クロスバリデーション法を用いた決定木モデル作成

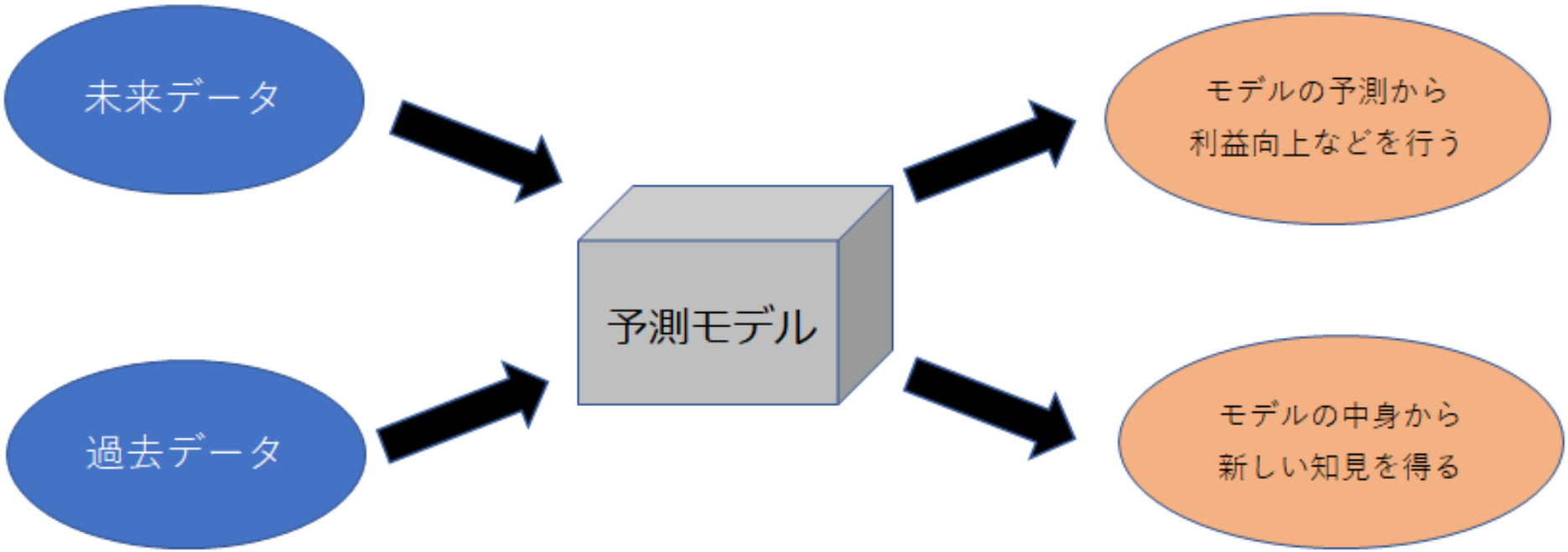
菅原 有一

芝浦工業大学 数理科学研究会

2019/11/01

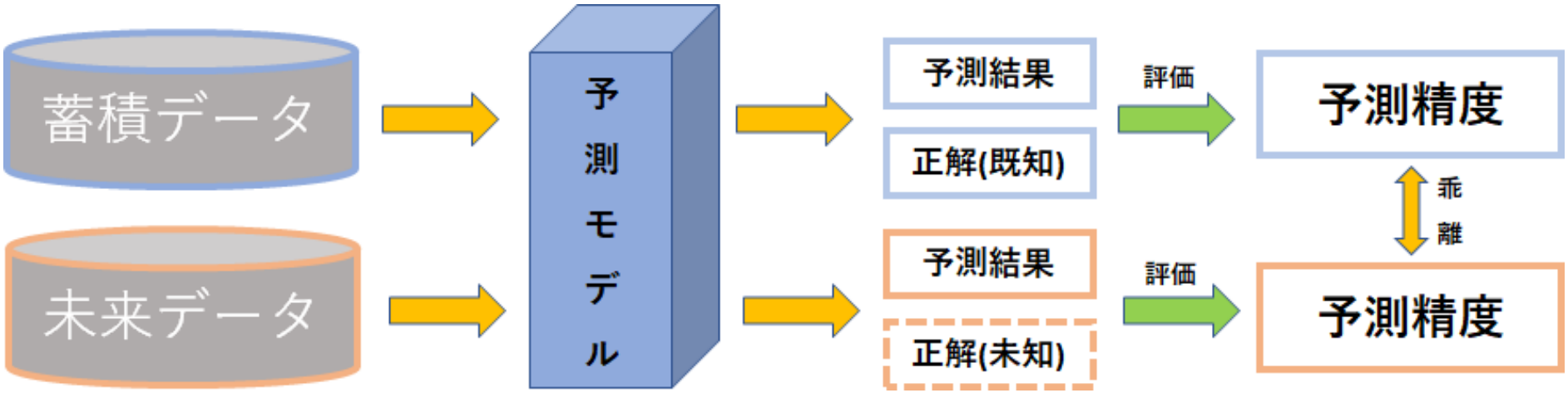
データ分析

データ分析とは.....
過去の蓄積データから未来を予測するためのモデルを作成すること.
またそのモデルを用いた予測からより良い結果になるような分析をすること.



モデルの精度評価

(モデルの精度) = (蓄積データの予測精度) ・ (乖離幅)
として考える必要がある.

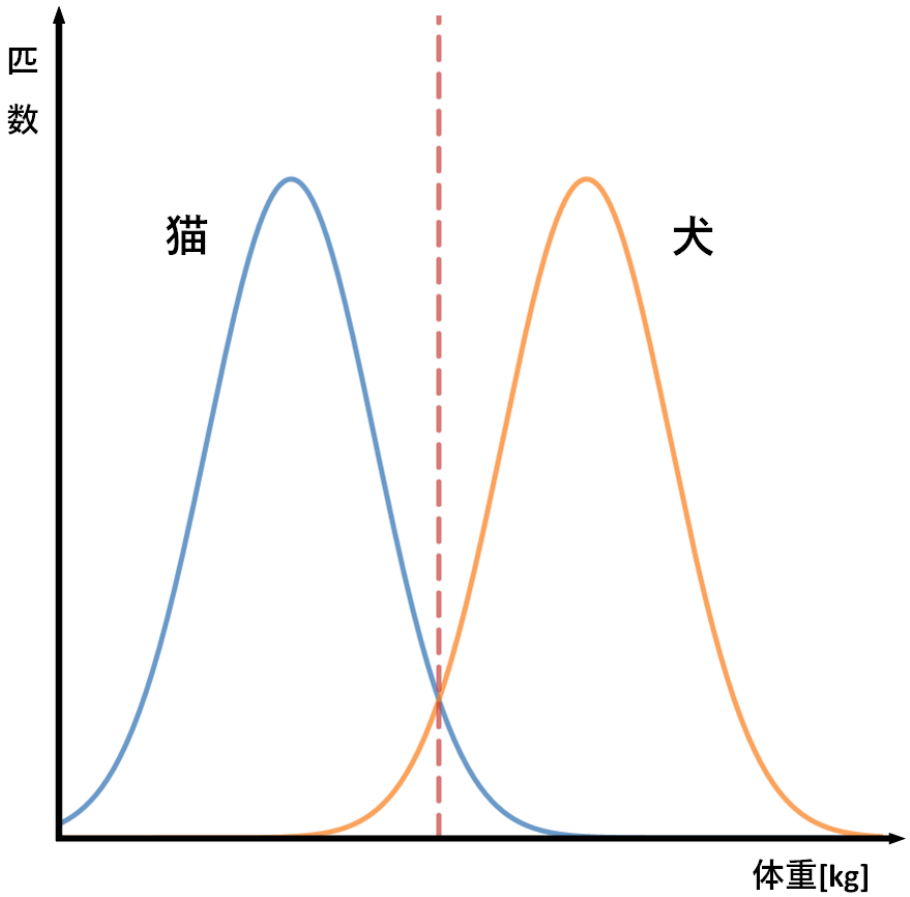


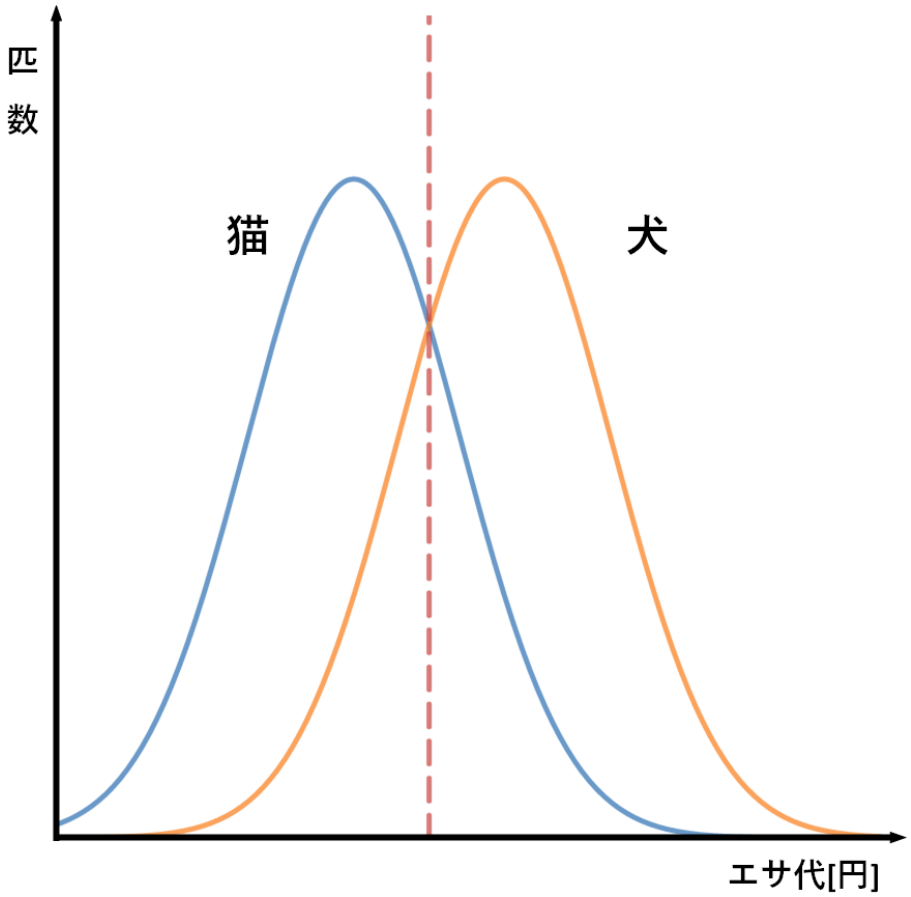
予測モデルの評価指標

評価指標	予測対象	値域
AUC	2 群分類, 例:医療診断, 契約	$0 \sim 1$
Log Loss	3 群以上の分類, 例:運転行動	$0 \sim \infty$
RMSE	個数, 例:売上個数, 降水量	$0 \sim \infty$
MAE	個数, 例:売上個数, 降水量	$0 \sim \infty$
NDCG	順序, 例:検索結果, 商品推薦	$0 \sim 1$
MAP@n	(2 群分類)^n ・ (順序), 例:購入商品	$0 \sim \sum_{i=1}^{n} 1/k$

AUC の精度評価方法

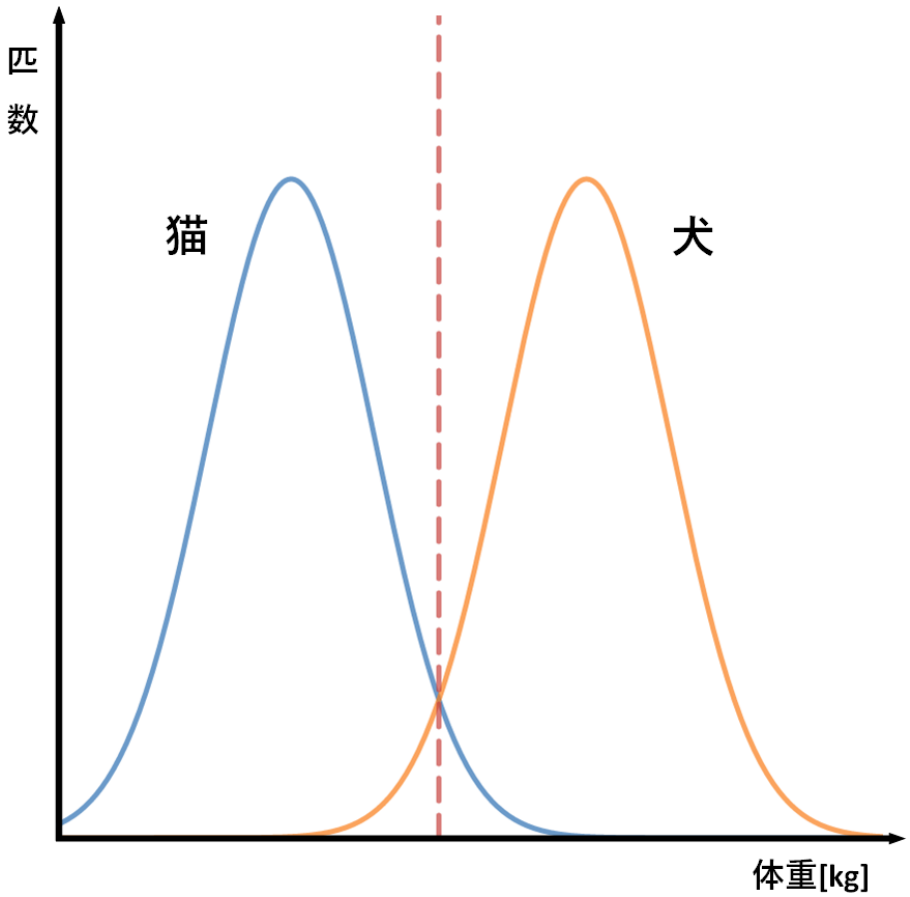
例:犬と猫の識別モデル
(左:体重で識別するモデル, 右:エサ代で識別するモデル)

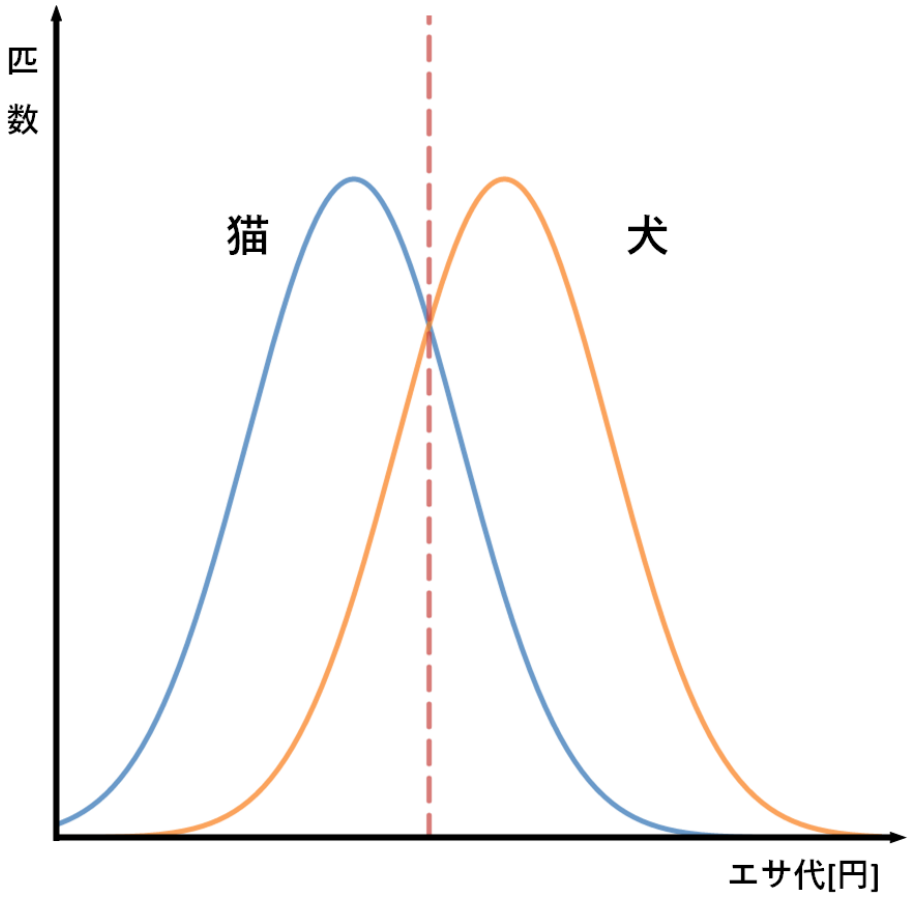




AUC の精度評価方法

例:犬と猫の識別モデル
(左:体重で識別するモデル, 右:エサ代で識別するモデル)





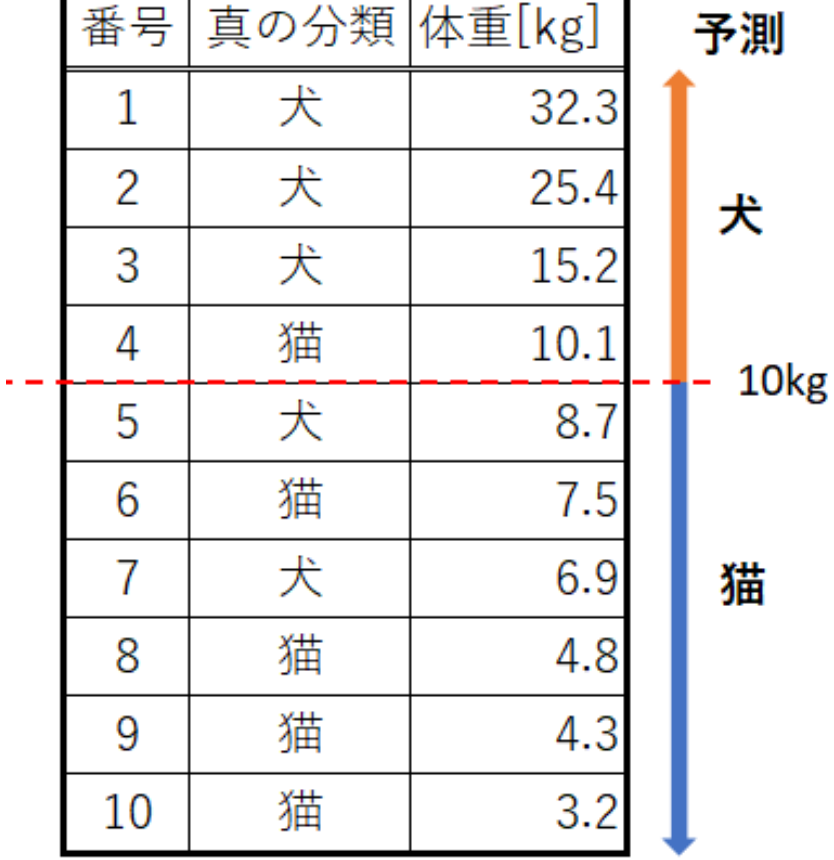
これの重なり具合から制度を評価する！

AUC の精度評価方法

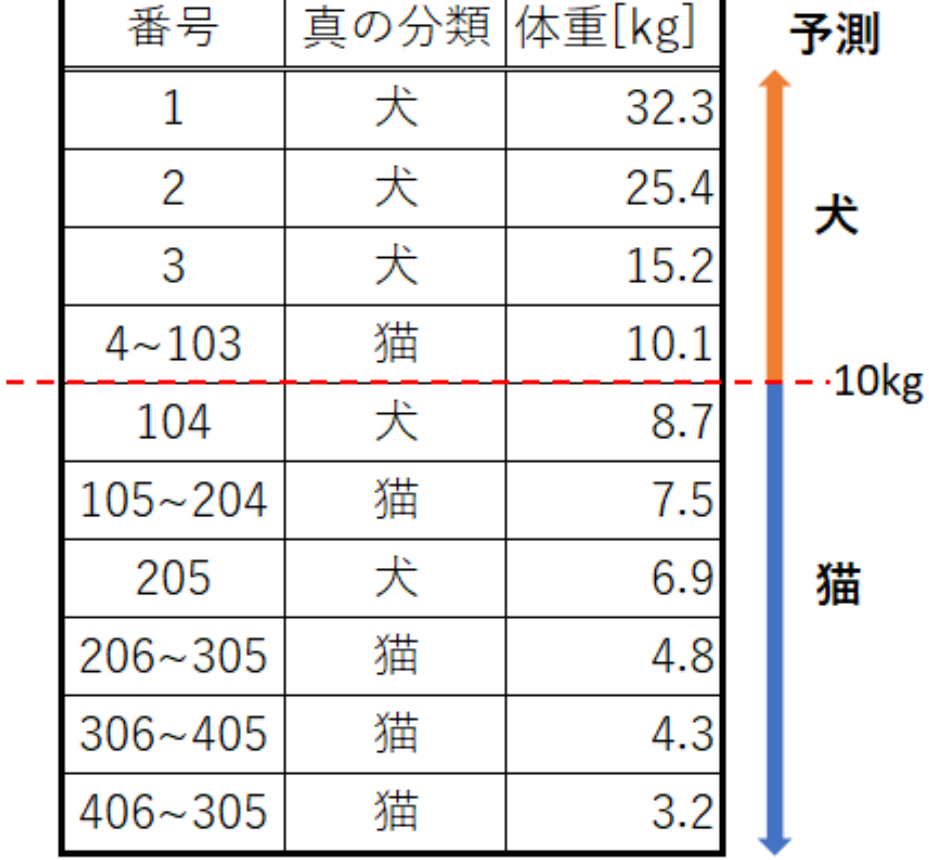
- 誤分類率

誤分類率　＝　(予測対象のうち, 誤って分類した割合)

番号	真の分類	体重[kg]
1	犬	32.3
2	犬	25.4
3	犬	15.2
4	猫	10.1
5	犬	8.7
6	猫	7.5
7	犬	6.9
8	猫	4.8
9	猫	4.3
10	猫	3.2



番号	真の分類	体重[kg]
1	犬	32.3
2	犬	25.4
3	犬	15.2
4~103	猫	10.1
104	犬	8.7
105~204	猫	7.5
205	犬	6.9
206~305	猫	4.8
306~405	猫	4.3
406~305	猫	3.2



AUC の精度評価方法

- 混同行列

混合行列とは, 下のような表のことを言う. この時, 真陽性率, 偽陽性率は次のものを表す.

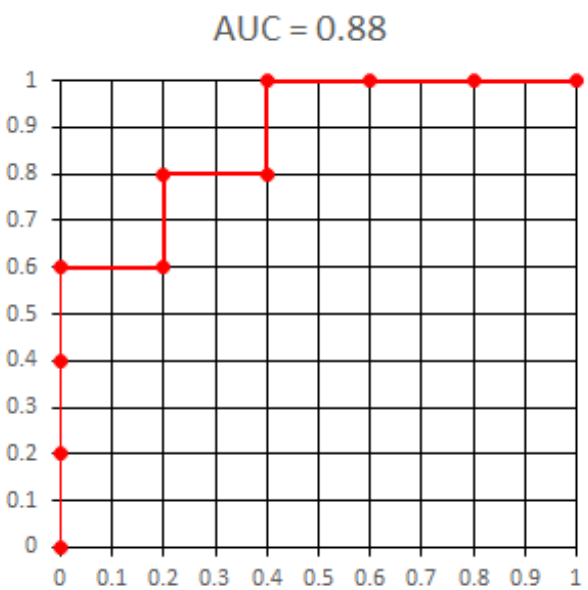
真陽性率 = 犬が犬と正しく分類された割合.
偽陽性率 = 猫なのに犬と誤分類された割合.

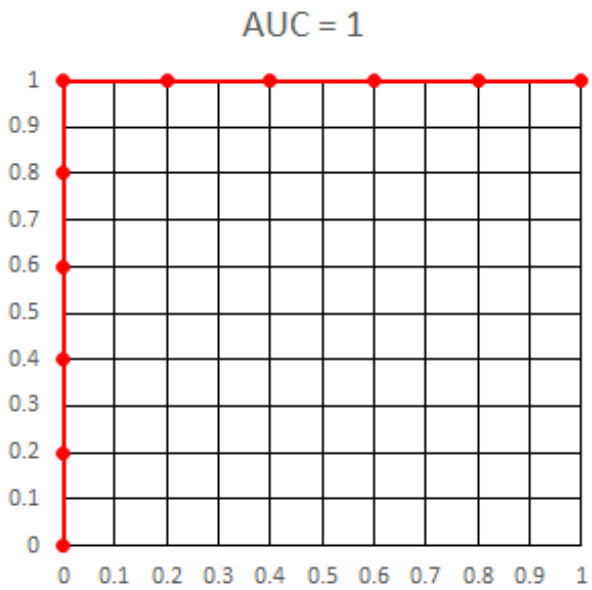
		予測の分類		
		犬	猫	合計
真の分類	犬	True Positeve(TP)	False Negative(FN)	Positive(P)
	猫	False Positeve(FP)	True Negative(TN)	Negative(N)

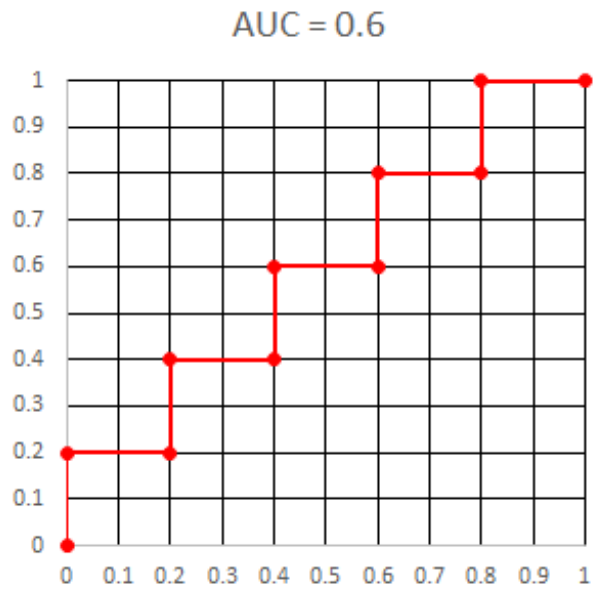
AUC の精度評価方法

- ROC 曲線 (受信者動作特性曲線)

ROC 曲線とは, 偽陽性率と真陽性率の変化をグラフで表したもの.







この ROC 曲線下の面積を AUC 値と定義する.

AUC の精度評価方法

- AUC と誤分類率の比較

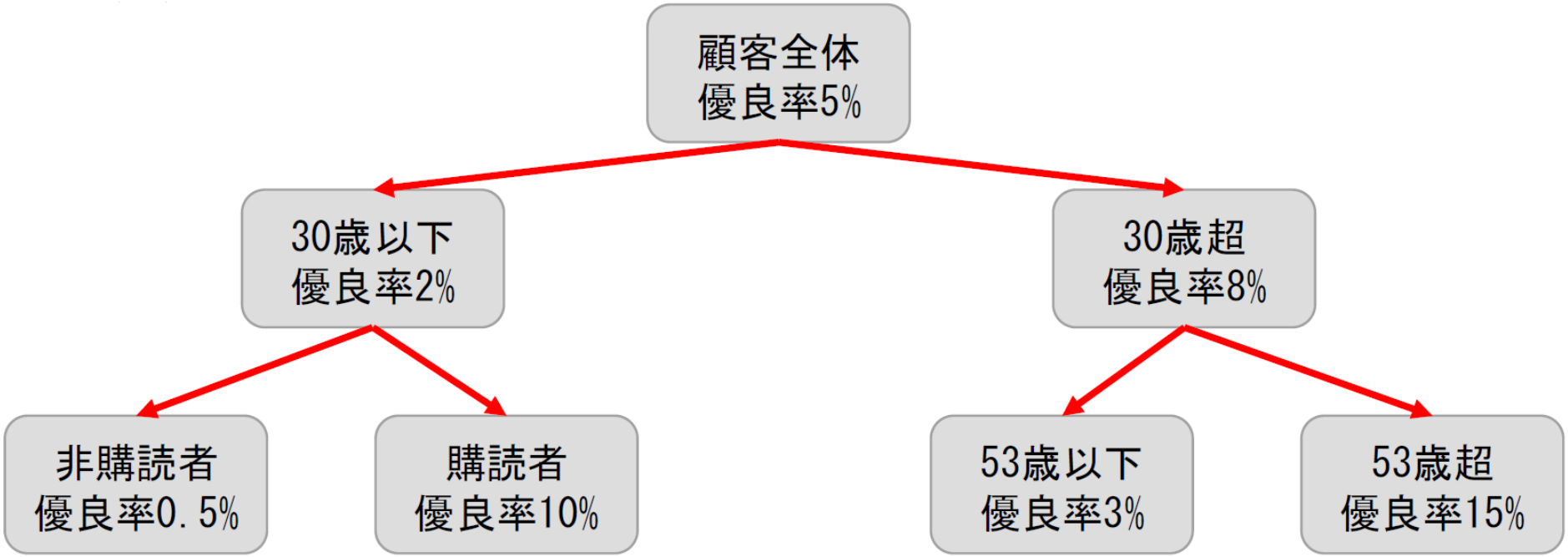
予測データ1		モデル	
		体重モデル	すべて猫と分類するモデル
指標	誤分類率	0.3	0.5
	AUC	0.88	0.5

予測データ2		モデル	
		体重モデル	すべて猫と分類するモデル
指標	誤分類率	0.2	0.01
	AUC	0.88	0.5

AUC のほうが安定して精度評価していることがわかる.

決定木モデルとは

決定木モデルとは下図のような図のこと.



図：決定木モデル

決定木モデルの作られ方

- 分割基準の選び方

データを全説明変数を虱潰しに探索して, 分割スコアのもっともよいもので分割する.

- スコア

分割スコアには Gini と Entropy がある.

$$\text{Gini} : I_G(S) := 1 - p_S^2 - (1 - p_S)^2$$

$$\text{Entropy} : I_E(S) := p_S log(p_s) - (1 - p_S) log(1 - p_S)$$

ただし, p_S はセグメント S の一方のクラス所属確立, 即ち, $p_S = \frac{n_i}{N}$

(ただし, N:トレーニングデータのサンプル数

n_i:クラス i に属するトレーニングデータの数)

決定木モデルの特徴

- 図で表現できて説明がしやすい.
  ⇒ビジネスでの利用 .
- 副次的に交互作用が生まれることがある.
  ⇒単純に重要度の低いデータを抜けばいいというわけではない .
- 複数変数の線形結合には弱い.
  ⇒精度を高めるために自分で新たな説明変数を作成するなどの工夫を必要とする場合もある.

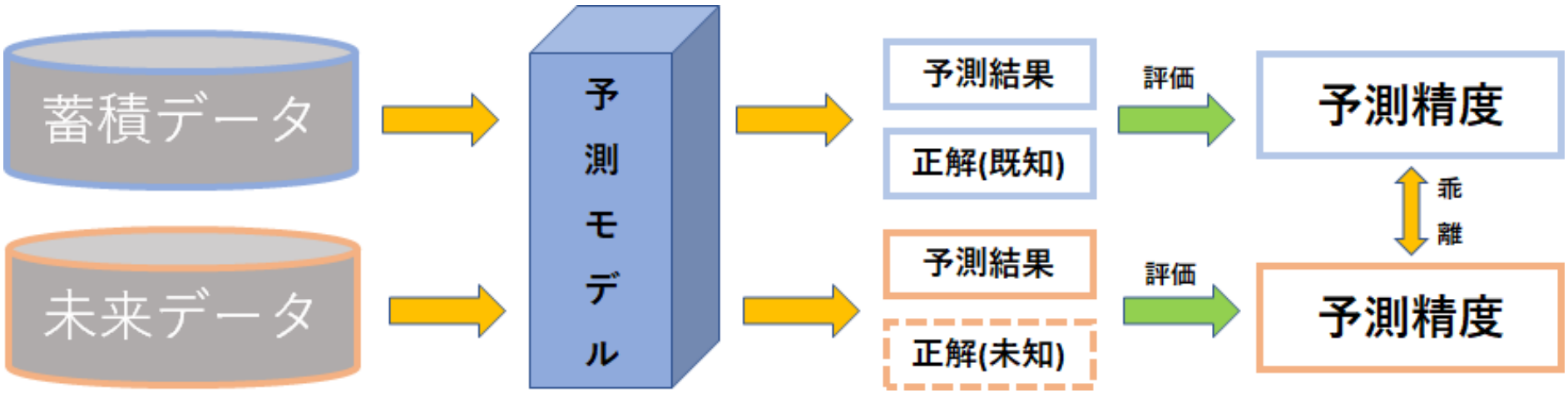
決定木モデルの学習パラメータ

決定木モデルの学習パラメータは次の 3 つがある.

- 木の深さ (mavdepth)
  分割の最大数
- 最小ノードサイズ (minbucket)
  セグメントに含まれるデータ数の最小値
- 枝刈の強さ (cp)
  意味のない分割をしないラインを決める

過学習

蓄積データの精度は高いが未知のデータでの精度が悪く, 2 つの予測精度の乖離が大きいとき, 予測モデルは「過学習 (過剰適合)」しているという.



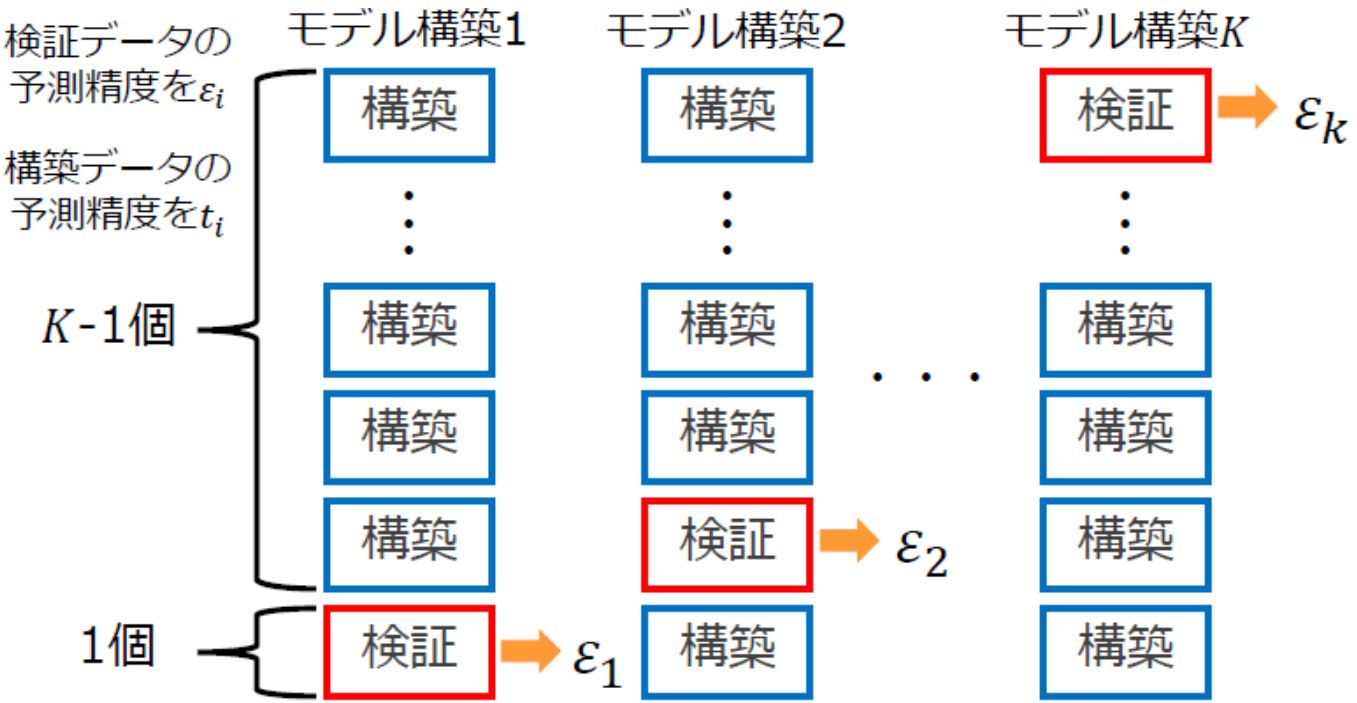
過学習

未知のデータに対する予測精度を上げたい！
⇒最適なモデルの複雑さを見つけたらいい！

過学習を対策する方法としては下記の方法があげられる.

- ホールドアウト法
  データの数が十分に多いときに有効.
- クロスバリデーション法 (K 分割交差検証法)
  今回はこっちを説明する.

クロスバリデーション法とは



検証データへの予測精度

$$\varepsilon = \frac{1}{K}\sum_i \varepsilon_i$$

構築データへの予測精度

$$T = \frac{1}{K}\sum_i t_i$$

この2つの予測精度を比較して過学習を防ぐ!

使用するデータ

ここで使用するデータは,
SIGNATE というサイトの【練習問題】銀行の顧客ターゲティング
(URL : `https://signate.jp`) を使用する.

R 言語の特徴

- よく Python と比較される
- 両方ともライブラリが豊富
- R はデータ分析に特化している
- それに対して Python は汎用性が高い言語

未知のデータのスコア上位 10 位の組み合わせ.

train	test	maxdepth	minbucket	cp	parms
0.926410712	0.905533167	21	14	1.00E-09	gini
0.9260673	0.905079642	13	14	1.00E-09	entropy
0.925896439	0.904790833	22	16	1.00E-08	entropy
0.921105049	0.904481269	18	18	1.00E-08	entropy
0.930072789	0.904369846	17	14	1.00E-07	entropy
0.924401266	0.904333604	15	16	1.00E-07	entropy
0.920277985	0.904277784	21	18	1.00E-07	gini
0.928844593	0.904164048	18	15	1.00E-09	entropy
0.925190769	0.903967297	16	16	1.00E-09	entropy
0.925770049	0.903952802	16	14	1.00E-09	gini

データ分析講座について

講師：佐野 遼太郎氏 (yukari17, Twitter : @yukari_data)

講座名：Kaggle Master によるデータ分析技術者養成講座

(全 5 回, 1 か月)

URL : https://yukari17.connpass.com/event/