第 7 講 回帰分析と分類モデル

１．データで未来・未知を予測する

データからパターンを抽出し、未来や未知のデータを予測する手法として、回帰分析と分類モデルが挙げられます。これらは、データサイエンスにおいて最も基本的かつ重要な予測技術です。回帰分析と分類モデルを理解することで、ビジネスにおける売上予測から迷惑メールの判別に至るまで、身近な課題をデータによって解決するための大きな一歩を踏み出すことが可能になります。

まず、回帰分析と分類の違いについて説明します。回帰分析は数値（連続値）を予測する手法で、分類はデータをカテゴリーに分ける手法です。目的が「数値」か「ラベル」かが主な違いです。下記に詳細な違いを示します。

項目	回帰分析（Regression）	分類（Classification）
目的	数値を予測する	カテゴリー（ラベル）を予測する
出力の例	価格、気温、売上などの連続値	スパム/非スパム、合格/不合格などのクラス
モデルの例	線形回帰、重回帰、リッジ回帰など	ロジスティック回帰、決定木、SVM など
評価指標	MSE（平均二乗誤差)、MAE、R^2など	Accuracy、Precision、Recall、F1 スコアなど
数学的性質	出力は連続値（実数）	出力は離散値（カテゴリー）
使う場面	売上予測、気温予測、年収予測など	顧客の属性分類、病気の診断、画像認識など

２．回帰分析

２．１　回帰の語源

「回帰」という言葉は、もともと「一周して元に戻ること」という意味があります。そのため、たとえば後ほど紹介する単回帰分析の式　$y = \beta 0 + \beta 1x$　を見ても、「どこが“元に戻る”のだろう？」と不思議に思われるかもしれません。回帰分析という言葉が使われるようになったのは、ある統計学者が親子の身長を分析し、極端な身長の子どもは平均身長に近づく傾向があることを発見しました。必ずしも遺伝せず、先祖返りのように平均値に戻っていく現象は「平均への収束→回帰」と呼ばれるようになりました。

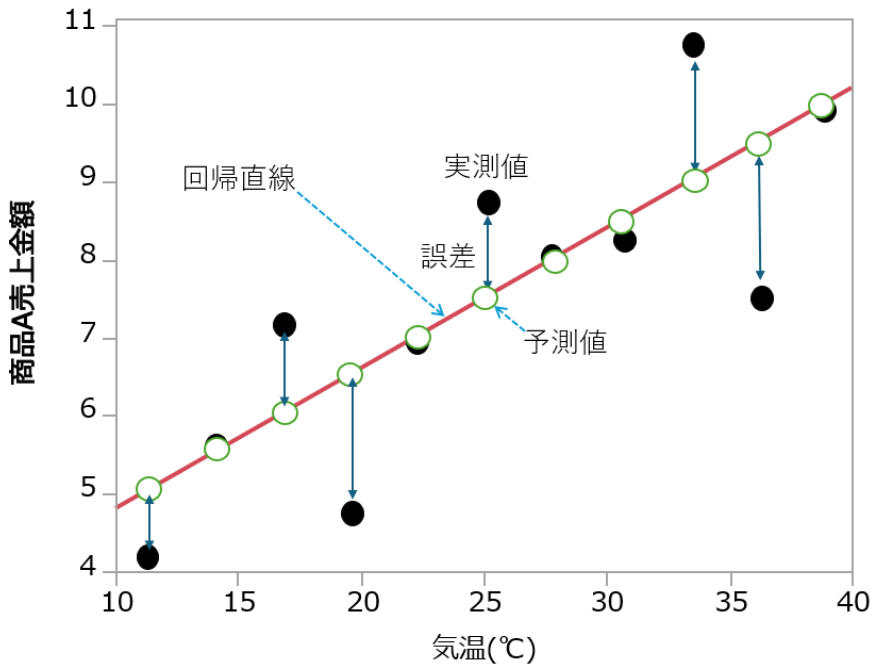
２．２　単回帰分析

最もシンプルな回帰分析の単回帰分析の特徴としては、説明変数x が 1 つの場合で、次の式が回帰式（以下、モデル）となります。

$$y = \beta 0 + \beta 1x$$

$\beta 0$ は方程式の切片とも呼び、$\beta 1$ は回帰係数です。

この$\beta 0$ と $\beta 1$ を推計する方法として使われているのが最小二乗法です。最小二乗法は、観測値と予測値の差である「残差」の二乗和が最小になるように$\beta 0$ と $\beta 1$ を推計する方法です。例えば、図の黒丸（●）は実際に測定された実測値で、その中に描かれている近似線（赤）は実測値との残差を最も小さくするために推計された予測値である白丸（○）を繋げた近似線、つまり回帰直線です。残差は実測値－予測値となります。

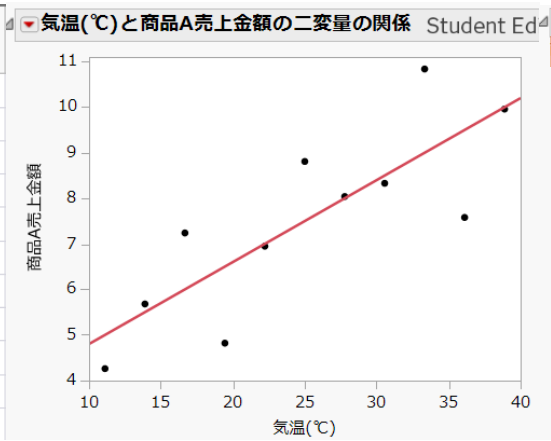


$\beta 0=$（yの平均値）$-\beta 1\times$（xの平均値）　$\beta 1=S_{xy}/S_{xx}$

S_{xy}: xとyの偏差の積の総和　S_{xx}: xの偏差平方和　で計算することができます。

例として、下図（左）のデータ表において、「気温（℃）」から「商品 A の売上金額（百万円）」を予測します。データよりアプリケーションソフト（SAS 社の JMP®）を使って計算した結果を下図に示します。このデータの場合、モデルは、商品 A 売上金額 $=3.00+0.18\times$気温　となります。つまり、気温が 1℃上昇すると、売上金額が 18 万円増えます。

気温(℃)	商品A売上金額
27.8	8.04
22.2	6.95
36.1	7.58
25	8.81
30.6	8.33
38.9	9.96
16.7	7.24
11.1	4.26
33.3	10.84
19.4	4.82
13.9	5.68



直線のあてはめ

商品A売上金額 = 3.0000909 + 0.1800183*気温(℃)

あてはめの要約

R2乗	0.666542
自由度調整R2乗	0.629492
誤差の標準偏差(RMSE)	1.236603
Yの平均	7.500909
オブザベーション(または重みの合計)	11

分散分析

要因	自由度	平方和	平均平方	F値
モデル	1	27.510001	27.5100	17.9899
誤差	9	13.762690	1.5292	p値(Prob>F)
全体(修正済み)	10	41.272691		0.0022*

パラメータ推定値

項	推定値	標準誤差	t値	p値(Prob>|t|)
切片	3.0000909	1.124747	2.67	0.0257*
気温(℃)	0.1800183	0.042443	4.24	0.0022*

尚、回帰分析においては、分散分析を行って、モデルが統計的に有意かどうか、残差の検定を行います。

2．3　重回帰分析

重回帰分析の特徴は複数の説明変数があることです。例えば、「面積」「築年数」「駅からの距離」などから「住宅価格」を予測したい場合、

モデルは、　価格 $=\beta 0+\beta 1\times$面積 $+\beta 2\times$築年数 $+\beta 3\times$距離　となります。このように、複数の要因が絡む現実的な予測に向いています。

２．４　多項式回帰

多項式回帰の特徴は、説明変数のべき乗を使って、曲線的な関係を表現します。例えば、面積のべき乗に比例して価格が変化する場合、
モデルは、　価格 ＝ $\beta 0 + \beta 1 \times$面積$+ \beta 2 \times$ 面積^2　となります。直線では表せない複雑な関係に対応できます。

２．５　リッジ回帰・ラッソ回帰

この２つの手法は、学習したデータに対しての精度は高いものの未知のデータに対しては同様の精度が出せない問題である「過学習」が起こりにくいように工夫されています。

２．６　線形以外の回帰

決定木回帰、ランダムフォレスト回帰、サポートベクター回帰など、様々な回帰分析手法があります。線形では表せない複雑な関係に対応でき、精度は高いのですが仕組みは複雑です。

３．分類モデルとは

３．１　分類の種類

分類は、与えられたデータを事前に定義されたカテゴリーに分けることを指します。例えば、メールが「スパム」か「スパムでないか」を判別するスパムフィルターや、画像が「犬」か「猫」かを判別するモデルが分類の例です。代表的な方法にはロジスティック回帰と決定木があります。ロジスティック回帰は、確率を使って分類します（例：スパムの確率が 70%ならスパムと判定）。決定木は、特徴量に基づいて「分岐」を繰り返し、最終的に分類を決める木のような構造のモデルです。直感的でわかりやすく、特徴量の重要度も見やすいのが魅力です。その他、サポートベクターマシン、決定木、ランダムフォレストなど、多くの手法があります。

３．２　ロジスティクス回帰

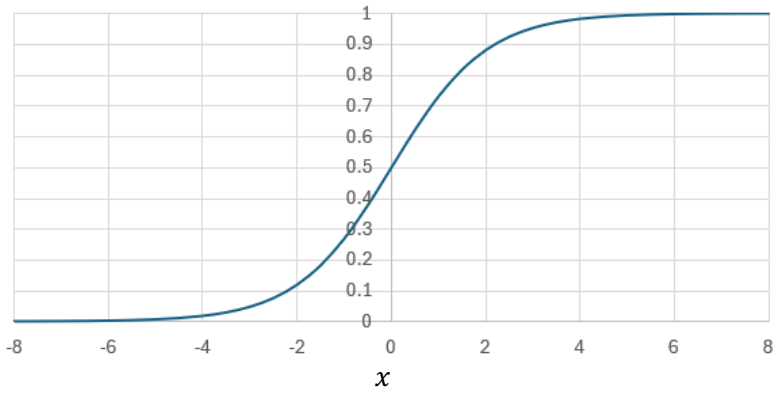
分類問題において、対象が「はい／いいえ」や「スパム／非スパム」といった二値のいずれかに属するかを判定する必要がある場面は多く見られます。こうした問題に対して有効な手法の一つがロジスティック回帰です。ロジスティック回帰は、線形回帰と似ていますが、ロジスティック回帰では、入力された特徴量の線形結合に対してシグモイド関数（ロジスティック関数）を適用することで、出力を 0 から 1 の範囲に収め、確率として解釈可能な値を得ることができます。

シグモイド関数は以下の式で表されます。

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

この関数の特徴は、入力値xが大きくなるにつれて出力が 1 に近づき、逆に小さくなると 0 に近づくという S 字型の滑らかな曲線を描く点にあります。下図の横軸がxで、縦軸が$\sigma(x)$です。xが 0 のとき、$\sigma(x)$は 0.5 と 0 と 1 のちょうど真ん中になります。これにより、モデルの出力を「あるクラスに属する確率」として自然に解釈できるようになります。

$\sigma(x)$



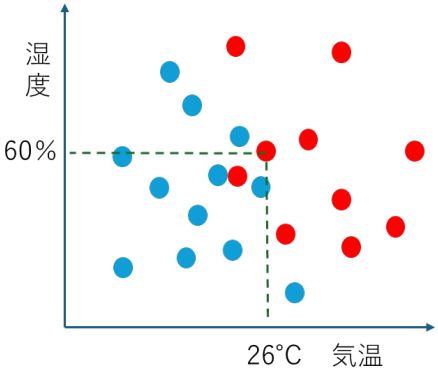
3．3　決定木

ロジスティック回帰が数学的な確率モデルを構築するのに対し、決定木はまったく異なる発想で分類問題に取り組みます。そのアプローチは非常に直感的で、人間が意思決定を行うプロセスに似ているため、結果の解釈が容易であるという点に大きな価値があります。

決定木の本質は、「もし○○ならこちら、そうでなければあちら」といったシンプルな条件分岐を、木の枝分かれのように繰り返すことで、データを分類していく手法です。

具体例として、A 君がある日に「暑いと感じるかどうか」を、その日の気温と湿度から予測するケースを考えてみましょう。A 君が「暑いと感じた日（赤点）」と「そうでない日（青点）」のデータが、気温と湿度を軸にした下図のグラフ上にプロットされています。

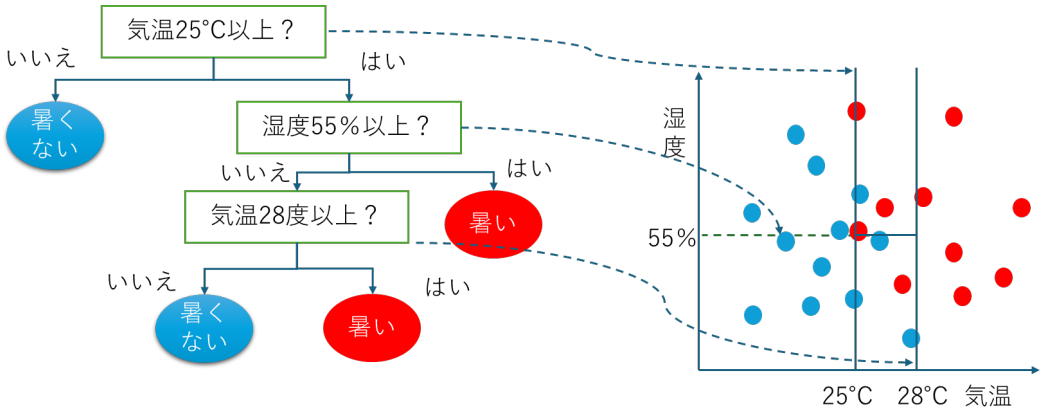
たとえば、気温が 26℃、湿度が 60%の日に、A 君は暑いと感じていました。この時点では、赤点と青点が混在しており、明確な分類パターンはまだ見えていません。



決定木は、混在したデータを最も効率的に分割できる質問（ルール）を自動的に探索し、見つけ出します。たとえば、まず全体に対して「気温は 25 度より高いか？」という質問が最も効果的であると判断し、データを 2 つのグループに分割します。次に、それぞれのグループに対して、さらに最適な分割ルール（たとえば、片方のグループでは「湿度は 55%より高いか？」）を再帰的に探索していきます。

このようなルールを順に適用することで、もともと混在していたデータ群に明確な境界線が引かれ、各領域は「暑い日」と判断される可能性が高いエリアと、「そうでない日」と判断されるエリアに分類されます。なお、一部のデータ点は正しく分類されていません。この点については、次章「モデルの評価」にて詳しく説明します。

これらの分割ルールは、下図（左）に示すように、気温と湿度に基づいて「暑いと感じるかどうか」を判断するツリー構造として表現できます。また、この判断に基づき、下図（右）のグラフには分類境界線を加えています。



このように決定木は、複雑なデータの中から人間にも理解しやすい明確なルールを自動で構築し分類を行います。

4．モデルの評価

4．1　概要

前述の回帰分析や分類モデルの性能を評価し改善するプロセスは、データ分析および予測モデル構築において重要です。例えば、正解率が 90%だからといって、モデルが優れているとは限りません。100 件のうち 90 件が『正常』で 10 件が『異常』なデータがあるとします。もし、すべてを『正常』と予測すれば、正解率は 90%になりますが、異常を 1 件も検出できていません。これでは本来の目的を果たしていないのです。

特に分類モデルの評価においては、単なる正解率だけでなく、より詳細な性能を把握するための複数の指標が存在します。どの指標を重視するかは、解決したい課題の性質によって異なります。例えば、医療診断では病気の見逃し（偽陰性）を防ぐことが、スパムメール判定では正常なメールの誤分類（偽陽性）を防ぐことが優先されます。

4．2　混同行列

混同行列は、分類モデルの予測結果を整理するための表で、以下は「2 クラス分類」の例です。

	実際が陽性（Positive）	実際が陰性（Negative）
判定が陽性	真陽性（TP: True Positive） 正しく陽性と予測 [スパムを正しくスパムと判定]	偽陽性（FP: False Positive） 誤って陽性と予測 [正当メールを誤ってスパムと判定]
判定が陰性	偽陰性（FN: False Negative） 誤って陰性と予測 [スパムを誤って正当メールと判定]	真陰性（TN: True Negative） 正しく陰性と予測 [正当メールを正しく正当メールと判定]

4．3　評価指標の種類

主に、下表の 4 つが評価の指標としてよく利用されます。

① 正解率（Accuracy）：全体の中で、正しく予測できた割合
　ただし、クラスの偏りがある場合（例：陽性が少ない）には注意が必要です。
② 精度（Precision）：「陽性と予測したもの」の中で、実際に陽性だった割合
③ 再現率（Recall）：「実際に陽性だったもの」の中で、正しく陽性と予測できた割合
④ F1 スコア：精度と再現率のバランスを取った指標
　精度と再現率の両方が高いときに高くなります。F1 スコアは 0～1 の間 の値を取り、1 に近いほど良いモデルです。特に「不均衡データ（陽性が少ない）」では、正解率よりも F1 スコアが信頼できます。

４．４　正しい指標の選び方

４．３に４つの評価指標を紹介しましたが、どのように使い分けすべきか、下表に示します。

指標	計算式	重視する点	ユースケース例
① 正解率 Accuracy	$\frac{TP + TN}{TP + FP + FN + TN}$	全体の正解率	クラスが均等
② 精度 Precision	$\frac{TP}{TP + FP}$	誤検出（FP）を減らす	スパムメール検出
③ 再現率 Recall	$\frac{TP}{TP + FN}$	見逃し（FN）を減らす	病気検出
④ F1 スコア	$\frac{2 \times Precision \times Recall}{Precision + Recall}$	Precision と Recall のバランス	不均衡データ

５．まとめ

・データによる予測には、大きく分けて「回帰（数値予測)」と「分類（カテゴリー分け)」の２種類があります。

・どんな問題にも効く完璧な万能モデルは存在せず、解決したい問題に応じて適切な手法を選ぶ必要があります。

・モデルの性能は「正解率」だけでなく、ビジネス目的（どの間違いが最も致命的か）に合った指標で正しく評価することが最も重要です。

データサイエンスと聞くと難しく感じるかもしれませんが、最も重要な一歩は、技術そのものではなく、皆さんが「解決したい課題は何か？」「最も避けたい間違いは何か？」を自問することです。