

Python勉強会@HACHINONE

第19章

機械学習はやわかり

お知らせ

Python勉強会@HACHINOHEでは、ジョン・V・グッターゲ『Python言語によるプログラミング入門ダクション』近代科学社、2014年をみんなで勉強しています。

この本は自分で読んで考えて調べると力が付くように書かれています。

自分で読んで考えて調べる前に、このスライドを見るのは、いわば**ネタバレ**を聞かされるようなものでもったいないです。

是非、本を読んでからご覧ください。

機械学習とは？

Python勉強会@HACHINOHE

- ・ アーサー・サミュエル、機械学習とは「明示的なプログラムを必要としない学習能力をコンピュータに与える分野」、1959年
- ・ 今日的には、「データにある暗黙のパターンから有益な推論を自動的に学習するプログラムを作成する分野」
- ・ 例: 線形回帰(第15章のバネのびや発射体の軌跡)
 - ・ データの集まりを代表するモデル(曲線)を自動的に学習
 - ・ データの予測にも利用できる

教師あり機械学習の例

Python勉強会@HACHINOHE

- トレーニング・データ: ラベル付けされたデータ、教師データ
背が高い: {エイブラハム・リンカーン, ジョージ・ワシントン, シャルル・ド・ゴール}
背が低い: {ベンジャミン・ハリソン, ジェームズ・マディソン, ルイ=ナポレオン}
- 特徴ベクトル: 不完全な情報

名前	国籍	役職	身長
エイブラハム・リンカーン	アメリカ	大統領	193
ジョージ・ワシントン	アメリカ	大統領	189
ベンジャミン・ハリソン	アメリカ	大統領	168
ジェームズ・マディソン	アメリカ	大統領	163
ルイ=ナポレオン	フランス	大統領	169
シャルル・ド・ゴール	フランス	大統領	196

- テスト・データ

トマス・ジェファーソン	アメリカ	大統領	189
-------------	------	-----	-----

背が高い？低い？を判定

機械学習の種類

Python勉強会@HACHINOHE

- 教師あり学習
 - k近傍法
 - 線形回帰
 - ロジスティック回帰
 - サポートベクターマシン
 - 決定木、ランダムフォレスト
 - ニューラルネットワーク
- 教師なし学習
 - クラスタリング
 - k平均
 - 階層型クラスタ分析
 - 期待値最大化法
 - 可視化と次元削減
 - 主成分分析
 - カーネル主成分分析
 - t分布型確率的近傍埋め込み法
 - 相関ルール学習
 - アプリオリ
 - eclat

Aurélien Géron 『scikit-learnとTensorFlowによる実践機械学習』 オライリージャパン、2018年より

特徴ベクトル

Python勉強会@HACHINOHE

- 特徴抽出: シグナルと雑音を分離、ふつう最適なシグナルはなく難しいタスク
 - 適当でない特徴を使うと
 - 正しくないモデルができる: 特徴の数が多過ぎたり
 - 計算時間が増える
- 例: プログラミングの授業の成績
 - シグナル: プログラミングの経験、数学能力
 - ノイズ: 性別
- 例: ワイン好き
 - シグナル: 年齢、居住国
 - ノイズ: 利き手

爬虫類を分類してみよう

Python勉強会@HACHINOHE

- 1つずつデータを加えて検討していくと？

爬虫類	名前	特徴量ベクトル				
		産卵	うろこ	有毒	変温	足の数
○	コブラ	○	○	○	○	0
○	ガラガラヘビ	○	○	○	○	0
○	ボア	×	○	×	○	0
○	アリゲータ	○	○	×	○	4
×	ヤドクガエル	○	×	○	×	4
×	サケ	○	○	×	○	0
○	ニシキヘビ	○	○	×	○	0

- 羊膜を持つか否かで爬虫類を分類できるが、その情報はなかった

判定の精度

Python勉強会@HACHINOHE

- 精度の表現

		判定	
		はい	いいえ
実際	はい	真陽性	偽陰性
	いいえ	偽陽性	真陰性
余計なものも判定			

- 「爬虫類」は、「うろこを持つ」「変温動物」というモデルの場合
 - 偽陰性はない: サンプルの爬虫類はこの条件を満たす
 - 偽陽性はある: サケはこの条件を満たすが爬虫類でない

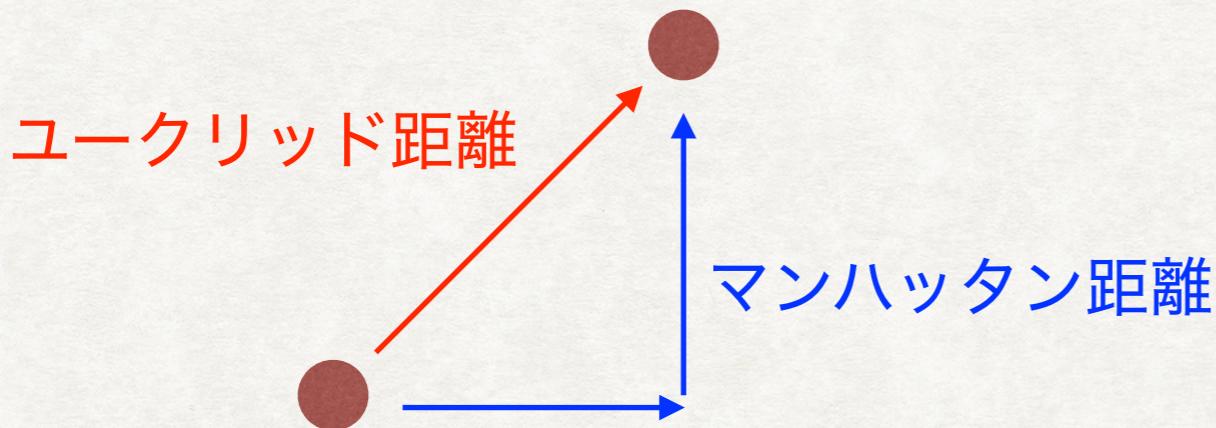
ミンコフスキー距離

Python勉強会@HACHINOHE

- ・ ミンコフスキー距離: 一般的な距離の表現

$$distance(V1, V2, p) = \left(\sum_{i=1}^{len} |V1 - V2|^p \right)^{\frac{1}{p}}$$

- ・ $p=1$: マンハッタン距離、縦横にしか動けない場合の距離
- ・ $p=2$: ユークリッド距離、斜めに動けるときの距離



動物のクラス図

Python勉強会@HACHINOHE

動物
名前
特徴量ベクトル
初期化(名前, 特徴量ベクトル)
名前を取得()
特徴量ベクトルを取得()
ユークリッド距離(他の動物)

動物間の距離: とりあえず3種類で

Python勉強会@HACHINOHE

- 特徴量ベクトル

名前	産卵	うろこ	有毒	変温	足の数
ガラガラヘビ	○	○	○	○	0
ボア	×	○	×	○	0
ヤドクガエル	○	×	○	×	4

- ユークリッド距離

	ガラガラヘビ	ボア	ヤドクガエル
ガラガラヘビ	--	1.414	4.243
ボア	1.414	--	4.472
ヤドクガエル	4.243	4.472	--

- ヤドクガエルは、ガラガラヘビやボアと似ていない
 - どちらかというとガラガラヘビに似ている

動物間の距離: アリゲータを追加

Python勉強会@HACHINOHE

- 特徴量ベクトル

名前	産卵	うろこ	有毒	変温	足の数
ガラガラヘビ	○	○	○	○	0
ボア	×	○	×	○	0
ヤドクガエル	○	×	○	×	4
アリゲータ	○	○	×	○	4

- アリゲータはヤドクガエルと近い
- おかしい??
- アリゲータとガラガラヘビの違い: 2件
- アリゲータとヤドクガエルの違い: 3件
- 足の数(0~4)の影響が他の因子(0, 1)より大きい?

- 距離

	ガラガラヘビ	ボア	ヤドクガエル	アリゲータ
ガラガラヘビ	--	1.414	4.243	4.123
ボア	1.414	--	4.472	4.123
ヤドクガエル	4.243	4.472	--	1.732
アリゲータ	4.123	4.123	1.732	--

動物間の距離: 足の本数を足のあるなしに変更

Python勉強会@HACHINOHE

- 特徴量ベクトル

名前	産卵	うろこ	有毒	変温	足
ガラガラヘビ	○	○	○	○	×
ボア	×	○	×	○	×
ヤドクガエル	○	×	○	×	○
アリゲータ	○	○	×	○	○

- 距離

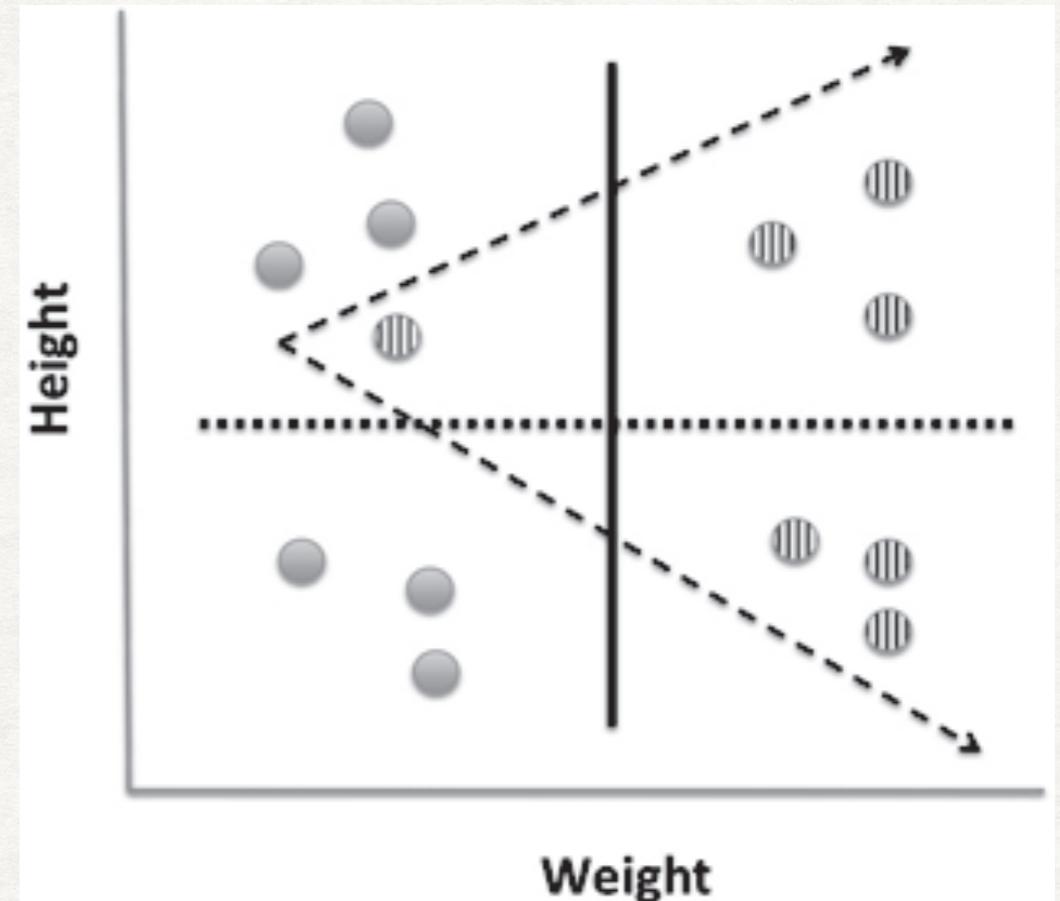
	ガラガラヘビ	ボア	ヤドクガエル	アリゲータ
ガラガラヘビ	--	1.414	1.732	1.414
ボア	1.414	--	2.236	1.414
ヤドクガエル	1.732	2.236	--	1.732
アリゲータ	1.414	1.414	1.732	--

- ヤドクガエルが他の動物間よりも遠い

分類: クラスタリング

Python勉強会@HACHINOHE

- 身長、体重、ストライプのシャツのデータ
- 線形(1直線)に分離できる
 - 身長: 上下
 - 体重: 左右
- 線形に分離できない
 - ストライプのシャツ



クラスタリング問題の定式化

Python勉強会@HACHINOHE

- クラスタリングは最適化問題
 - ある制約のもとに、目的関数を最適化する
- 目的関数
 - 同じクラスターに属している標本間の「距離(相違)」を最小化する
 - 距離の基準として、特徴ベクトルの分散が考えられる
- 計算方法: クラスターの大きさを考慮していないバージョン

• 標本の平均 $mean(c) = sum(V)/len(V)$

C: クラスターの集合全体

c: クラスター

V: 標本の特徴ベクトルのリスト

e: 標本

• 標本の分散 $variance(c) = \sqrt{\sum_{e \in c} distance(mean(c), e)^2}$

• 相違性 $dissimilarity(C) = \sum_{c \in C} variance(c)$

※単純に結果だけ見るとヤバい。k個の標本をk個のクラスターに分類すると相違性は0

標本とそのクラスターのクラス図

Python勉強会@HACHINOHE

標本	クラスター
名前	標本群
特徴量ベクトル	標本の型
ラベル	重心
初期化(名前, 特徴量ベクトル, ラベル)	初期化(標本群, 標本の型)
特徴量ベクトルの次元を取得()	更新(標本群)
特徴量ベクトルを取得()	クラスターに含まれる標本群を取得()
ラベルを取得()	クラスターに含まれる標本群のサイズ()
名前を取得()	重心を取得()
ユークリッド距離(他の標本)	重心を計算()
文字列化()	分散を取得()
	文字列化()

k平均法

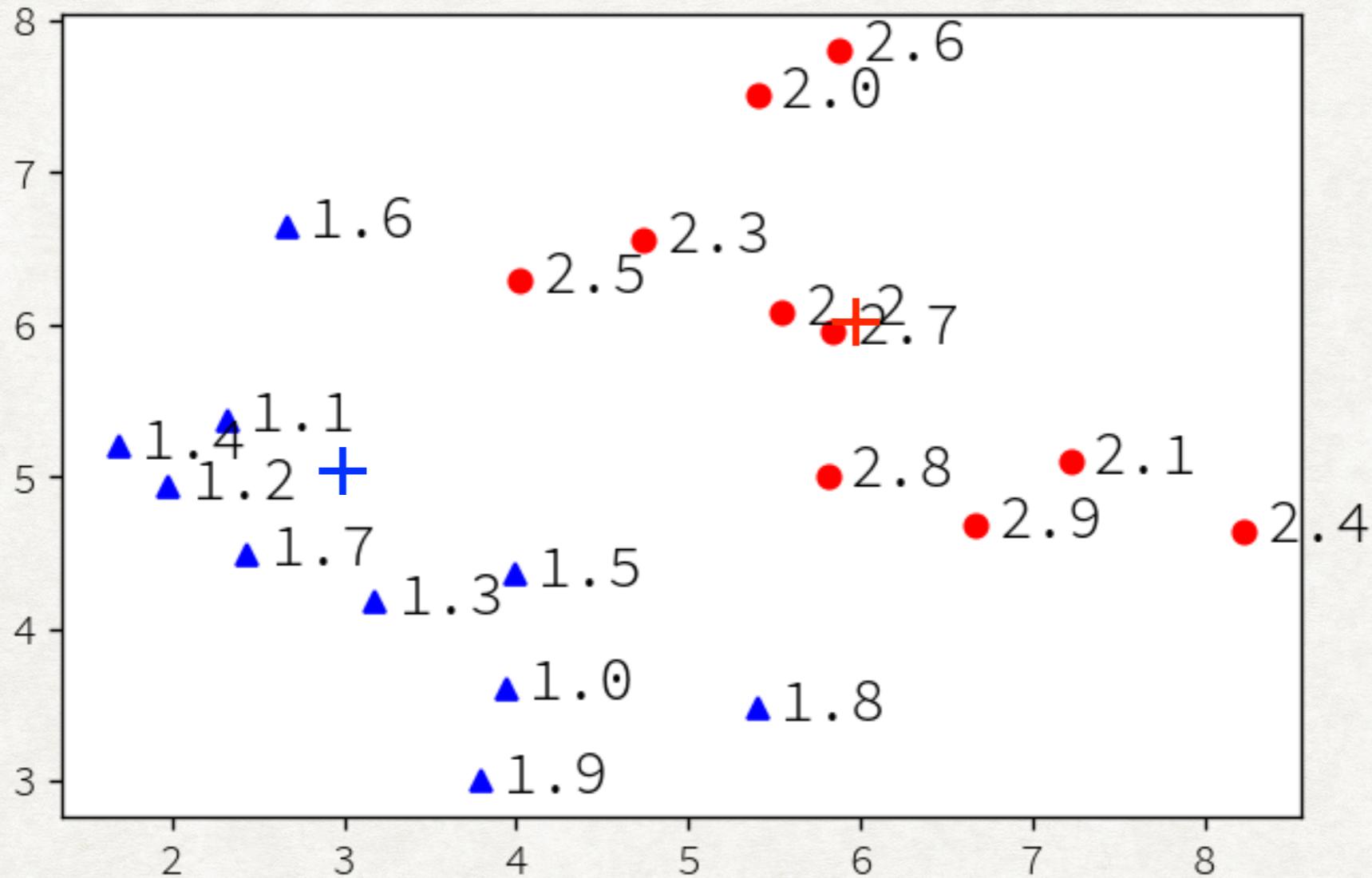
Python勉強会@HACHINOHE

- 標本をk個のクラスターに分類するタスク
 - 各標本は、一番近いクラスターの重心のクラスターに属す
 - クラスターの集合全体の相違性が最小化されている
- k平均法の計算方法: 貪欲法の一種
 - クラスターの重心の初期値として、k個の標本を選ぶ: 結果はこの選択により変動
 - くりかえす
 - それぞれの標本を最も重心が近いクラスターに所属させる $O(knd)$
 - d は標本間の距離を計算するのに必要な計算量のオーダーで次元に依存
 - クラスターの重心を計算しなおす $O(n)$
 - 個々の重心が変化したか調べ、変化しなければ終了 $O(k)$

標本

Python勉強会@HACHINOHE

- (3, 5)と(6, 6)を中心にガウス分布に沿う乱数で標本を生成



k平均法の結果と指練習

Python勉強会@HACHINOHE

- 1回の試行

- 2つの分布をうまく分離

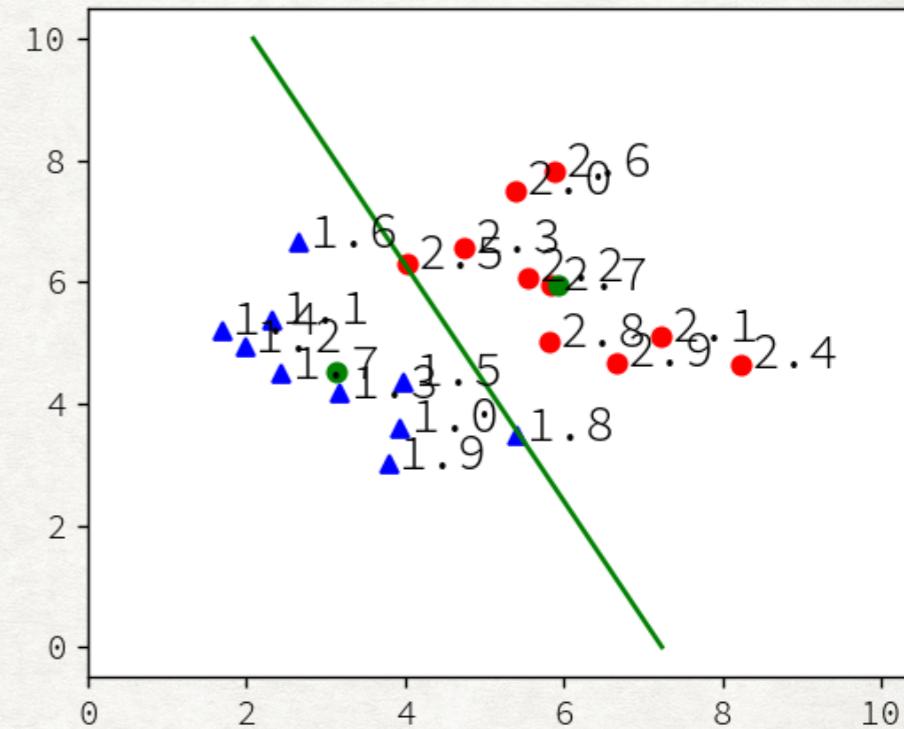
最終結果

クラスター：重心は[5.93613865 5.96069975]、要素：

2.0, 2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 2.7, 2.8, 2.9

クラスター：重心は[3.14170883 4.52143963]、要素：

1.0, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9



- 40回の試行

- 2つの分布を分離しないが、相違性は小さい

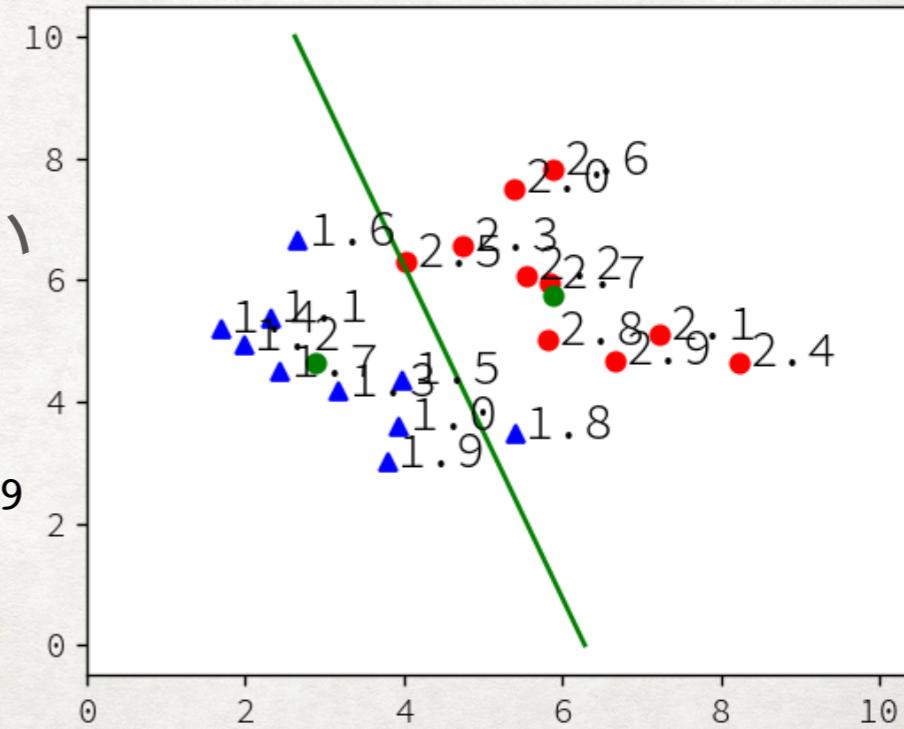
最終結果

クラスター：重心は[5.88777325 5.73417409]、要素：

1.8, 2.0, 2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 2.7, 2.8, 2.9

クラスター：重心は[2.89032989 4.63838655]、要素：

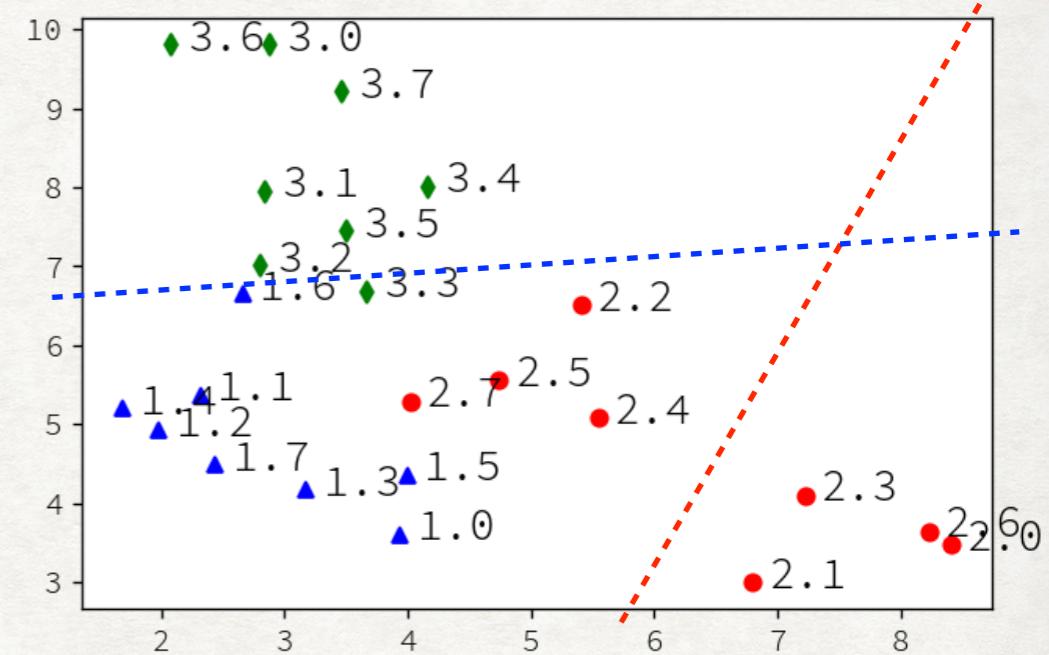
1.0, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.9



k平均法: 2つと3つに分類

Python勉強会@HACHINOHE

- (3, 5)、(6, 5)、(3, 8)を中心に生成
- 2つに分類
 - クラスター: 重心は[7.66239972 3.55222681]、要素: 2.0, 2.1, 2.3, 2.6
 - クラスター: 重心は[3.36736761 6.35376823]
要素: 1.0, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 2.2, 2.4, 2.5, 2.7, 3.0, 3.1, 3.2, 3.3, 3.4, 3.5, 3.6, 3.7
- 3つに分類
 - クラスター: 重心は[7.66239972 3.55222681]
要素: 2.0, 2.1, 2.3, 2.6
 - クラスター: 重心は[3.10687385 8.46084886]
要素: 3.0, 3.1, 3.2, 3.4, 3.5, 3.6, 3.7
 - クラスター: 重心は[3.50763348 5.21918636]
要素: 1.0, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 2.2, 2.4, 2.5, 2.7, 3.3



※図中の分離の直線は適当にひいている

k平均法: 6つに分類

Python勉強会@HACHINOHE

- (3, 5)、(6, 5)、(3, 8)を中心に生成
- 6つに分類

最終結果

クラスター: 重心は[7.66239972 3.55222681]

要素: 2.0, 2.1, 2.3, 2.6

クラスター: 重心は[2.80974427 9.60386549]

要素: 3.0, 3.6, 3.7

クラスター: 重心は[3.70472053 4.04178035]

要素: 1.0, 1.3, 1.5

クラスター: 重心は[2.10900238 4.99452866]

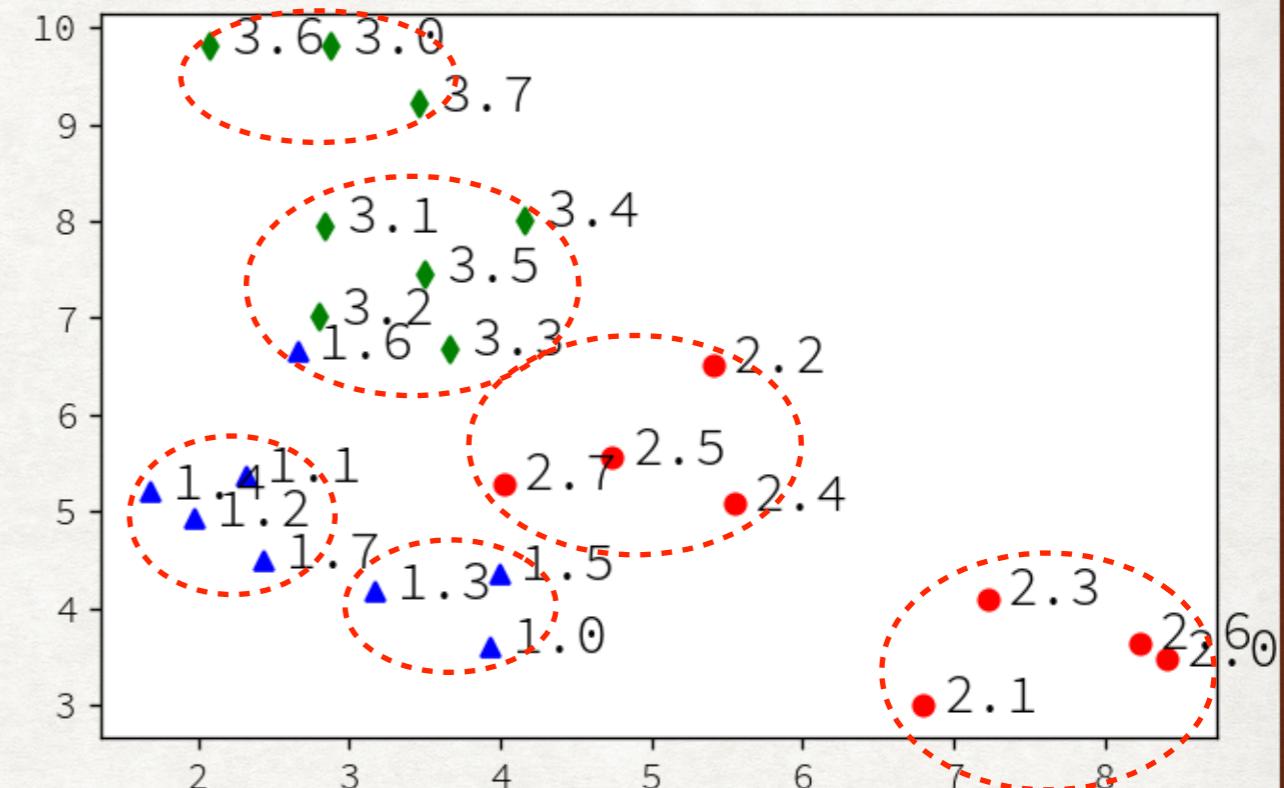
要素: 1.1, 1.2, 1.4, 1.7

クラスター: 重心は[4.92742554 5.60609442]

要素: 2.2, 2.4, 2.5, 2.7

クラスター: 重心は[3.27637435 7.28932247]

要素: 1.6, 3.1, 3.2, 3.3, 3.4, 3.5



いくつに分類するべきか?
クラスターの個数を増やすと過剰適合
が起こる。

\

k平均法: 草食、肉食、雑食の単純な分類

Python勉強会@HACHINOHE

- 特徴ベクトル
 - 上前歯、上犬歯、上前臼歯、上臼歯、下前歯、下犬歯、下前臼歯、下臼歯、平均体重
- 単純な分類結果: 平均体重で分類されているっぽい
ウシ、エルク、ムース、アシカ
草食動物 3 、肉食動物 1 、雑食動物 0

アナグマ、クーガー、イヌ、キツネ、モルモット、ジャガー、カンガルー、ミンク、モグラ、マウス、ヤマアラシ、ブタ、ウサギ、アライグマ、ラット、コウモリ、スカンク、リス、ウッドチャック、オオカミ
草食動物 4 、肉食動物 9 、雑食動物 7

クマ、シカ、オットセイ、ハイイロアザラシ、ヒト、ライオン
草食動物 1 、肉食動物 3 、雑食動物 2

k平均法: 草食、肉食、雑食のスケーリングした分類

Python勉強会@HACHINOHE

- 特徴の大きさが、個々に異なるので、平均0、標準偏差1になるようにスケーリング
- スケーリングなし

ウシ、エルク、ムース、アシカ

草食動物 3、肉食動物 1、雑食動物 0

アナグマ、クーガー、イヌ、キツネ、モルモット、ジャガー、カンガルー、ミンク、モグラ、マウス、ヤマアラシ、ブタ、ウサギ、アライグマ、ラット、コウモリ、スカンク、リス、ウッドチャック、オオカミ

草食動物 4、肉食動物 9、雑食動物 7

クマ、シカ、オットセイ、ハイイロアザラシ、ヒト、ライオン

草食動物 1、肉食動物 3、雑食動物 2

- スケーリングあり

アナグマ、クマ、クーガー、ウシ、シカ、イヌ、キツネ、オットセイ、ハイイロアザラシ、エルク、ヒト、ジャガー、ライオン、ミンク、モグラ、ムース、ブタ、アライグマ、コウモリ、アシカ、スカンク、オオカミ

草食動物 4、肉食動物 13、雑食動物 5

カンガルー、ウサギ、リス、ウッドチャック

草食動物 2、肉食動物 0、雑食動物 2

モルモット、マウス、ヤマアラシ、ラット

草食動物 2、肉食動物 0、雑食動物 2