機械学習エンジニアコース
Week4 Session

－ 機械学習の基礎 －

DIVE INTO CODE

2019年7月31日（水）

鈴木 達哉

今日の流れ

1. チェックイン・KPT（担当：鈴木）
2. 講義（担当：鈴木）
3. お昼休み
4. ペアプログラミング（担当：遠藤）
5. KPT・チェックアウト（担当：遠藤）

構成

1. 提言
2. 導入
3. 今日の目的
4. 授業前課題の確認
5. 授業課題
6. 質疑応答

提言

未来は目指すものであり、創るものだ。

安宅 和人

“シン・ニホン”AI×データ時代における□本の再□と□材育成（ https://www.mof.go.jp/pri/research/conference/fy2017/inv2017_04_02.pdf ）

導入 – カリキュラムの確認

ゴールから逆算して設計されたカリキュラムになっています。数歩先を見据え、走りながら考えてください。

就職

機械学習エンジニアになる。

Term3（10月）

問題を定義して、時間内に解決できる。

Term2（9月）

現在の問題を認識し、既存の解決策を適用できる。

Term1（8月）

古典的理論を知り、定石を身につける。

事前学習（7月）

道具を活かす思考を身につける。

導入 – カリキュラムの確認

Term3（10月）

問題を定義して、時間内に解決できる。

- 調査
- 仮説を立てる
- 条件を知る
- SQL
- データセット作成
- Docker
- Raspberry Pi
- 公開

導入 – カリキュラムの確認

Term2（9月）

現在の問題を認識し、既存の解決策を適用できる。

- 深層学習
- 画像認識
- 自然言語処理
- 論文読解
- コードリーディング
- OSS
- フレームワーク

導入 – カリキュラムの確認

Term1（8月）

古典的理論を知り、定石を身につける。

- 教師あり学習
- 教師なし学習
- スクラッチ
- Kaggle

導入 – カリキュラムの確認

事前学習（7月）

道具を活かす思考を身につける。

- プログラミング（Python）
- 機械学習のための数学
- 探索的データ分析
- 機械学習の基礎
- オブジェクト指向

導入 – 大切な考え方

今月は、道具を活かす思考を身につける。

	◎ Good	× Not Good
1	**「何があればできるだろう」と自分に問う**	「まだ習ってないからなあ」と立ち止まる
2	**「本当にあっているのか」と疑う**	「◯◯に書いてあったから」と信じ込む
3	**「まずはやってみよう」と手を動かす**	「もっと分かってからやろう」と慎重になる

今日の目的

学びの目的。目的が、人の役割と必要な学びを明確にする。明確な学びは、成長実感と自信につながる。

	目的とすること	目的としないこと
1	**仲間とプログラムの考え方を学ぶ**	関数をたくさん覚える
2	**機械学習の基礎を知る**	課題を早く完成させる
3	**新人ビジネスマンの気持ちになる**	

今日の目的:機械学習の基礎を知る

「機械学習の基礎を知る」

そもそも機械学習の基礎とは何か。
モデル作りにこだわることではない。

- **機械学習の一連の流れの実装と実行を繰り返す**

今日の目的:新人ビジネスマンの気持ちになる

「新人ビジネスマンの気持ちになる。」

分析ツールを使う上で大切にしたい姿勢。以下のようなイメージを持ってみる。

- **自分は新人ビジネスマン**
  - **ドメイン知識がない**
  - **データはある**
- **会社で先輩や上司にホウレンソウする**
  - **ビジネスゴールがある**

授業前課題の確認

授業前課題の解説を行います。

授業課題

DIVER 授業前課題で身につけた力を活用して、より実践的な問題にチャレンジ！

1. **信用情報の学習**
   a. **コンペティション内容の確認**
   b. **学習と検証**
   c. **テストデータに対する推定**
   d. **特徴量エンジニアリング**

参考情報
- HomeCredit_columns_description.csv
  https://www.kaggle.com/c/home-credit-default-risk/data

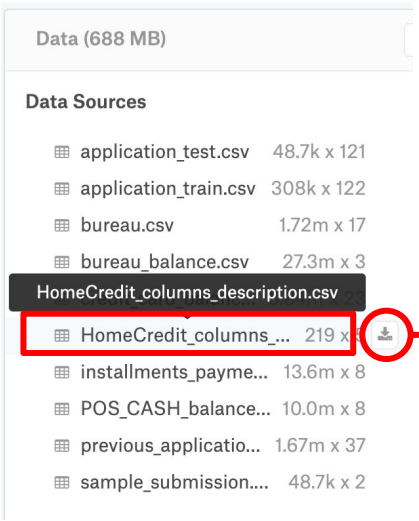
授業課題

Kaggleコンペティションに取り組むフロー

1. 問題提起を理解する
2. 評価指標を理解する
3. PublicとPrivateの比率を確認する
4. EDAする
5. First Submissionする
6. 指標値の関数をつくる
7. 前処理する
8. Feature Engineeringする
9. Trainingする
10. 指標値で評価する
11. Submissionする

授業課題

【注意事項】HomeCredit_columns_description.csv を見て、データセットの列についての説明を確認しよう。



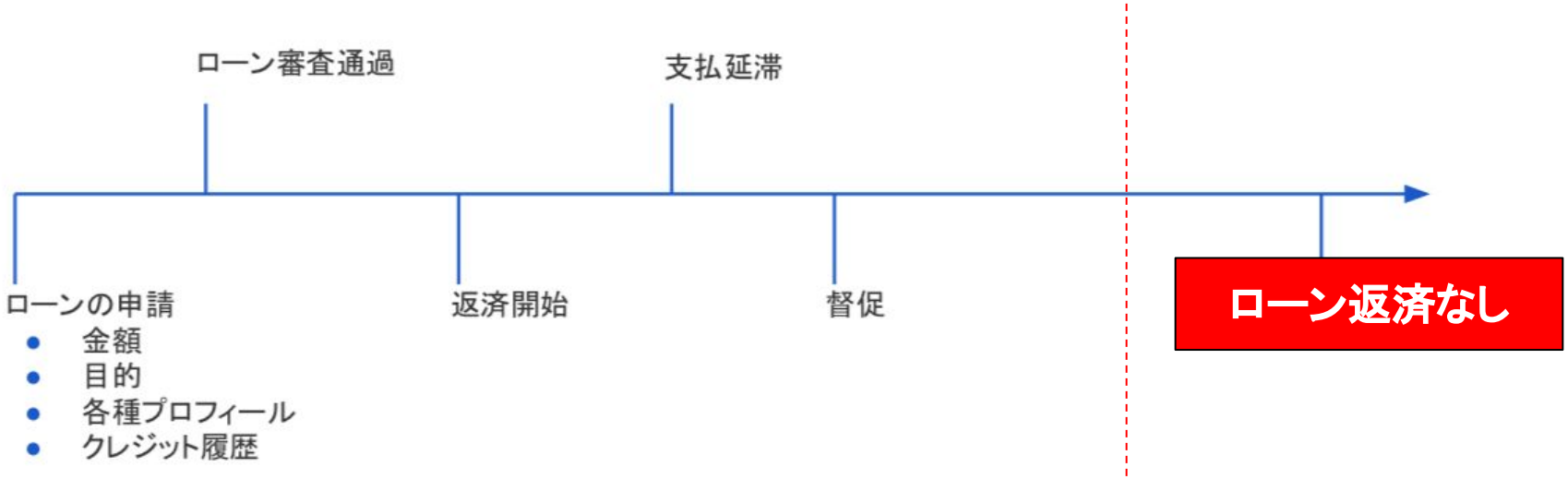
	A	B	C	D	E
1		Table	Row	Description	翻訳
2	1	application_{trai	SK_ID_CURR	ID of loan in our sample	私たちのサンプル中のローンのID
3	2	application_{trai	TARGET	Target variable (1 - client with payment difficulties	ターゲット変数（1 - 支払が困難なクライアント：彼/彼女は
4	5	application_{trai	NAME_CONTRACT_TYPE	Identification if loan is cash or revolving	識別ローンは、現金または回転であれば
5	6	application_{trai	CODE_GENDER	Gender of the client	クライアントの性別
6	7	application_{trai	FLAG_OWN_CAR	Flag if the client owns a car	フラグクライアントが車を所有している場合
7	8	application_{trai	FLAG_OWN_REALTY	Flag if client owns a house or flat	旗クライアントは、家やフラットを所有している場合
8	9	application_{trai	CNT_CHILDREN	Number of children the client has	子供の人数クライアントが持っています
9	10	application_{trai	AMT_INCOME_TOTAL	Income of the client	クライアントの収入
10	11	application_{trai	AMT_CREDIT	Credit amount of the loan	ローンの与信額
11	12	application_{trai	AMT_ANNUITY	Loan annuity	ローン年金
12	13	application_{trai	AMT_GOODS_PRICE	For consumer loans it is the price of the goods fo	消費者ローンのために、それは、ローンが与えられるため、
13	14	application_{trai	NAME_TYPE_SUITE	Who was accompanying client when he was appl	彼は融資を申請されたとき、誰が、クライアントに付随して
14	15	application_{trai	NAME_INCOME_TYPE	Clients income type (businessman, working, mat	クライアントの収入の種類（実業家、作業、産休、...）
15	16	application_{trai	NAME_EDUCATION_TYPE	Level of highest education the client achieved	最高の教育の水準クライアントを実現
16	17	application_{trai	NAME_FAMILY_STATUS	Family status of the client	クライアントの家族の状況
17	18	application_{trai	NAME_HOUSING_TYPE	What is the housing situation of the client (renting	クライアントの住宅事情は何である（親と同居、レンタル、
18	19	application_{trai	REGION_POPULATION_RELATIVE	Normalized population of region where client lives	クライアントが住んでいる地域の正規化された人口（高い数
19	20	application_{trai	DAYS_BIRTH	Client's age in days at the time of application	塗布時の日中のクライアントの年齢
20	21	application_{trai	DAYS_EMPLOYED	How many days before the application the person	何日前に申請人は、現在の雇用を開始しました
21	22	application_{trai	DAYS_REGISTRATION	How many days before the application did client c	何日の適用前に、クライアントは自分の登録を変更しました
22	23	application_{trai	DAYS_ID_PUBLISH	How many days before the application did client c	何日の適用前に、クライアントは、彼がローンを適用したと
23	24	application_{trai	OWN_CAR_AGE	Age of client's car	クライアントの車の時代
24	25	application_{trai	FLAG_MOBIL	Did client provide mobile phone (1=YES, 0=NO)	クライアントは、携帯電話を提供しました（1 = YES、0 = N

Googleスプレッドシートで **GOOGLETRANSLATE関数** を使おう！
https://support.google.com/docs/answer/3093331?hl=ja&authuser=0

授業課題

特徴量の増やし方

- その専門分野の論文を読むことからはじめる
- ドメイン知識を勉強する
  - 業界人にヒアリングする
- 時系列にデータを追ってみる

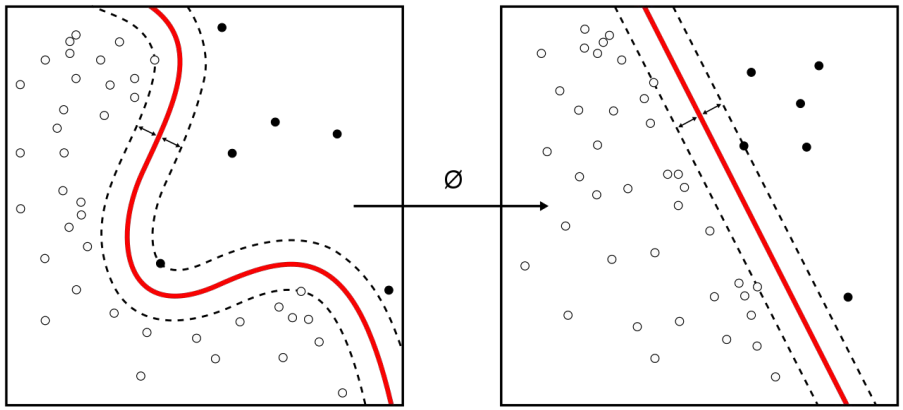


DataRobot Essentialsハンズオントレーニング資料より

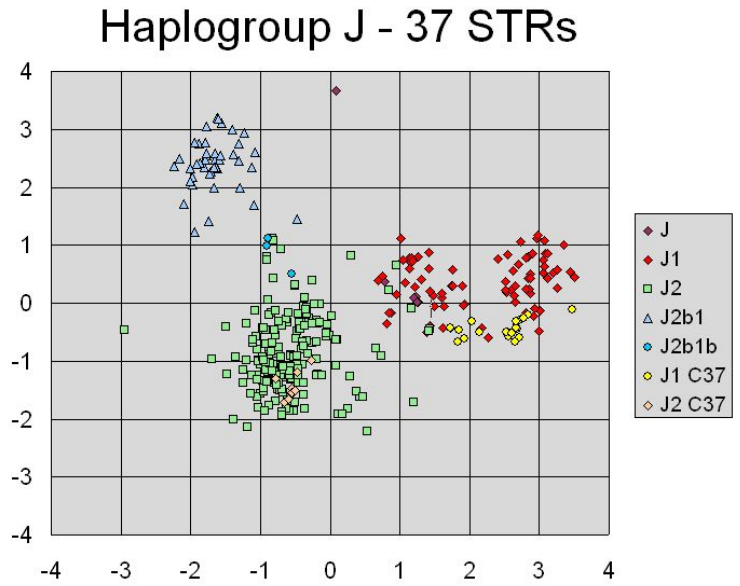
授業課題

特徴量の減らし方

- 教師なし学習を活用する
  - 主成分分析（Principal Component Analysis）を行う



Wikipediaより



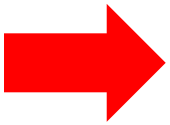
授業課題

不均衡データは、なぜ欠損しているのかの仮説を立てよう。

1. ランダムなもの

2. 事実上欠損している
   a. データがない、営業していない

3. 恣意的に欠損している
   a. 年収アンケートに答えない

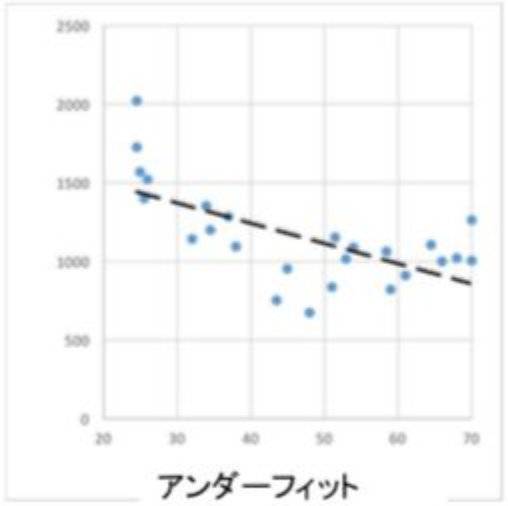


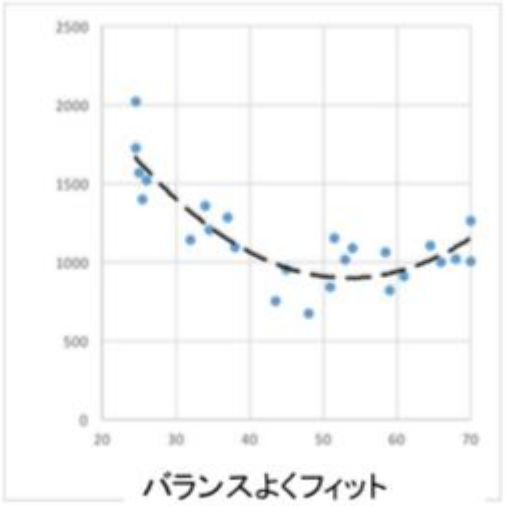
欠損事由に応じてダミーで補うか否かを考えよう

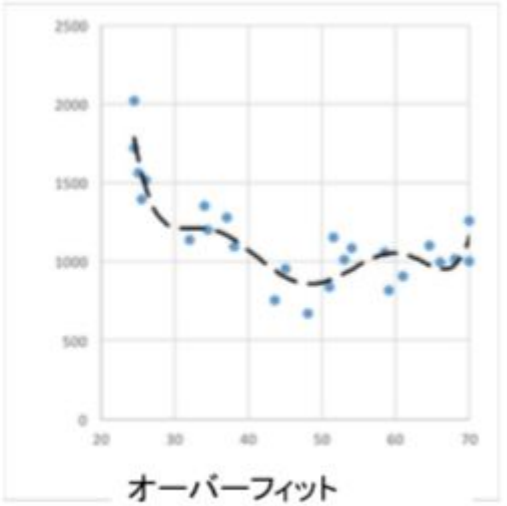
授業課題

汎用的に使えるよう学習している状態が好ましい。過学習を防ぎ、汎化性能を高める。

学習不足 ← **適度** → **丸暗記**







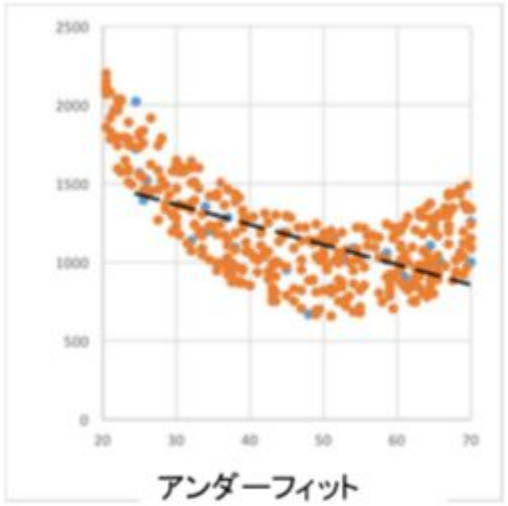
DataRobot Essentialsハンズオントレーニング資料より

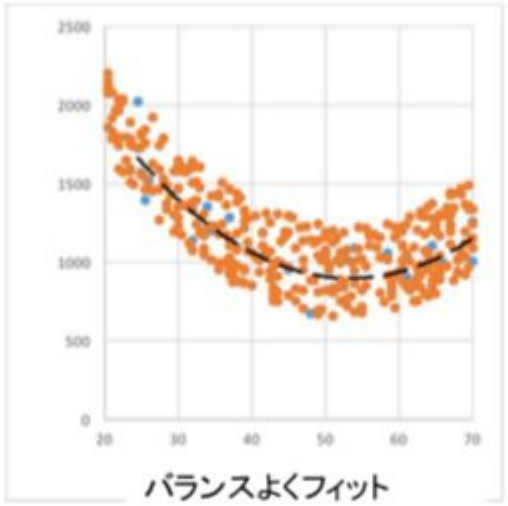
別名：過学習

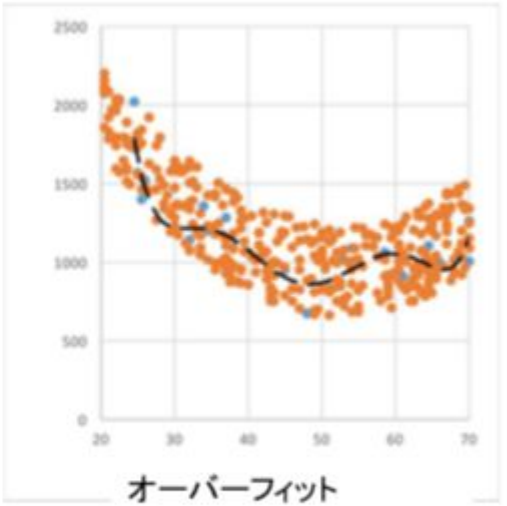
授業課題

未知のデータを使った予測では、学習不足でも、丸暗記状態でも、予測精度は下がる。

学習不足 ← **適度** → **丸暗記**





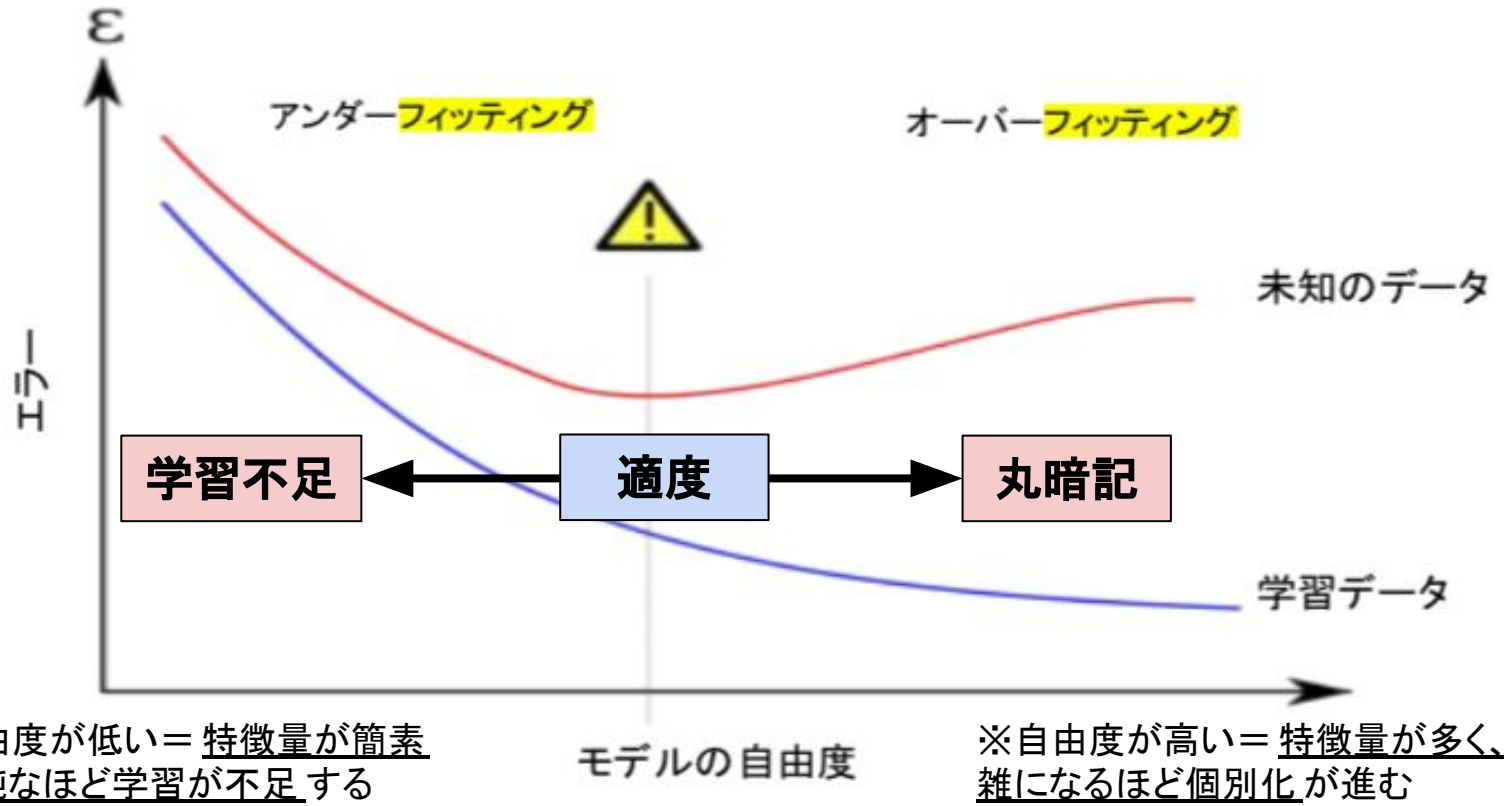


DataRobot Essentialsハンズオントレーニング資料より

”過ぎたるは、及ばざるがごとし。” 汎化性能があることが重要

授業課題

適度な精度になっているかどうかは、学習に使ったデータと未知のデータを予測してみた精度（エラー率）の差違で認識できる。

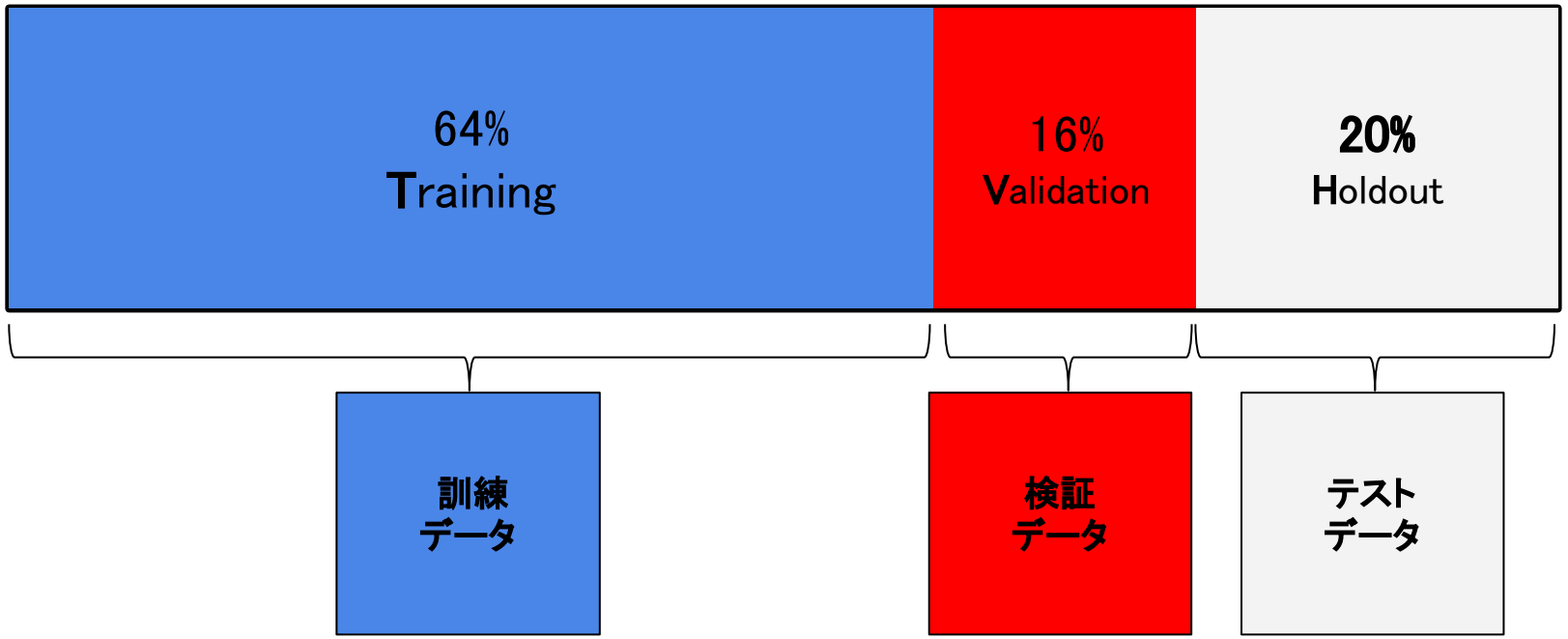


※自由度が低い＝特徴量が簡素で単純なほど学習が不足する

※自由度が高い＝特徴量が多く、複雑になるほど個別化が進む

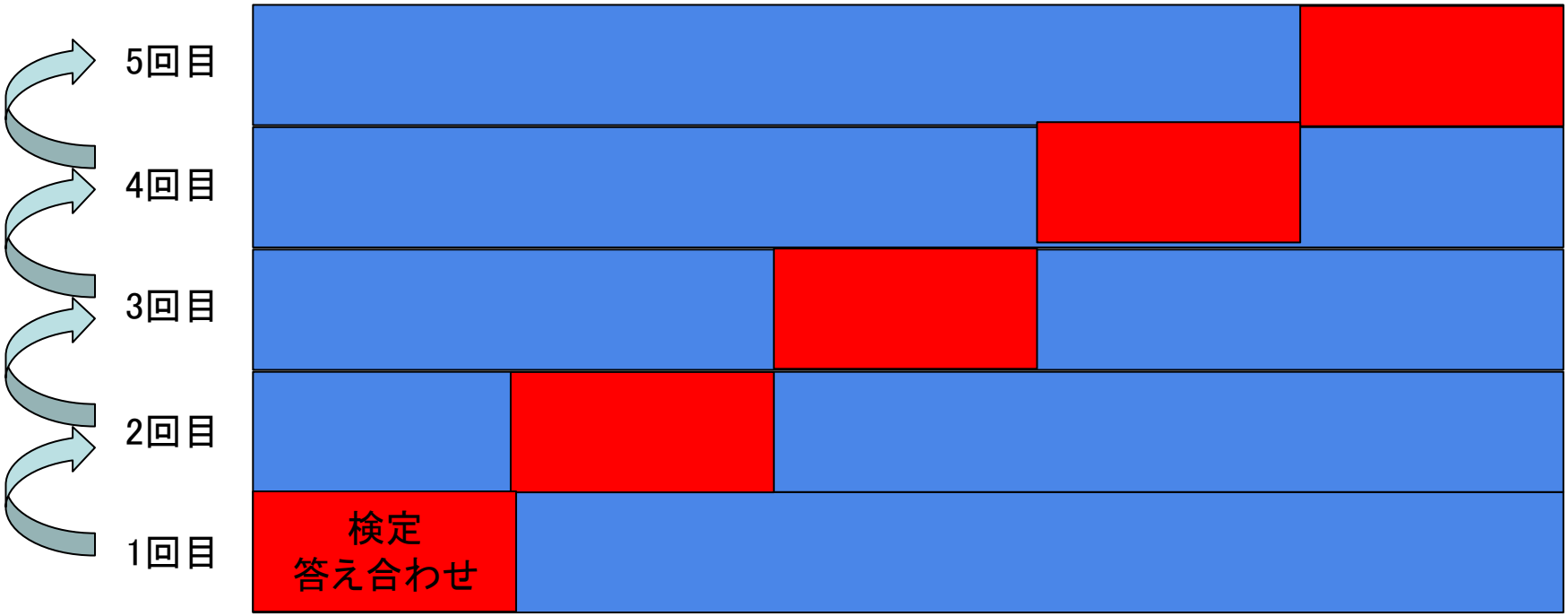
授業課題

アンダー・オーバーフィッティングを回避するために、データを分割して使用する。以下は分割の一例。Kaggleではテストデータは初めから別にあるため、訓練データの一部を検証データとする。



授業課題

訓練用と検証用データの分割を複数試す交差検証「クロスバリデーション」を行うことが理想的。（Sprnt1で扱う）



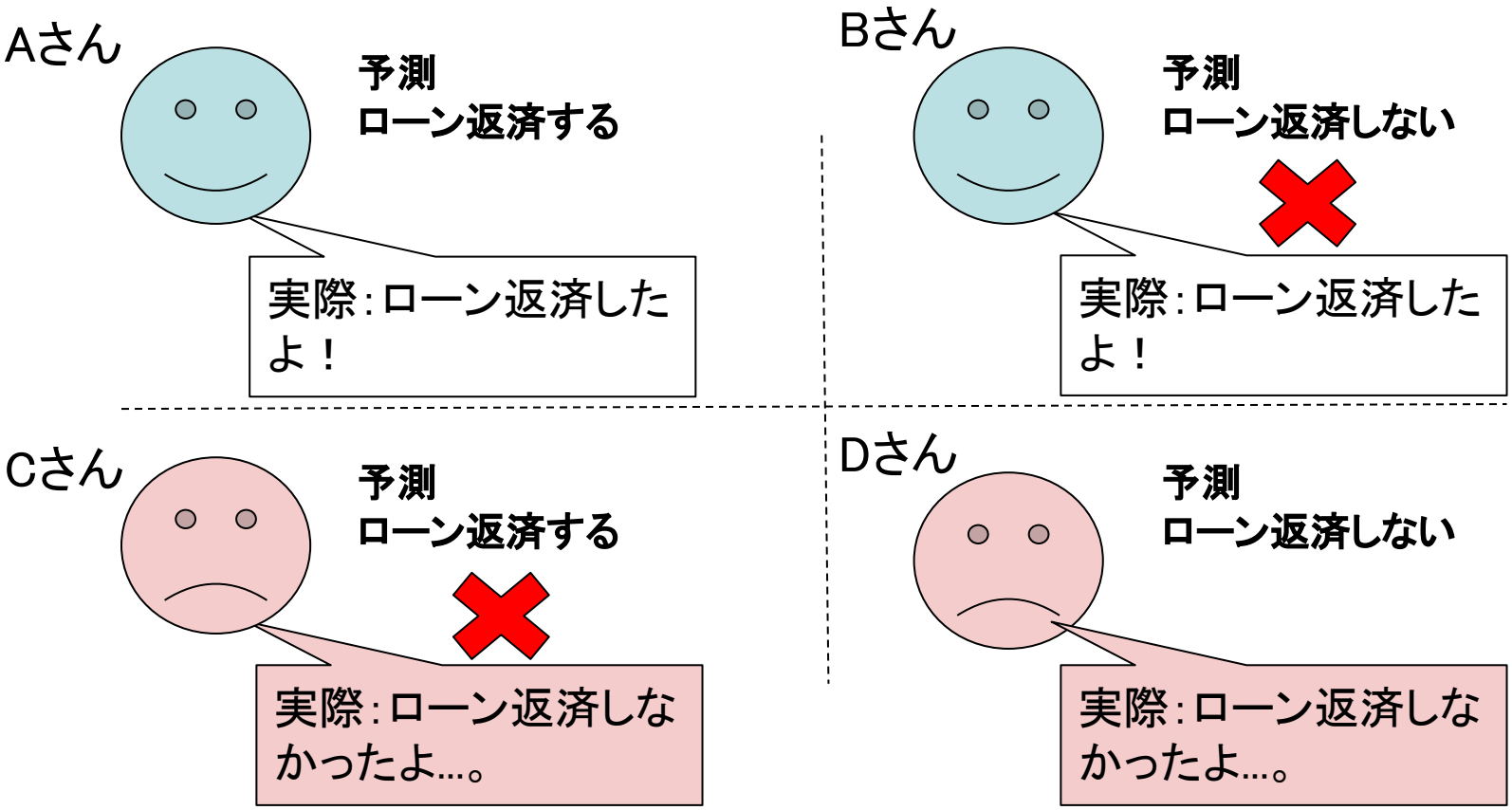
授業課題

分類問題の精度を考えてみよう。”正解”の率が良い？

	正解	予測値	しきい値				
	0か1	（確率）	0.1	0.3	0.5	0.7	0.9
1	1	0.7	1	1	1	1	0
2	0	0.3	1	1	0	0	0
3	0	0.8	1	1	1	1	0
4	0	0.4	1	1	0	0	0
5	0	0.9	1	1	1	1	1
6	0	0.5	1	1	1	0	0
7	0	0.7	1	1	1	1	0
8	0	0.2	1	0	0	0	0
9	0	0.5	1	1	1	0	0
10	0	0.1	1	0	0	0	0
		正解率	10%	30%	50%	70%	80%

授業課題

精度の指標値は、正解率以外にもある。



授業課題

実際と予測を照合して精度確認ができる表「混同行列」。

混同行列 （ Confusion Matrix ）		予測	
		ローン返済する （– 陰性）	ローン返済しない （+ 陽性）
実際	ローン返済する （– 陰性）	正解、 陰性だった	不正解、 陽性ではなかった
	ローン返済しない （+ 陽性）	不正解、 陰性ではなかった	正解、 陽性だった

授業課題

「混同行列」の TN, FN, FP, TP をおさえておこう。

混同行列 (Confusion Matrix)		予測 ※主語になる	
		ローン返済する (– 陰性)	ローン返済しない (+ 陽性)
実際	ローン返済する (– 陰性)	本当に陰性だった True Negative (TN)	陽性ではなかった False Positive (FP)
	ローン返済しない (+ 陽性)	陰性ではなかった False Negative (FN)	本当に陽性だった True Positive (TP)

授業課題

よく使う評価指標 Precision , Recall をおさえておこう。

指標	英名	和名	計算式	利用観点
TPR	**Precision**	適合率	TP ÷（TP + FP）	**Positiveな予測したもののうち、実際に正解だったものの比率**
FPR	**-**	偽陽性率	FP ÷（FP + TN）	**実際にはNegativeなもののうち、Positiveと誤って予測したものの比率**

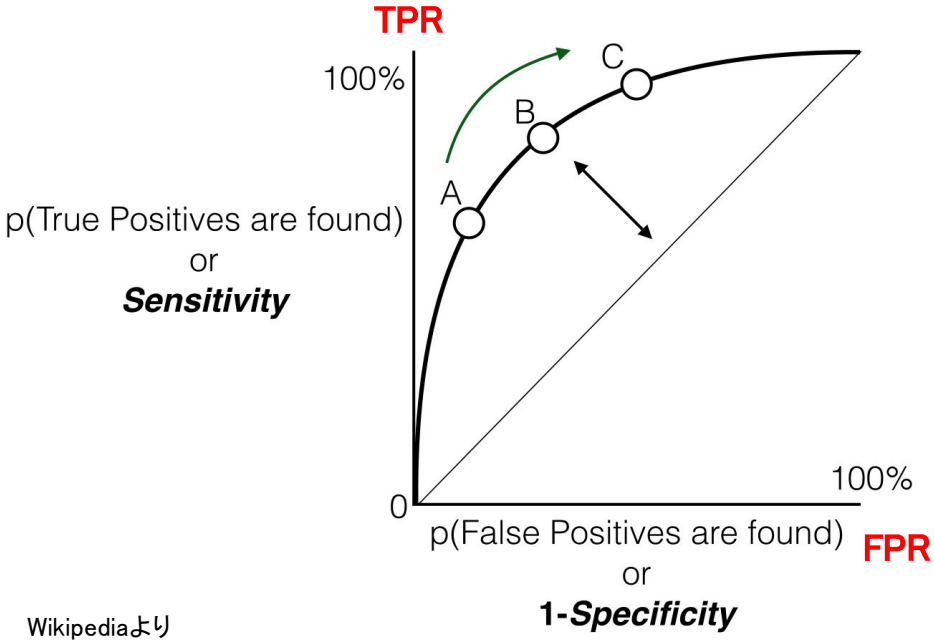
TPR		Predict	
		N	**P**
	N	TN	FP
	P	FN	**TP**

FPR		Predict	
		N	**P**
	N	TN	**FP**
	P	FN	TP

授業課題

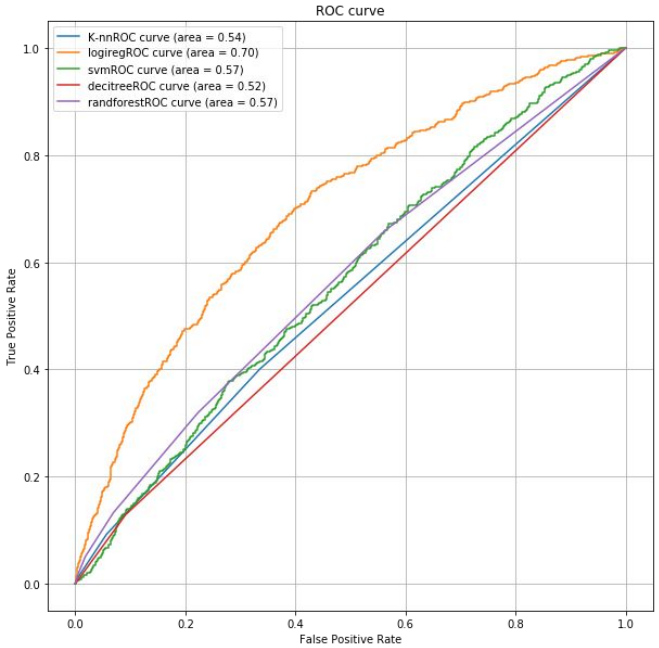
TPRを縦軸、FPRを横軸にとった評価用曲線「ROC曲線」をモデルの性能評価によく使う。この面積がAUC。

RoC曲線の図式



Wikipediaより

モデルの性能評価



授業課題

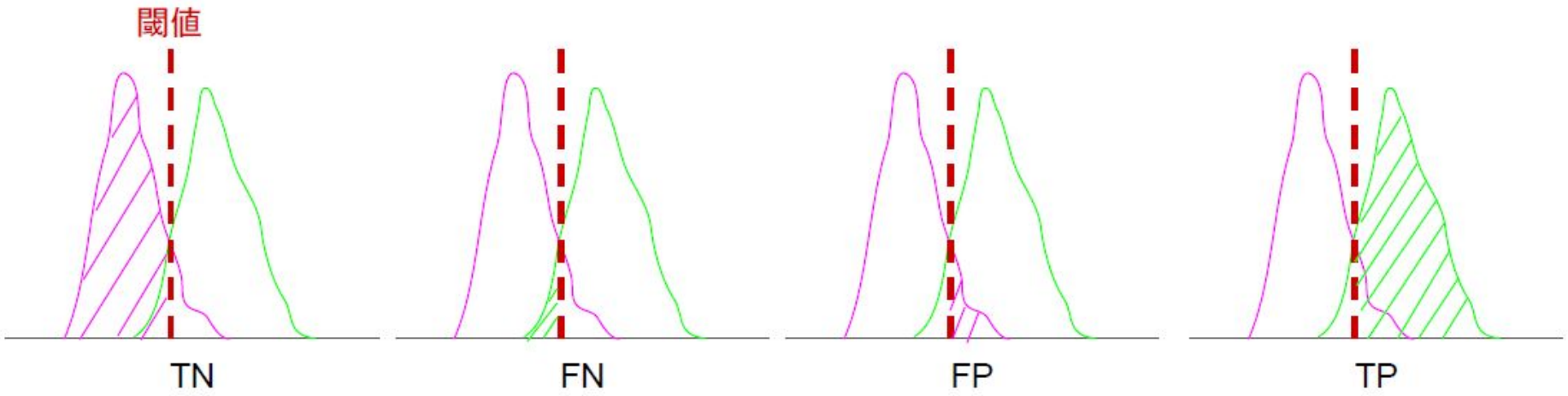
混同行列で表される指標は、予測分布のグラフとしても表すことができる。予測精度をしきい値で高めるイメージを持とう。

		Predict	
		N	P
	N	TN	FP
	P	FN	TP

		Predict	
		N	P
	N	TN	FP
	P	**FN**	TP

		Predict	
		N	P
	N	TN	**FP**
	P	FN	TP

		Predict	
		N	P
	N	TN	FP
	P	FN	**TP**



(再掲)今日の目的

学びの目的。目的が、人の役割と必要な学びを明確にする。明確な学びは、成長実感と自信につながる。

	目的とすること	目的としないこと
1	**仲間とプログラムの考え方を学ぶ**	関数をたくさん覚える
2	**機械学習の基礎を知る**	課題を早く完成させる
3	**新人ビジネスマンの気持ちになる**	

（再掲）提言

未来は目指すものであり、創るものだ。

安宅 和人

“シン・ニホン”AI×データ時代における□本の再□と□材育成（ https://www.mof.go.jp/pri/research/conference/fy2017/inv2017_04_02.pdf ）

質疑応答

講義は以上です。

ここまでで疑問に思った点はありますか？