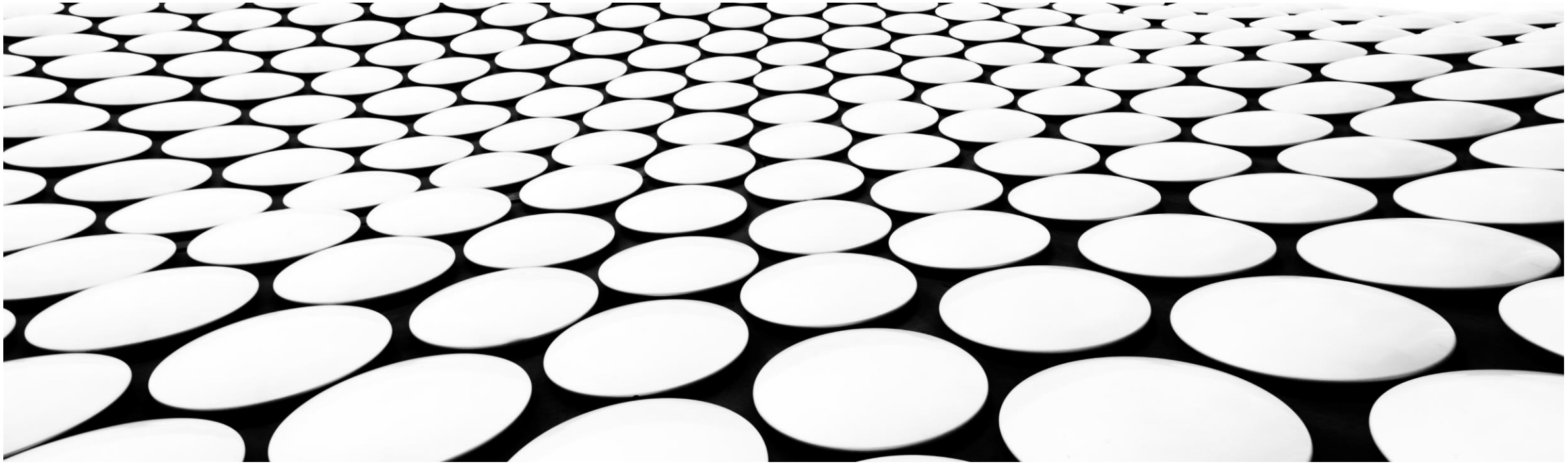
第7講 回帰分析と分類モデル



第7講 回帰分析と分類モデル
目次

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ある
- 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
- 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

回帰分析と分類の違い

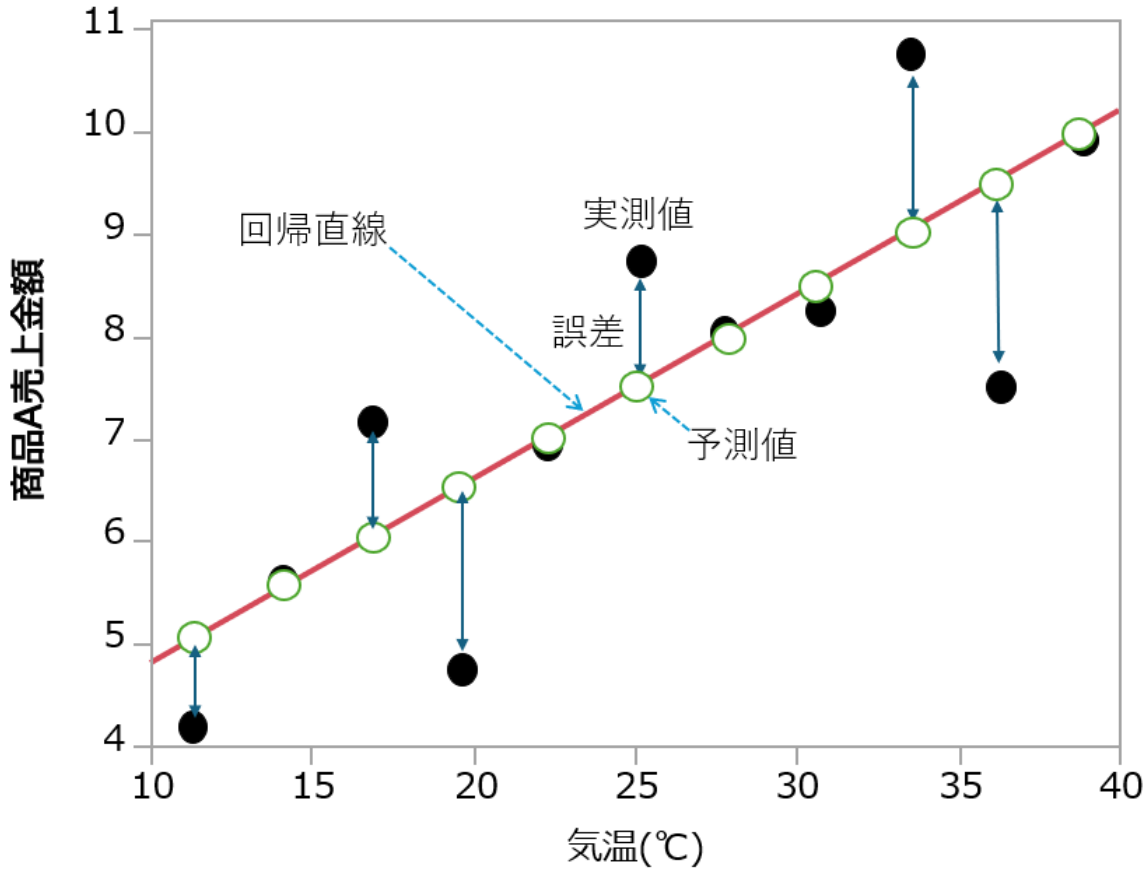
項目	回帰分析（REGRESSION）	分類（CLASSIFICATION）
目的	数値を予測する	カテゴリー（ラベル）を予測する
出力の例	価格、気温、売上などの連続値	スパム/非スパム、合格/不合格などのクラス
モデルの例	線形回帰、重回帰、リッジ回帰など	ロジスティック回帰、決定木、SVMなど
評価指標	MSE（平均二乗誤差）、MAE、R^2など	ACCURACY、PRECISION、RECALL、F1スコアなど
数学的性質	出力は連続値（実数）	出力は離散値（カテゴリー）
使う場面	売上予測、気温予測、年収予測など	顧客の属性分類、病気の診断、画像認識など

２．回帰分析

1. 回帰の語源
2. 単回帰分析
3. 重回帰分析
4. 多項式回帰
5. リッジ回帰・ラッソ回帰
6. 線形以外の回帰

２．回帰分析
単回帰分析　最小二乗法

- 回帰直線　$y = \beta 0 + \beta 1 x$
- 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β0とβ1を推計する方法
- 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

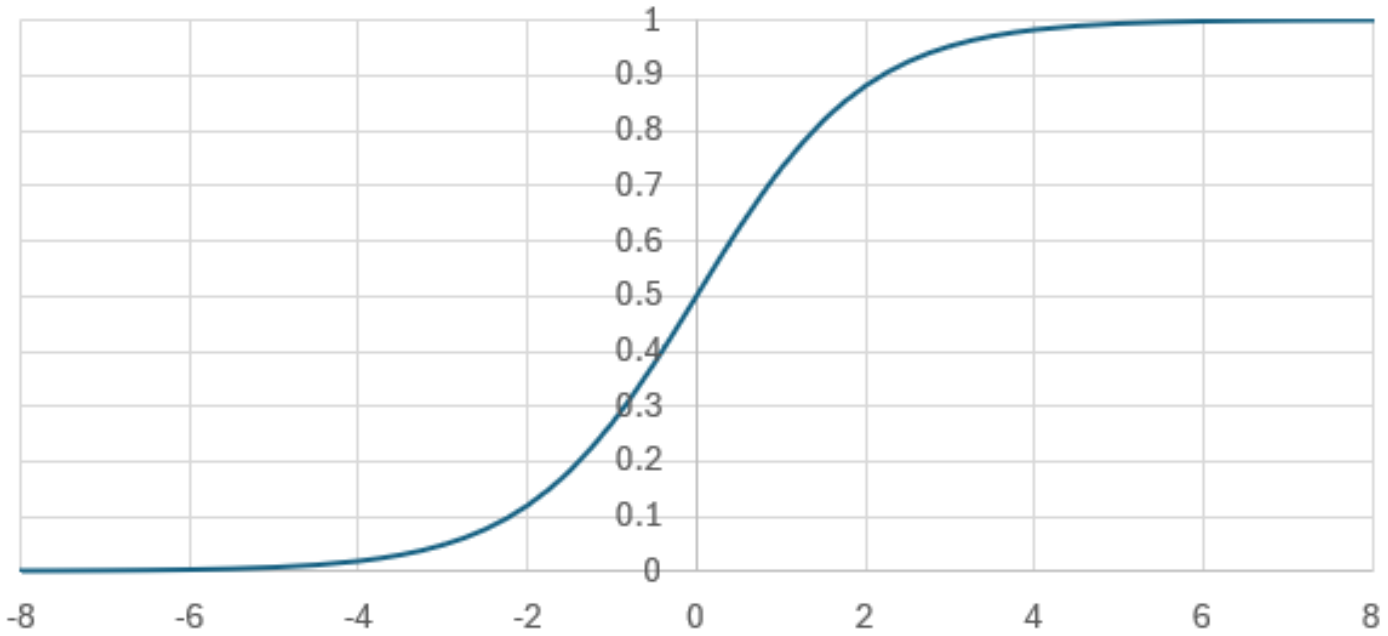


３．分類モデルとは
分類の種類

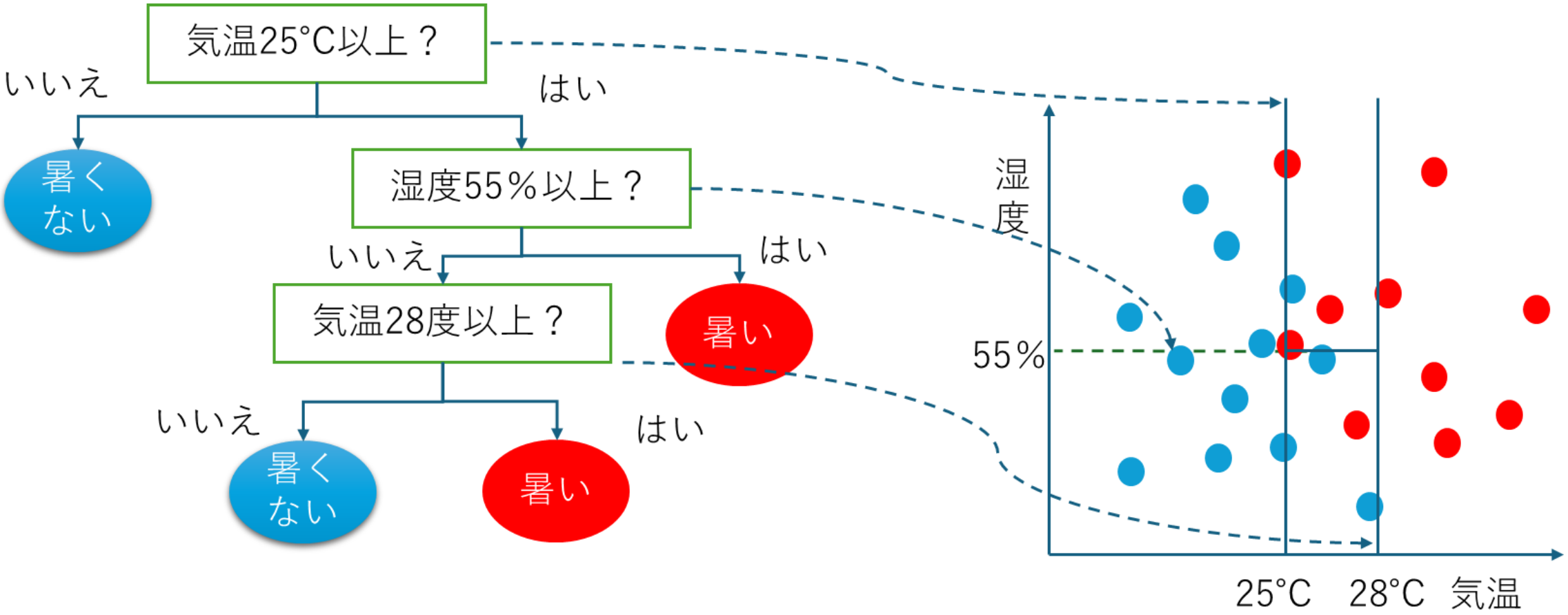
- 分類は与えられたデータを事前に定義されたカテゴリーに分けること
- 分類の種類として、
  ロジスティック回帰、決定木、サポートベクターマシン、ランダムフォレストなどがある

３．分類モデルとは
ロジスティクス回帰・シグモイド関数

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



３．分類モデルとは
決定木の事例



４．モデルの評価
混同行列

	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．モデルの評価
正しい指標の選び方

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
①精度 PRECISION	$\frac{TP}{TP + FP}$	誤検出（FP）を減らす	スパムメール検出
①再現率 RECALL	$\frac{TP}{TP + FN}$	見逃し（FN）を減らす	病気検出
①F1 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種類がある
- 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る
- 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