EGEM勉強会

用語集

用語	説明
トランザクションデータ	業務に伴って発生した出来事の詳細を記録したデータのこと。登録や変更、削除等の手続き処理で蓄積されていくもの。（例：受注データ・履歴データなど）
インポート	他のアプリケーションで作成した形式の異なるデータやファイル等を変換して読み込むこと。
結合	複数のSASデータセットを決められたルールに従って統合すること。内部結合（結合する両方のテーブルで一致する行のみ含む）、左外部結合・右外部結合（片方の指定したテーブルのすべての行を含む）などがある。
連続変数	値として表すことができ、連続的な値をとる変数のこと（例：長さ、時間、温度など）。
カテゴリ変数	1つ1つのデータを区別・分類するために用いられている変数のこと。それぞれのデータに大小関係や優劣はなく、単純にデータを分類するために用いられる。
オブザベーション	行、レコードのこと。
正規分布	確率分布の一種で、データの分布が平均値を頂点とした左右対称の山形で表示されるもの。平均±標準偏差の範囲に全体の約68%、平均±標準偏差×2の範囲に約95%、平均±標準偏差×3の範囲に約99%が含まれる等の特長がある。
演算子	計算の内容を表す記号のこと。算術演算子（「+」「−」「*」「/」など）、比較演算子（「=」「<」「>=」など）、論理演算子（「AND」「OR」「NOT」）などがある。
ラベル	変数より細かい説明を記述する目的として使われるもの。変数名とは別にラベルを指定することができ、日本語も使用可能。
出力形式	データの表示形式（データフォーマット）のこと。入力されたデータはそのままで表示形式を変えるもの。
入力形式	データを読み込む際の形式のこと。

用語集

用語	説明
クラスター	英語で「集団」「群れ」のことで、似た性質のものが集まっている様子を表す。クラスター分析とは、異なる性質のものが混ざり合った集団から、互いに似た性質を持つものを集め、クラスターを作る方法。
K-means法	予めいくつのクラスターに分けるかを決め、決めた数の塊にサンプルを分割する非階層クラスター分析の代表的手法。階層クラスター分析と違い、サンプル数が大きいビッグデータを分析する際に適している。
多変量解析	多数の変数間の相互の関係性をとらえるために使われる統計的手法の総称。重回帰分析、判別分析、因子分析、クラスター分析など多岐に渡る分析手法がある。
信頼度（確信度）	前提（Aを買う）が起きた場合に結果（Bを買う）が起きる割合のこと。前提と結果の相関、関連の強さを表す。
サポート（支持度）	前提（Aを買う）と結果（Bを買う）が同時に起こる場合が全トランザクションに占める割合のこと。併売する顧客が全体のどの程度の割合なのかを表す。
期待信頼度	全てのデータの中で結果（Bを買う）の割合。B単独の人気を判断する。
リフト値	前提（Aを買う）が起きた場合に結果（Bを買う）が起きる割合は、全てのデータの中で結果（Bを買う）の割合よりどれだけ多いかを倍率で示したもの。リフト値が低ければ、商品Bは単独（の理由）で売れており、商品Aの商品との関連性よりも商品B特有の理由で売れていると考えられる。
WORKライブラリ	一時データライブラリのこと。WORKライブラリに保存したSASデータセットは、SASを一旦終了すると消去される。SAS終了後も保存しておきたいデータセットは、WORK以外のライブラリ（永久データライブラリ）に保存する必要がある。
要約統計量	基本的なデータ特性を表す統計値。平均値や最大、最小値、標準偏差などがある。基本統計量とも呼ばれる。

用語集

用語	説明
ハンドリング	取り扱い、処理、操作、対処、対応などの意味を持つ英単語。ソフトウェアやプログラミングなどの分野で、特定の状況や対象について、対応する処理を行うことをハンドリングという。
クレンジング	データ分析しやすいデータに整形すること。
SASデータセット	SAS固有のいずれかのファイル形式で内容が格納されたファイル。SASデータセットには次の2種類があります。SASデータファイルとSASデータビューです。SASデータファイルは、データ値に加えて、そのデータに関連付けられているディスクリプタ情報を含みます。SASデータビューには、ディスクリプタ情報と、他のSASデータセットまたはソフトウェアベンダのファイル形式で格納されたファイルからデータ値を取り出すために必要となるその他の情報のみが含まれます。
SASライブラリ	SASデータセットが集まったデータの貯蔵庫のようなものです。
ライブラリ参照名	SASライブラリに一時的に関連付けられる名前。SASファイルの完全名は、ピリオドで区切られた2つの語から構成されます。最初の語はライブラリ参照名であり、これはライブラリを表します。2番目の語は、特定のSASファイルの名前になります。たとえば、VLIB.NEWBDAYの場合、ライブラリ参照名VLIBは、ファイルNEWBDAYが格納されているライブラリを表しています。ライブラリ参照名を割り当てるには、LIBNAMEステートメントを使用するか、またはオペレーティングシステムのコマンドを使用します。
変数	SASデータセットまたはSASデータビュー内の列。各変数のデータ値は、すべてのオブザベーションの単一の特性を表します。各SAS変数は、名前、データタイプ(文字または数値)、長さ、出力形式、入力形式、ラベルという属性を持ちます。
欠損値	無回答や非該当など集計から除去する値のことです。
外れ値	統計において他の値から大きく外れた値である。 測定ミス・記録ミス等に起因する異常値とは概念的には異なるが、実用上は区別できないこともある。
解析	テキストを成分語、フレーズ、マルチワード語、句読点、およびその他のタイプの情報に分割する目的でテキストを分析すること。
モデル	入力から出力を計算する公式またはアルゴリズムです。データマイニングモデルには、入力変数が与えられた場合、ターゲット変数の条件付き分布に関する情報が含まれています。
精度	精密さの度合。

用語集

用語	説明
相関係数	2つの確率変数間の相関（類似性の度合い）を示す統計学的な指標のこと。Aの値とそれに対応するBの値をグラフ化した場合に、右上がりの直線となるものを「正の相関」、右下がりの直線となるものを「負の相関」と呼ぶ。数値は−1から+1の間となり、数値が0に近づくほど相関関係が希薄になる。数値が0の場合は「相関がない」、つまりAの数値が変化してもBの数値に影響がないということになる。ただし、相関係数は順序尺度であり、間隔尺度ではない。このため相関係数を単純に比較することは意味がない。 たとえば自動車では、搭載するガソリンの量が多いほど走行可能距離が伸びる「正の相関」、走行距離が伸びるほどガソリンは減るので「負の相関」となる。なお、基本的に単位はつけない。 相関係数を把握することで、Aの数値によってBの数値を予測することができる。
回帰分析	結果の数値と、その要因の数値から、それぞれの関係を予測する分析手法のこと。比較的容易な分析手法で、ひとつの要因から結果を予測する「単回帰分析」と、複数の要因からひとつの結果を求める「重回帰分析」がある。 たとえば、既存顧客に新たなダイレクトメールを送る際、過去のダイレクトメールへの反応履歴（結果）と、送付対象者の年齢、収入、住居地域、購入金額、購入履歴など複数の顧客属性（要因）との因果関係を重回帰分析することで今回発送するダイレクトメールの反応数を推定することができます。なお、結果の数値は「目的変数」や「従属変数」と呼ばれ、要因となる数値は「説明変数」と呼ばれる。
ロジスティック回帰分析	見込み顧客が製品を買ってくれるかどうか、キャンペーンに反応するかどうか、など 将来のYES/NO を予測するときに使える手法。
ROC (Receiver Operating Characteristic)曲線	二値変数（YES/NO　例. 実際に購買した／しなかった）と連続変数（例. 購買可能性%予測値）との関係の強さを評価する方法。例えば連続変数のあるカットオフの値を設定し、それ以上をYES=購買する、それ未満をNO=購買しない、と予測した場合の陽性率（予測=YES、実際=YES）、偽陽性率（予測=YES、実際=NO）を取得する。カットオフの値を動かすことで陽性率、偽陽性率がどのように変化するかをグラフ上に曲線として表現し、その曲線で連続変数と二値変数の関係の強さを評価する。縦軸に陽性率を、横軸に偽陽性率をとった場合に、曲線の左上方向へのふくらみが大きいほど、変数間の関係が強いと判断できる。

用語集

用語	説明
ツリー分析（決定木）	観察対象データの集団を、従属変数（結果：購買の有無、解約の有無など）に対し最も効率よく分類できる独立変数（原因）によって次々と分割し、木の枝のように分岐・整理していく分析手法。データの集団を効率よく分類・整理し、ルール抽出・生成や予測モデル構築などに利用される。たとえば、商品を購入する／しないに最も強く影響する要素を探る際などに用いられる。 マーケティング分野では、最も高い反応が期待できる顧客グループに対して販促計画を練るなどの目的で使われる。 ツリー分析では、視覚的に分析結果を把握できるとともに計算方法が比較的簡単で、モデル作成しやすいことが特徴とされる。
因子	ある結果をひき起こすもとになる要素。現象の要因を構成している作用素または力。
最尤法（さいゆうほう）	統計的推定の際、実際に得られた標本があるとき、それが得られる確率が最大になるような母数の値をその推定値とする手法。