することは，「真のモデルでなく有用なモデルを求め，予測の精確さによってモデル選択を行う」という予測モデリングの考えにまさしく沿っている.

Iwahiro

GLMの理屈の要点

- いろいろな教科書に書いてあるように，線形モデルの回帰係数のベクトルβは，$\beta=(X^TWX)^{-1}X^TWy$という式から，計算機を使えば（ランク落ちしていない限り）高速に精確に一意に求まる．右辺の各記号は，学習データに基づくもので，Xは計画行列，Wは各対象の重みを表す対角行列，yは目的変数のベクトル．
- ここで大事なのは，$(X^TWX)^{-1}X^TW$という形の計算は（計算機を前提とすれば）非常に「よい」性質をもっているということである．
- GLMはその「よさ」を活用する．

Iwahiro

GLMの理屈の要点

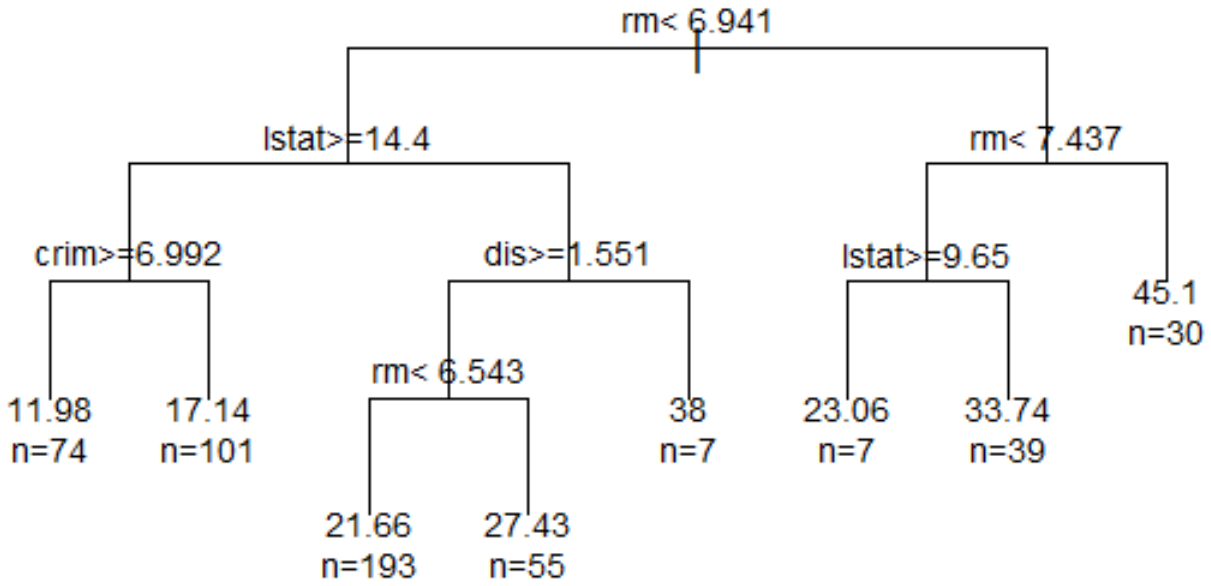
- GLMの回帰係数のベクトルβは，下記の更新式に基づいて計算（反復計算し，収束してきたら停止）することができる．ただし，gをリンク関数とすれば，$\mu^{(r)}=g^{-1}(X\beta^{(r)})$であり，$W_H^{(r)}$,$W_F^{(r)}$,$D^{(r)}$はいずれも$\mu^{(r)}$をもとに簡単に計算できる対角行列である（細かいことをいえば，下式はニュートン法の場合で，Fisher's scoring法の場合は$W_H^{(r)}=W_F^{(r)}$であり,ほかにも種々の準ニュートン法がある).

$$\beta^{(r+1)}=\beta^{(r)}+(X^TW_H^{(r)}X)^{-1}X^TW_F^{(r)}D^{(r)}(y-\mu^{(r)})$$

- それゆえ，GLMは高速に精確に一意に実行することができる．ちなみに，反復計算の収束は早く，データが大きくなければ，ほんの数回の反復で済む場合も多い.

Iwahiro

決定木とは何か



決定木（回帰木）の例

Iwahiro

回帰木という手法を再発見してみる

学習データ（標本サイズn）からランダムにとってきた観測対象iの目的変数y_iの値を当てたい．評価基準は2乗誤差(誤差の平方）とする．

- 何のヒントもないときは，どういう値をいうべきか→全体の平均値$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$をいえばよい．そのとき，2乗誤差の期待値に標本サイズを乗じたもの（誤差の平方和）は，分散分析でいう全体平方和$SS = \sum_{i=1}^{n}(y_i - \bar{y})^2$となる．
- ある特徴量x_jについて「$x_j \geq a$」かと聞いてyesの場合，どういう値をいうべきか．noの場合はどうか→yesなら$x_j \geq a$を満たす対象の平均値をいい，noなら$x_j < a$を満たす対象の平均値をいう．

Iwahiro

回帰木という手法を再発見してみる

- その場合，何のヒントもないときよりも，誤差の平方和はどれくらい減るか→新たな平方和はいわゆる群内平方和なので，もととの差は2つの群の群間平方和(これは計算が簡単).
- 補足

$A = \{i | x_{ji} \geq a\},\ A^c = \{i | x_{ji} < a\},\ n_A = |A|,\ n_{A^c} = |A^c| = n - n_A,\ \bar{y}_A = \frac{1}{n_A}\sum_{i\in A} y_i,\ \bar{y}_{A^c} = \frac{1}{n_{A^c}}\sum_{i\in A^c} y_i$とすれば，

群内平方和$\mathrm{WSS} = \sum_{i\in A}(y_i - \bar{y}_A)^2 + \sum_{i\in A^c}(y_i - \bar{y}_{A^c})^2$

群間平方和$\mathrm{BSS} = n_A(\bar{y}_A - \bar{y})^2 + n_{A^c}(\bar{y}_{A^c} - \bar{y})^2$

群内平方和WSS +群間平方和$\mathrm{BSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2 =$ 全体平方和SS

Iwahiro

回帰木という手法を再発見してみる

- 特徴量 $x_1,\cdots,x_p$ に関してyes/noで答えられる単純な質問を1個だけして，その答えをもとに当てるときはどういう質問をしたらよいか．単純な質問とは，どれか1つの特徴量に関して，その特徴量がある値以上かどうかを聞くものだとする→そうした質問の候補は，各特徴量のユニーク値の個数をそれぞれ $m_1,\cdots,m_p$ とすると， $m_1+\cdots+m_p-p$ 個ある．それらの候補のうち，できあがる2群の群間平方和が最大のものを選べばよい．
- 要は，

2乗誤差を小さくしたい→群内平方和の最小化

→群間平方和（計算が楽）の最大化

Iwahiro

回帰木という手法を再発見してみる

- 木の構造になるように単純な質問をあらかじめ用意して群を分けていくこととし，最初の質問は上で決めたものとするとき，次の枝分かれの質問はどういうものとしたらよいか→最初の質問で枝が2本になっており，質問の候補は高々2(m_1+…+m_p-p)個ある．それらの候補のうち，できあがる3群の群間平方和が最大のものを選べばよい．
- 同様のことを繰り返せば，回帰木ができる．ずっとやると，群の個数は観測対象の個数に一致し，過適合が起きる．どこで停止したらよいだろうか→群間平方和の増え具合が一定率（`rpart`のデフォルトは`cp = 0.01`）を下回ったらもう分割しないようにする．

Iwahiro

決定木の利用とランダムフォレスト

- 決定木は枝分かれが少ないと適合不足，多いと過適合になりがちで，単独では（わかりやすさでは優れているが）予測の点では劣る.
- その原因は，単純な枝分かれのうちで最適のものを一歩一歩トップダウンで貪欲に選んでいるために，結果的にできあがる全体としての予測力が高まりにくくなっていることにある.
- とはいえ，複雑な枝分かれ規則を採用すると，せっかくの高速な計算の魅力が損なわれる.
- そんな中，互いの相関が小さいたくさんの弱い学習器の平均をとることとしたものは強い学習器となるという事実が知られていた.

Iwahiro

決定木の利用とランダムフォレスト

だったら，互いに相関の小さい，弱い決定木をたくさん作って（いわば森にして），平均をとればよいのではないか.

↓

では，相関が小さくなるようにするにはどうすればよい？

↓

ランダム化したらどうだろうか？

↓

うまくいった！

ランダムフォレストの誕生

Iwahiro

決定木の利用とランダムフォレスト

相関の低い，弱い決定木を作るための2つの工夫

1. 学習データの標本そのものは使わず，ブートストラップ標本（標本から対象を，標本サイズと同じ個数だけ復元抽出して作る標本）をもとに決定木を作る．これにより，平均的には6割強程度の対象しか使われない．こうしてブートストラップをもとに平均をとる手法を一般にバギング（Bootstrap AGGregatING）という．
2. 枝分かれの際に候補とする特徴量は，枝分かれのたびに一定割合（`randomForest`で回帰の場合のデフォルトは1/3）でランダムに抽出し，それらの特徴量のみを候補として枝分かれの規則を選ぶ．

Iwahiro

新手法開発の例：AGLM

- 講師が2017年に発案した手法AGLMを，日本アクチュアリー会の研究チームが開発.
- Rパッケージ`aglm`（いまではCRAN登録済）.
- 同手法に関して2020年3月に執筆した次の論文が2021年のHachemeister賞を受賞.

  AGLM: A Hybrid Modeling Method of GLM and Data Science Techniques

- 損保アクチュアリー分野での表彰だが，手法としては，生保などその他の分野でも十分に利用可能.

Iwahiro

AGLMの受賞理由

- この論文で主張するハイブリッドモデリングアプローチは，データサイエンス技術による正確性の向上とGLMの優れた説明力の両方をバランスよく得たいという実際的な要求に応えようとしている.
- 自動車保険の例を用いて，既存のモデリング手法に対するAGLMの利点を定性的かつ定量的に評価している.
- 論文とともに統計ソフトウェアRのパッケージも用意しており，AGLMを簡単に使用できるように配慮している.
- 論文では自動車保険の価格設定の例を示しているが，提示される概念は他の損害保険分野，更にはその他の応用にも発展することが期待される.
- 全体として，委員会は，この論文はすべてのアクチュアリーにとって興味深いものであり，実用的な使用と独創的な思考を含み，大変優れた論文であると評価した.

Iwahiro

おわりに

Iwahiro

本講演で伝えたかったこと（再掲）

- 世の中でデータサイエンスの人気は高く，講演者が関係している業界でも同様である.
- データサイエンスのスキルはいろいろとあるが，数理に強い人が，この分野のさまざまな手法や技術を数理的に理解することを通して習得しようとするのは，かなり有効なアプローチだと思う.
- また，専門的な研究者の道に進まなくても，数理的な発想をもとに実用的な研究開発をすることは十分に可能であるし，社会的価値が高いことだと思う.
- 数理が好きならば，これらのことに携わるのは，有利であったり価値が高かったりするのに加え，きっと楽しいことだと思う.

Iwahiro

ご清聴ありがとうございました.

Iwahiro