

ブートキャンプ° for Data Science

2回目講義資料

Tryeting Inc. ながえゆうき

アジェンダ

- イン트로ダクション
 - この講義の趣旨の再確認
 - 前回の宿題の確認
 - プ・レ・ゼ・ン
- 早速やってみよう
 - アヤメのデータで、let's classification (Rでの分類問題)
 - API経由で予測モデルを活用する (サービス化・Bash-shell)
- 宿題

アジェンダ

- イン트로ダクション
 - この講義の趣旨の再確認
 - 前回の宿題の確認
 - プ・レ・ゼ・ン
- 早速やってみよう
 - アヤメのデータで、let's classification (Rでの分類問題)
 - API経由で予測モデルを活用する (サービス化・Bash-shell)
- 宿題

さて、どうでしたか？

今回の一連の講義は、レベルアップのためのもの

 ついていけない

 わからないことがあればきく

諦めないで、頑張りましょう！

本授業のゴール

<ゴール>

高度な統計学や機械学習理論とスキルを体系立てて理解し、ビジネス課題に適用できる人材育成を目指します。具体的に以下ことができるようになることを目指します。

- 高度なデータ分析技術を用いてビジネス課題解決ができる
- 「なぜこの分析手法を使うか」を技術的な側面とビジネス側面から判断できる
- PythonやRを使って求められている課題に対し適切な分析手法を実行できる
(やっていけないことも理解している)
- 分析結果を関係者にわかりやすく説明できる

<特徴>

- 体系立てて学習することで、データ分析技術の引き出しを増やします。
- 小規模クラスのため、クラス内での議論や講師とやりとりを多くすることで内容理解を深めることができます。
- 理解を深めるためのクイズやハンズオンで演習(宿題含む)を数多く実施し、理解するだけでなく手を動かせることを目標にします。

本授業のゴール

<ゴール>

高度な統計学や機械学習理論とスキルを体系立てて理解し、ビジネス課題に適用できる人材育成を目指します。具体的に以下ことができるようになることを目指します。

- 高度なデータ分析技術を用いてビジネス課題解決ができる
- 「なぜこの分析手法を使うか」を技術的な側面とビジネス側面から判断できる
- PythonやRを使って求められている課題に対し適切な分析手法を実行できる
(やっていけないことも理解している)
- 分析結果を関係者にわかりやすく説明できる

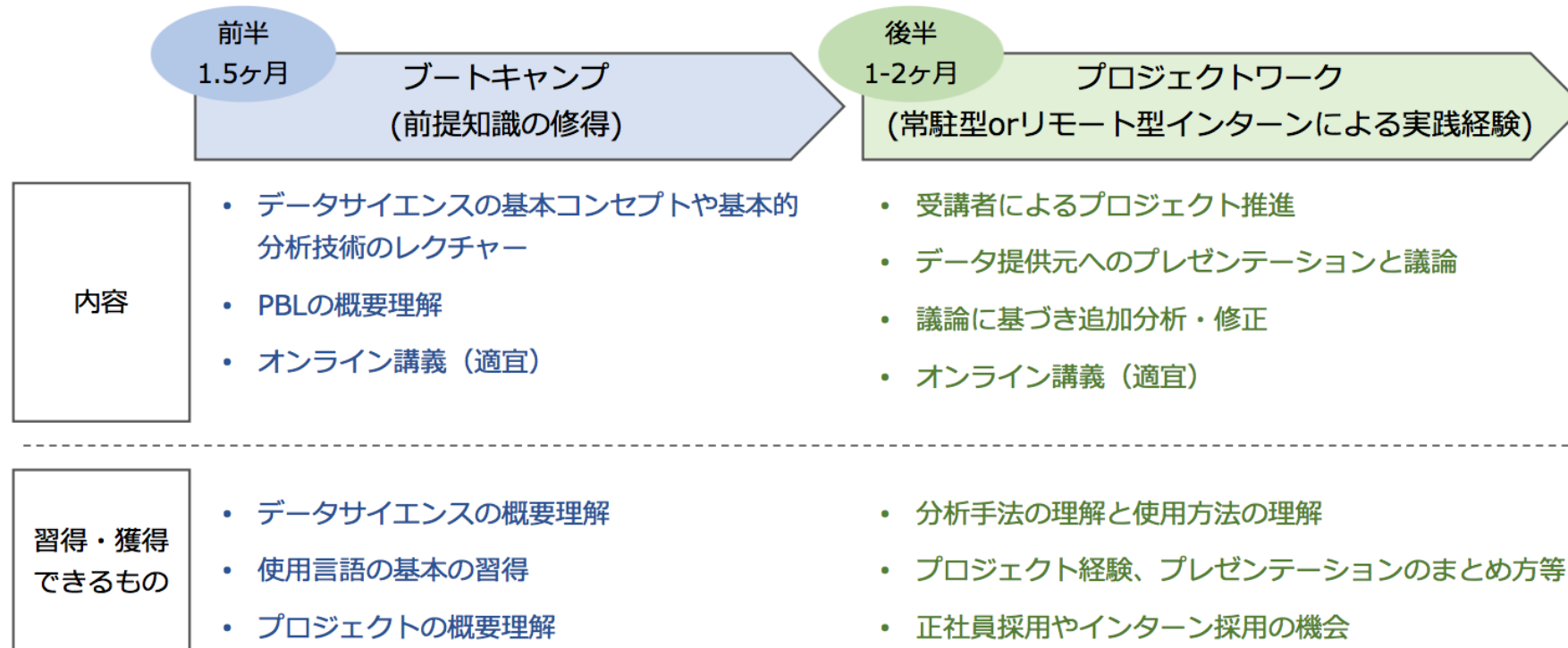
<特徴>

- 体系立てて学習することで、データ分析技術の引き出しを増やします。
- 小規模クラスのため、クラス内での議論や講師とやりとりを多くすることで内容理解を深めることができます。
- 理解を深めるためのクイズやハンズオンで演習(宿題含む)を数多く実施し、理解するだけでなく手を動かせることを目標にします。

プロジェクトの進行：こんな感じ

概要

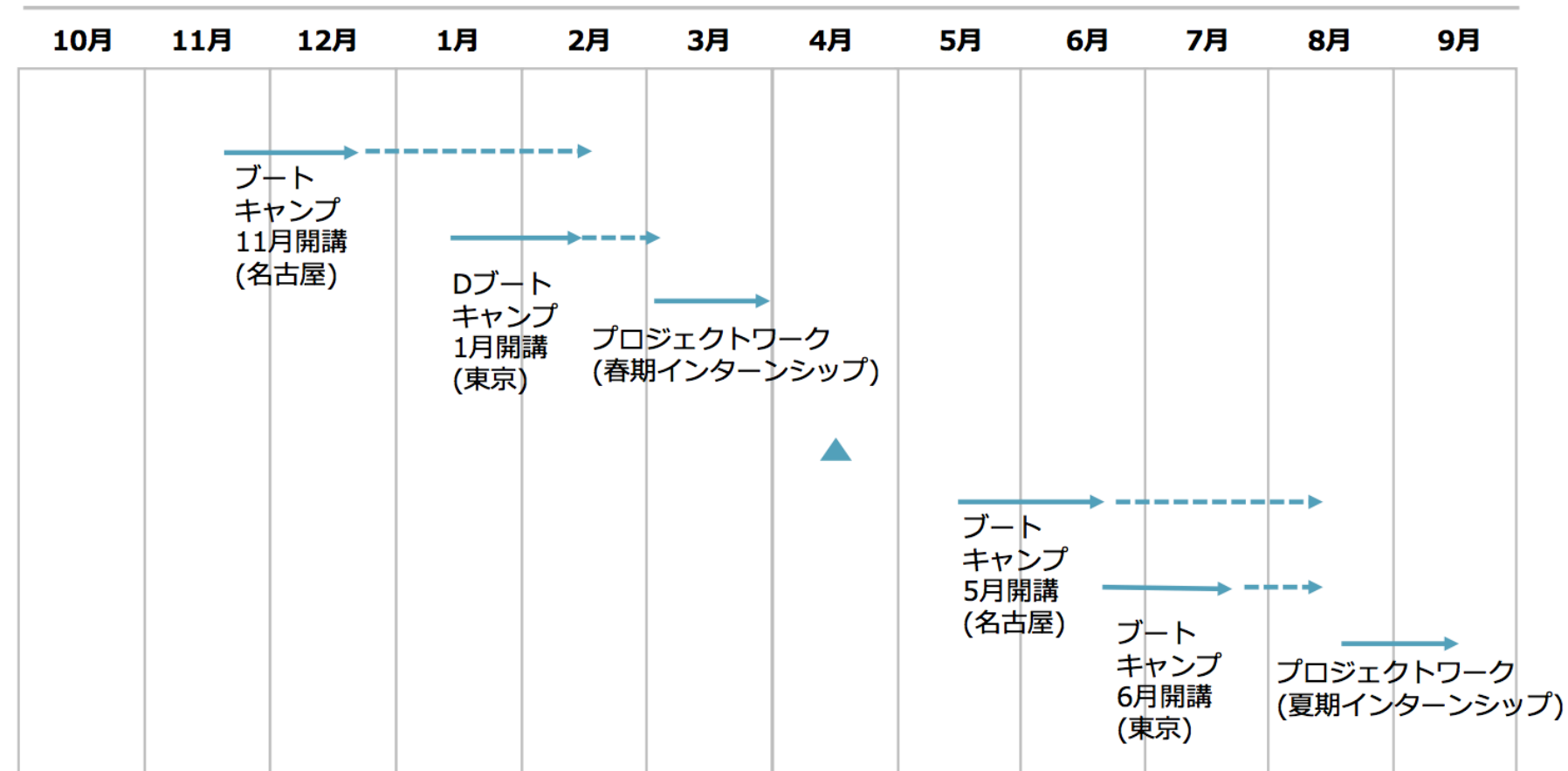
- 合計約2～3ヶ月のPBLプログラム
秋期開講のブートキャンプ ⇒ 春期のプロジェクトワーク
春期開講のブートキャンプ ⇒ 夏期のプロジェクトワーク
- メンターによるコーチングやピアディスカッションの場を提供
- プロジェクトと並行して名古屋大学のオンライン講座を提供



プロジェクトの俯瞰図

初年度は春期と夏期の年2回の会員企業によるプロジェクトワークを実施。それぞれに向け、東京・名古屋で事前のトレーニングプログラムを開講。

■初年度のスケジュール



要は…

頑張ると、企業が目をつけてくれる

就活もセットのデータサイエンスビジネス応用の授業。

(スポンサーが見ている)

ちなみに

氏名	所属	学年	経験のあるプログラミング言語、経験年数等について記入してください。
川添亮太郎	医学系研究科	D2	
筒井大貴	生命農学研究科	ポスドク	プログラミングに入るのかは不安ですが、Rなら遺伝子発現解析などで1年ほど使っていました
河合 翼	生命農学研究科	博士課程前期課程2年	シェルスクリプト(Unix)と若干R, それぞれ約1年です。
佐藤 真	工学研究科有機・高分子化学専攻	研究員	Fortran (簡単なデータ解析に使う程度)
岡本一輝	工学研究科	修士1年	Fortran 半年
山口毅志	生命農学研究科	博士課程2年	R言語を有意差検定程度で使用 (3年間)
植田格弥	名古屋大学	D2	Python 3カ月程度
山崎 匠	工学研究科 航空宇宙工学専攻	M2	C (授業でやったことがある程度), Python (積分計算に使用)
天本 義史	工学研究科	博士研究員	Python 1年
胡 蘇紅(こ そこう)	人文学研究科	D1	全くわからないんです。
櫻井 幹記 (さくらい よしき)	工学研究科計算理工学専攻	修士2年	fortran90,4年程 C言語,2年程
今井祐太	名古屋大学 創薬科学研究科	M1	R 一年半ほど
土屋 健斗	生命農学研究科		1 Java半年、python1年、VB半年
アラム タニムル	理学研究科	D3	
松田 光司	理学研究科 生命理学専攻 動物器官	研究員	
加藤優介 (かとうゆうすけ)	理学研究科物質理学専攻 (物理系)	博士後期課程2年	
松本 恵	工学研究科 化学・生物工学専攻 (生)	D3	C, C++, Java(企業にて7年間、組込み開発)、R(研究にて6年間)
陳璐	文学研究科	D3	CSS
林 卓弥	名古屋大学大学院理学研究科	修士課程2年	C言語1年、FORTRAN言語2年
鈴木俊章	創薬科学研究科	D3	
今枝立至	工学	D2	C言語、Fortran
金山英誠	農学部		4 matlab、python、共に半年程度

ながえゆうき 工学研究科 D2 R(原理主義者), python, bash-shell, Cなどなど

ちなみに、プログラムかけることに価値はない。どう使って、価値を認めさせるか。(ナーバスにならなくてもいい)

この講義でのルール(暫定版)

- 楽しむ(リタイアは、ご自由に)
- プログラミングそのものを教えるつもりない(ggrks)
- かといって、理論を教えるつもりない(教科書読んで)
- 社会人として扱います
- わからないことがあったら聞く(だいたいなんでも答える)

この講義でのルール(暫定版)

- 楽しむ(リタイアは、ご自由に)
- プログラミングそのものを教えるつもりない(ggrks)
- かといって、理論を教えるつもりない(教科書読んで)
- 社会人として扱います
- **わからないことがあったら聞く(だいたいなんでも答える)**

アジェンダ

- イン트로ダクション
 - この講義の趣旨の再確認
 - 前回の宿題の確認
 - プ・レ・ゼ・ン
- 早速やってみよう
 - アヤメのデータで、let's classification (Rでの分類問題)
 - API経由で予測モデルを活用する (サービス化・Bash-shell)
- 宿題

宿題

なんでもいいからオープンデータを見つけてきて、MLで予測を試みよう

- pptx 1枚スライド(16:9)で作成・次回発表1分ピッチ・質疑3分
 - 入力データの構成は？
 - アウトプットの構成は？
 - わかると何が嬉しい？
- 作成ファイル
 - ピッチファイル：{name}_1_work.pptx
 - 作成コード：{name}_1_work.R
 - 作成MLモデルオブジェクト：{name}_1_work.obj
 - 印刷資料人数分

アジェンダ

- イン트로ダクション
 - この講義の趣旨の再確認
 - 前回の宿題の確認
 - プ・レ・ゼ・ン
- 早速やってみよう
 - アヤメのデータで、let's classification (Rでの分類問題)
 - API経由で予測モデルを活用する (サービス化・Bash-shell)
- 宿題

このirisデータ、一体何？

- 米国の植物学者Edgar Andersonによって1936年に発表された研究のなかで、もともとの計測が行なわれている (Anderson, 1936)。数値データが報告されたのは、英国の統計学者Ronald Aylmer Fisherの1936年の論文 (Fisher, 1936)。
- setosaに比べて, versicolorとvirginicaの形態はよく似ている。setosaとの違いは, virsinicaのほうが大きく, versicolorで小さい。versicolorは, setosaとvirginicaの雑種に由来しているかもしれない。

このirisデータ、一体何？

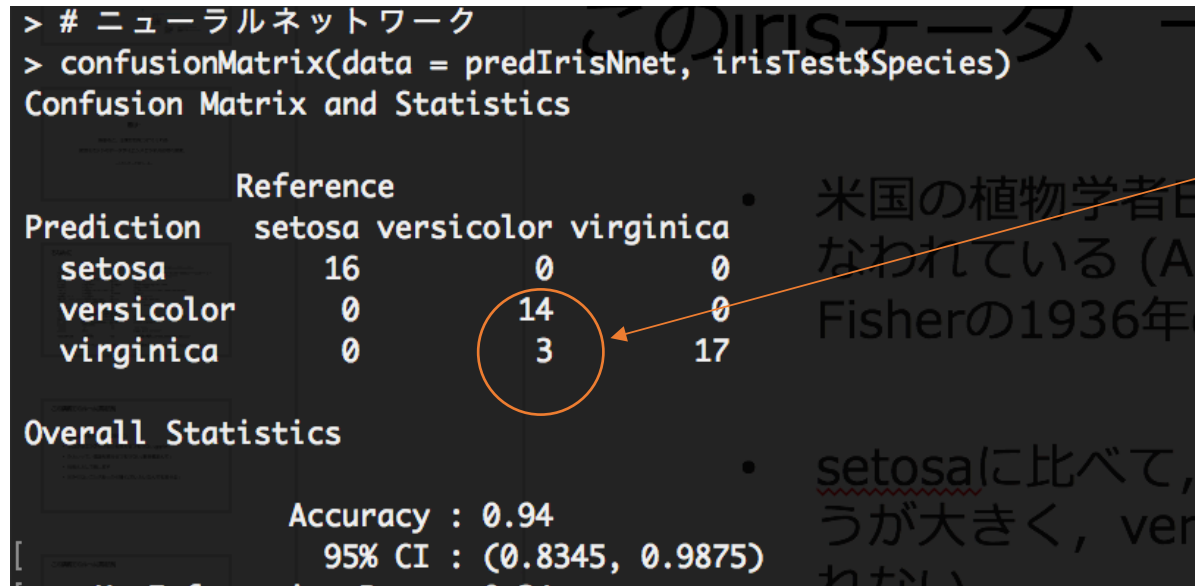
- 米国の植物学者Edgar Andersonによって1936年に発表された研究のなかで、もともとの計測が行なわれている (Anderson, 1936)。数値データが報告されたのは、英国の統計学者Ronald Aylmer Fisherの1936年の論文 (Fisher, 1936)。
- setosaに比べて、versicolorとvirginicaの形態はよく似ている。setosaとの違いは、virginicaのほうが大きく、versicolorで小さい。versicolorは、setosaとvirginicaの雑種に由来しているかもしれない。

```
> # ニューラルネットワーク
> confusionMatrix(data = predIrisNnet, irisTest$Species)
Confusion Matrix and Statistics
```

	Reference		
Prediction	setosa	versicolor	virginica
setosa	16	0	0
versicolor	0	14	0
virginica	0	3	17

```
Overall Statistics
```

Accuracy : 0.94
95% CI : (0.8345, 0.9875)



これ。アルゴリズムではどうしようもない

きちんとデータの本質を知って、解析をすることが必要である好例。

アジェンダ

- イン트로ダクション
 - この講義の趣旨の再確認
 - 前回の宿題の確認
 - プ・レ・ゼ・ン
- 早速やってみよう
 - アヤメのデータで、let's classification (Rでの分類問題)
 - API経由で予測モデルを活用する (サービス化・Bash-shell)
- 宿題

アジェンダ

- イン트로ダクション
 - この講義の趣旨の再確認
 - 前回の宿題の確認
 - プ・レ・ゼ・ン
- 早速やってみよう
 - アヤメのデータで、let's classification (Rでの分類問題)
 - API経由で予測モデルを活用する (サービス化・Bash-shell)
- 宿題

- 下記のお客様のご要望を聞いて、利益最大化に貢献できそうな方法を考案し、実現できるシステムの構成を、機械学習を交えてご提案せよ(コードは書かなくてもいい)
- 小売店舗のオーナー
- POSによる簡単な購買履歴データを持っている(後述)
- 「データは集めているだけ」
- 「どのお客様にクーポンを配ればいいのかわからない」
- 「単価を上げたい」

次回予告

イケてるWEBサービスをコマンドラインで作ろうぜ(サービス)の巻

使用言語 : R, bash-shell