勾配ブースティングでも変わらない決定木のお気持ち

1. 決定木系アルゴリズムにおける特徴量エンジニアリング
2. 簡単な実験を通して決定木の気持ちに触れる
3. 考察とまとめ

イントロ

機械学習における特徴量エンジニアリング

- 機械学習の活用において、最も重要かつ、最も奥深い領域
- 「データサイエンスにおける匠の技」と言及する人もいる
- 過去のすいすい会でも発表済み
  - 第４回「自動機械学習での特徴量の作り方」
    （https://www.youtube.com/watch?v=Ms52EnCRk8g）

決定木系アルゴリズムでの特徴量エンジニアリング

Kaggler※の間では以下のような名言が知られている

※Kaggler = 世界的に有名なコンペサイトKaggleの最前線で戦うデータサイエンティストたち

「決定木の気持ちになって考える」

でも、決定木の”お気持ち”ってなんなの？

でも、決定木の”お気持ち”ってなんなの？

Q. 相当経験がないと気持ちを察するのは難しいんじゃ、、、

A. そんなことはありません

例えば、すでにある特徴量同士を四則演算しただけで
決定木が一気に学習しやすくなる事も

▼

簡単な実験を通して決定木の”お気持ち”に触れてみる

注意点

- あくまでも決定木系アルゴリズムの挙動を確認するための簡単な実験

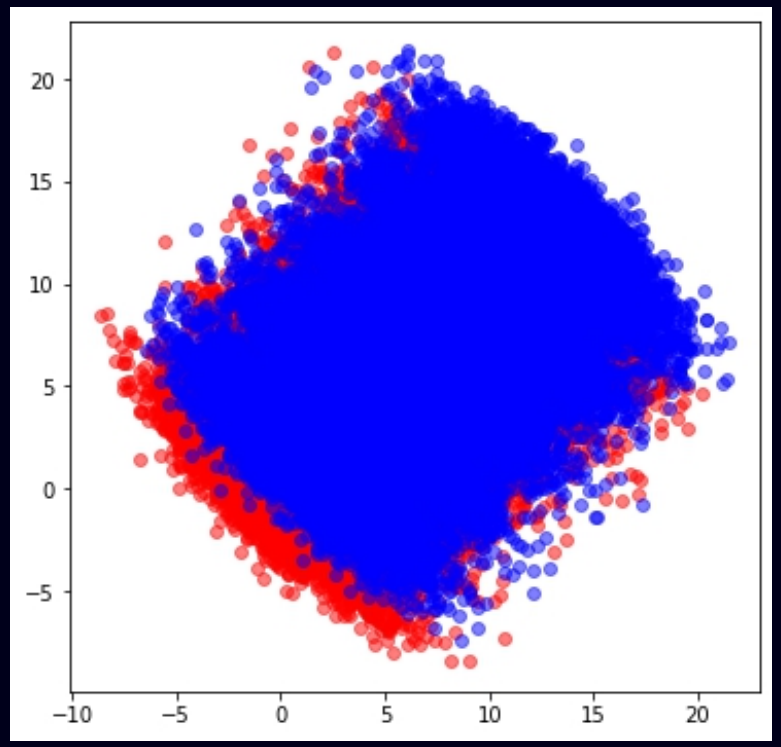
- ハイパーパラメータはForecstFlowで最適化
  - 今回の実験に特化したチューニングではない

- 機械学習に関する知識をある程度前提

実験

（恣意的ではあるが）以下のようなシチュエーションを考えてみる

y



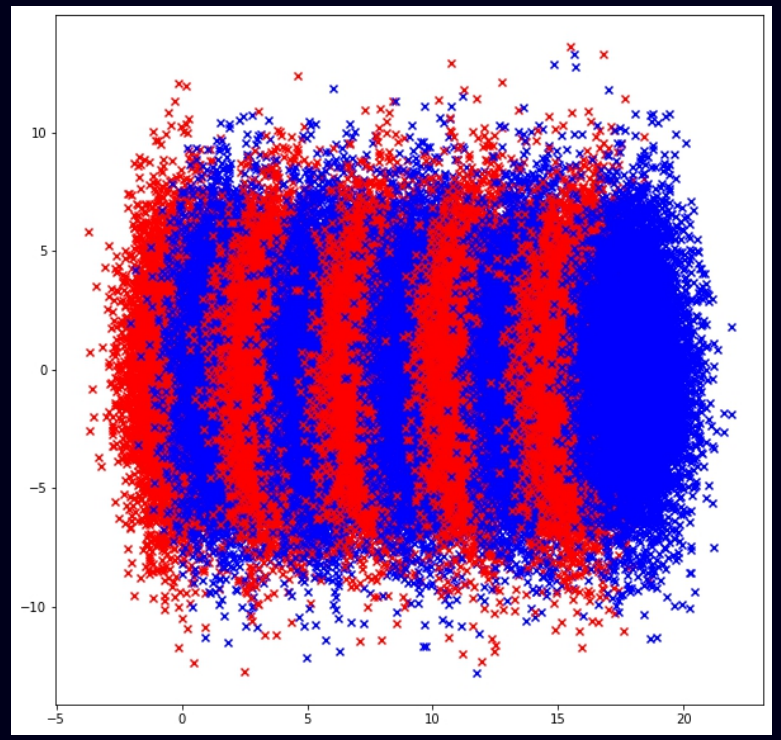
x

・特徴量：x、yの2種類
・target（正解ラベル）：rとb（2値分類）
・r５万, b5万の計10万レコード

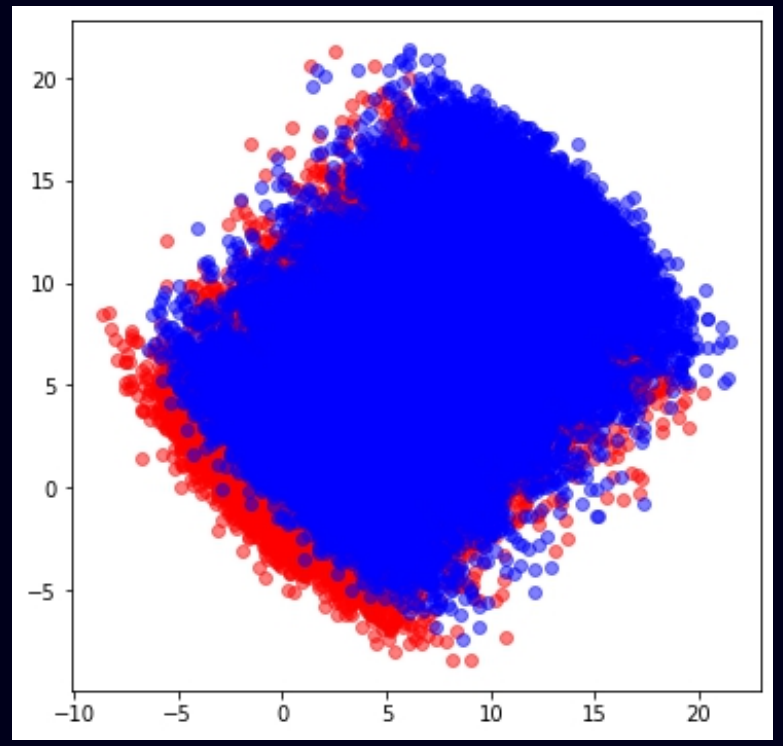
x	y	target
-0.91	-1.74	r
-1.02	-0.20	r
-0.09	-1.02	r
9.43	-2.69	b
9.32	-4.83	b
8.34	-1.64	b

・
・
・

（恣意的ではあるが）以下のようなシチュエーションを考えてみる



反時計回りに45°回転



y

x

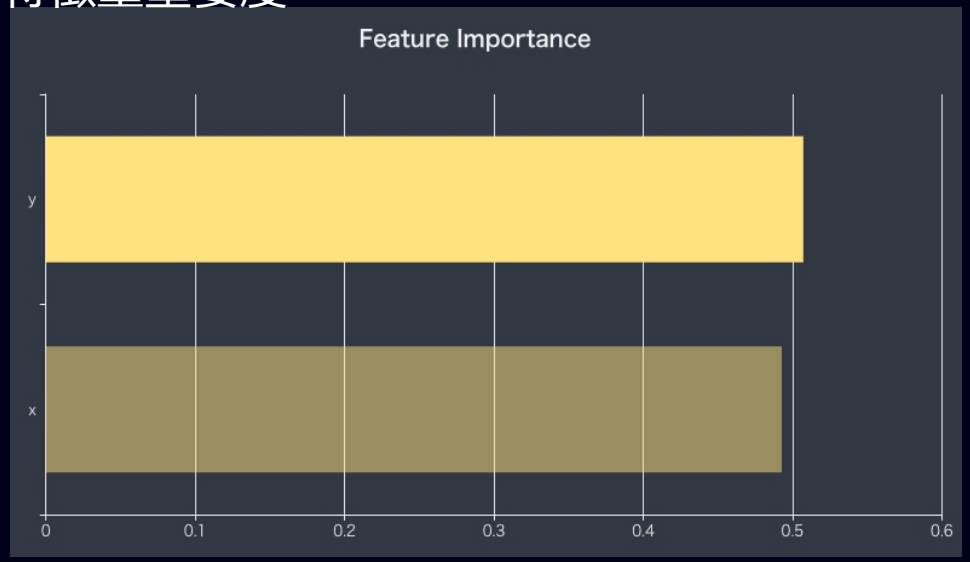
中心を変えた縦に長い2次元正規分布を準備
正解ラベルが交互になるように10グループ

ForecastFlowに入れて訓練→精度確認

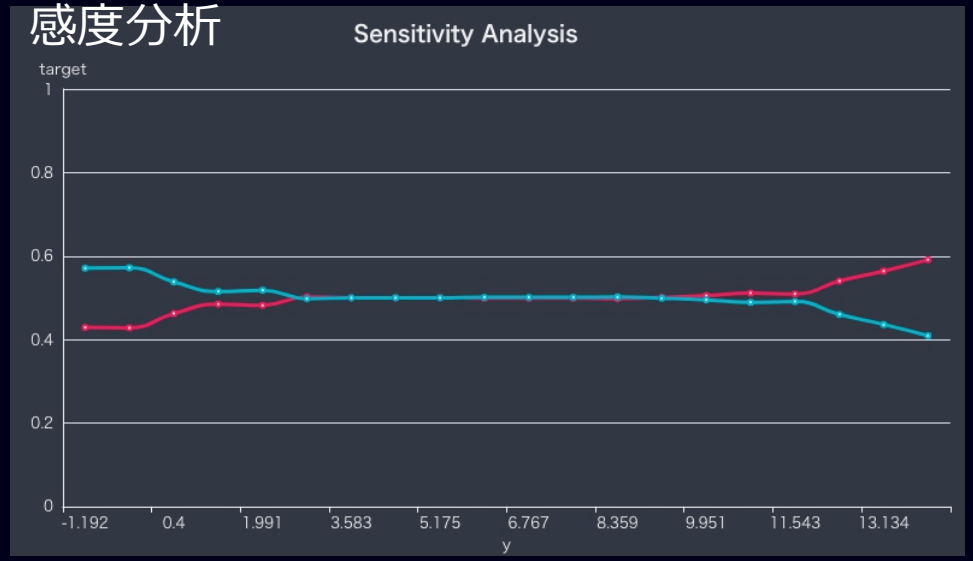
精度

Label Name	F1 Score
r	0.564
b	0.548

特徴量重要度



感度分析

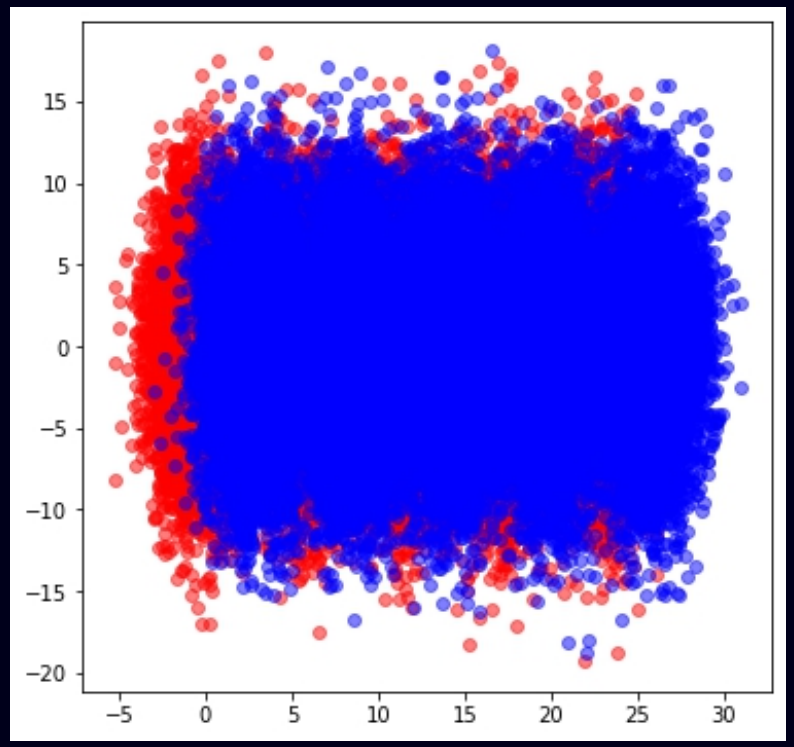


精度は良くない、、、

【理由】
（誤解を恐れずにいうと）
決定木は斜めの線を引くことが苦手だから

xとyを四則演算した特徴量を追加

x-y



x+y

新たに追加した特徴量

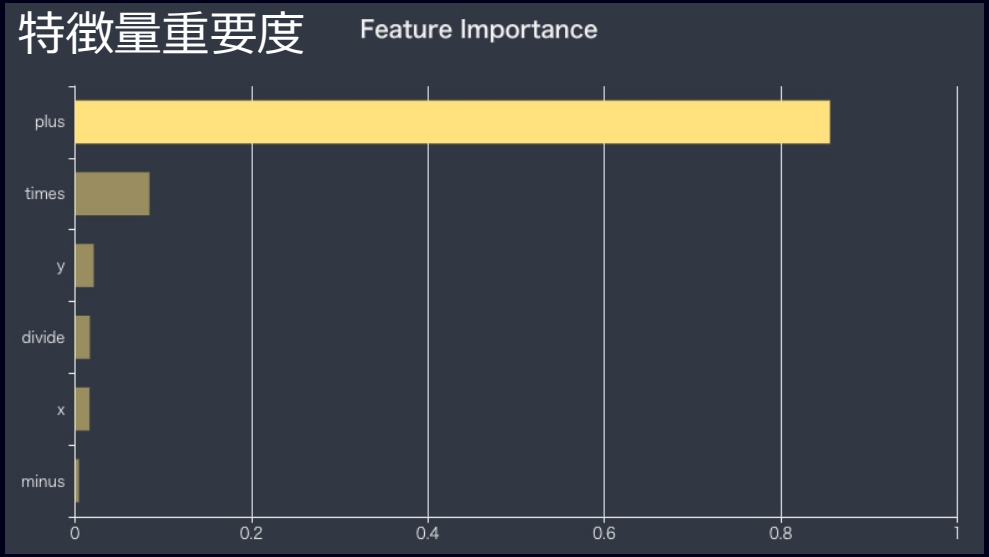
x	y	x+y	x-y	xy	x/y	target
-0.91	-1.74	-2.65	0.82	1.59	0.53	p
-1.02	-0.20	-1.23	-0.82	0.21	4.99	p
-0.09	-1.02	-1.11	0.93	0.09	0.09	p
9.43	-2.69	6.74	12.11	-25.3	-3.51	n
9.32	-4.83	4.48	14.15	-45.0	-1.93	n
8.34	-1.64	6.70	9.98	-13.6	-5.08	n

ForecastFlowに入れて訓練→精度確認

精度

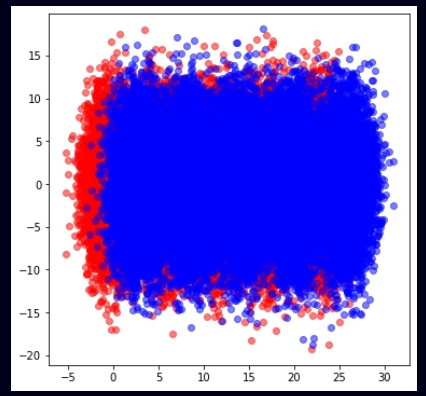
Label Name	F1 Score
r	0.717
b	0.719

特徴量重要度

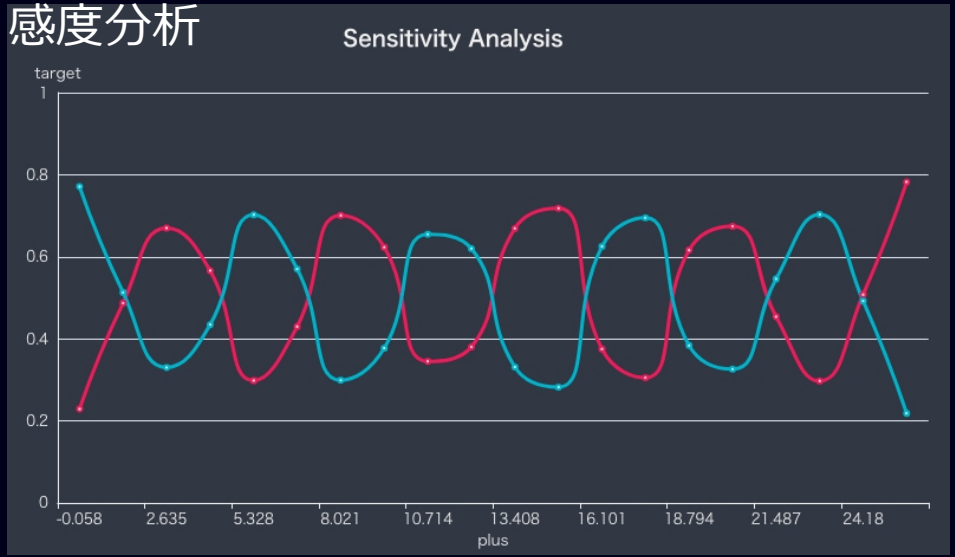


精度が劇的に改善！！

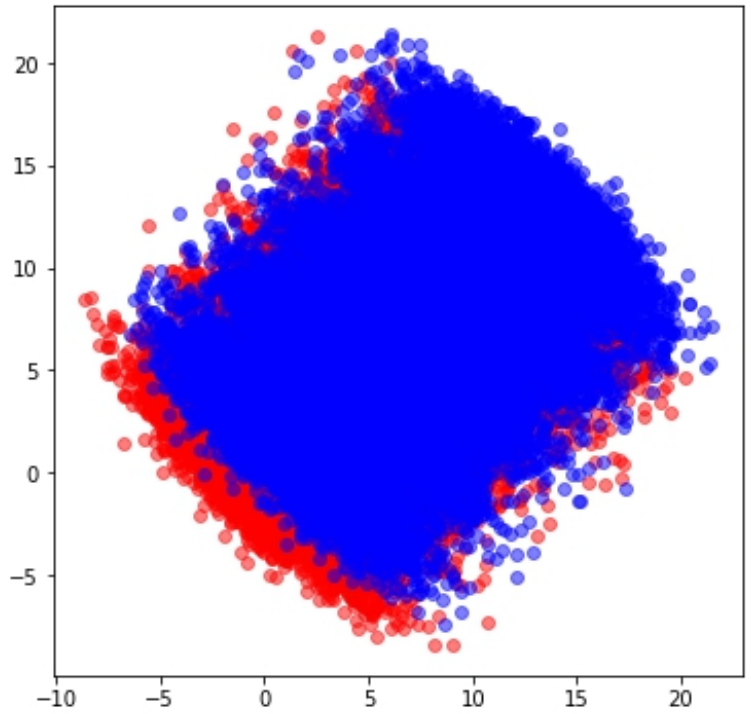
【理由】
x+yで見ると縦に分割線を
引けるようになったから

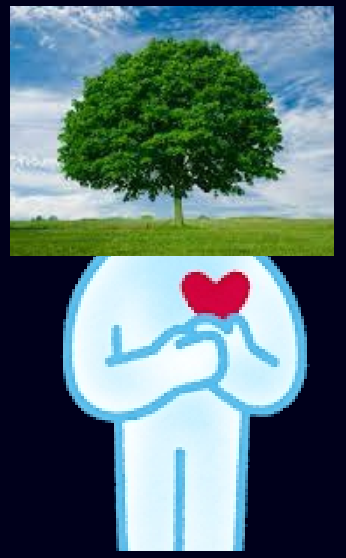


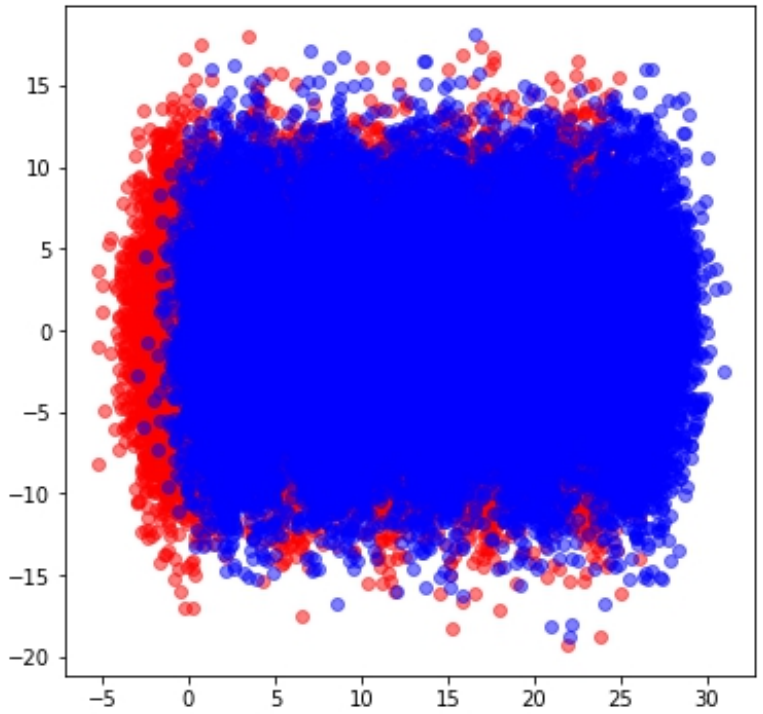
感度分析



決定木の”お気持ち”



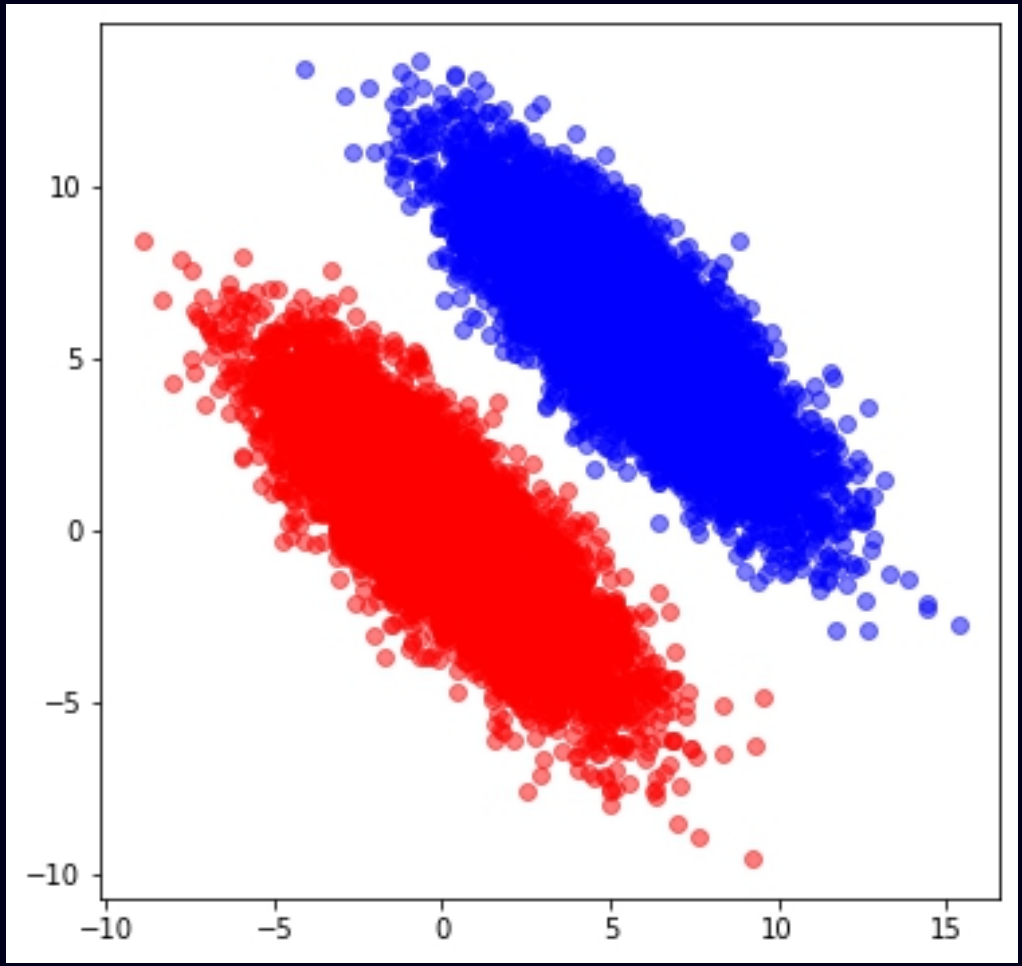




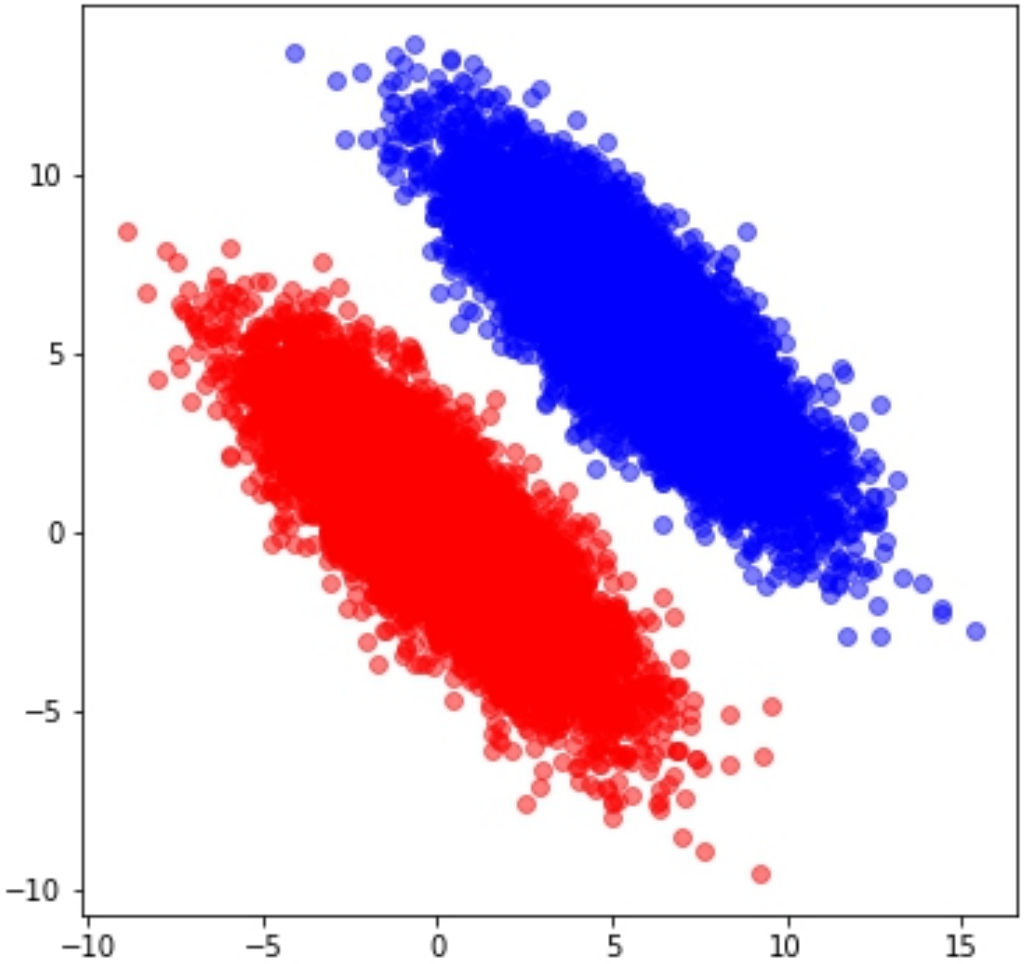
理屈の説明

さらに単純化した以下のシチュエーションを考える

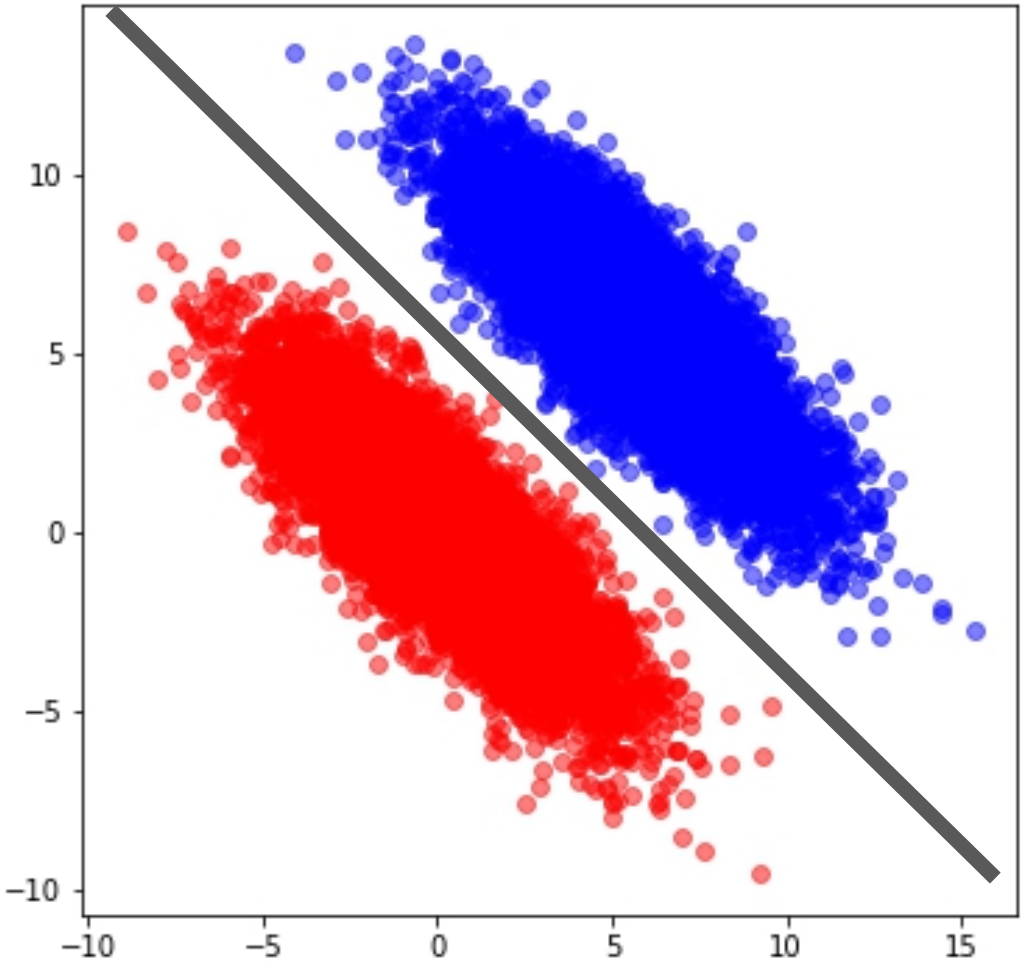
- この時決定木はどうやってこの2つのグループを分類するか？



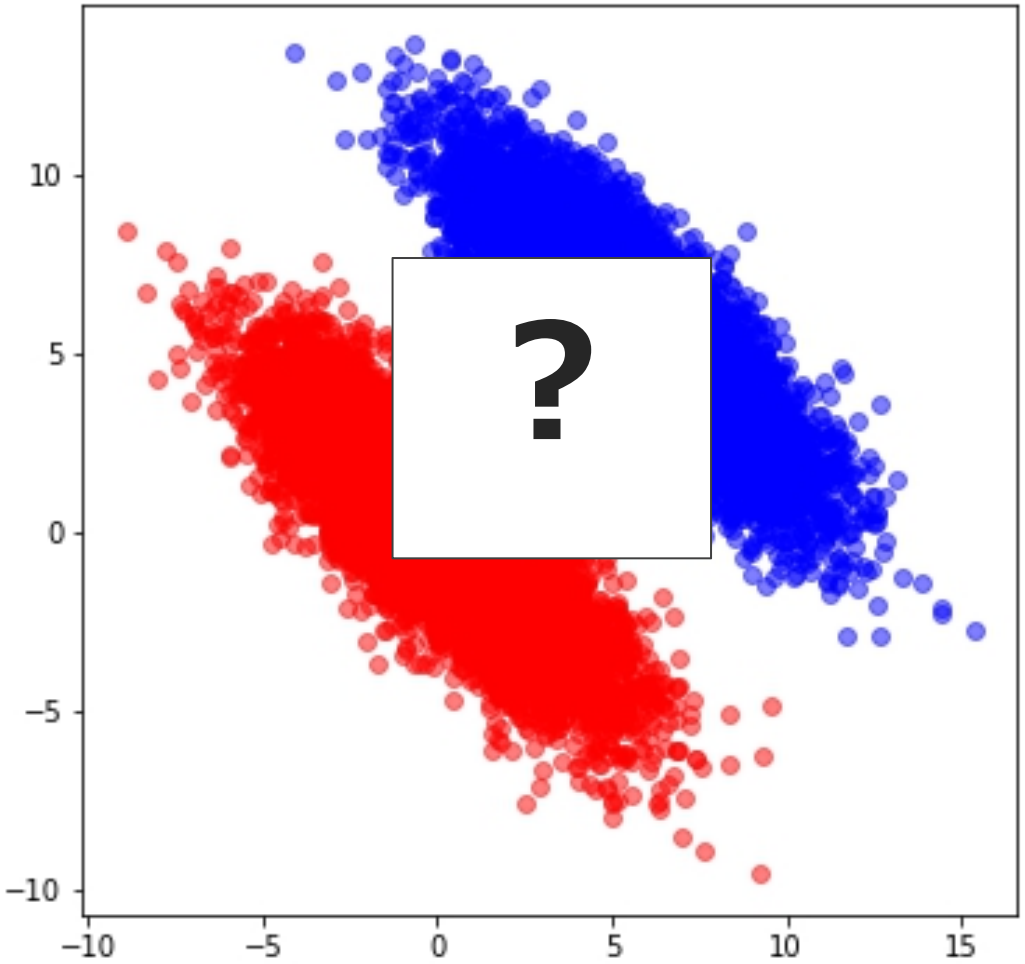
人間の場合は、、、



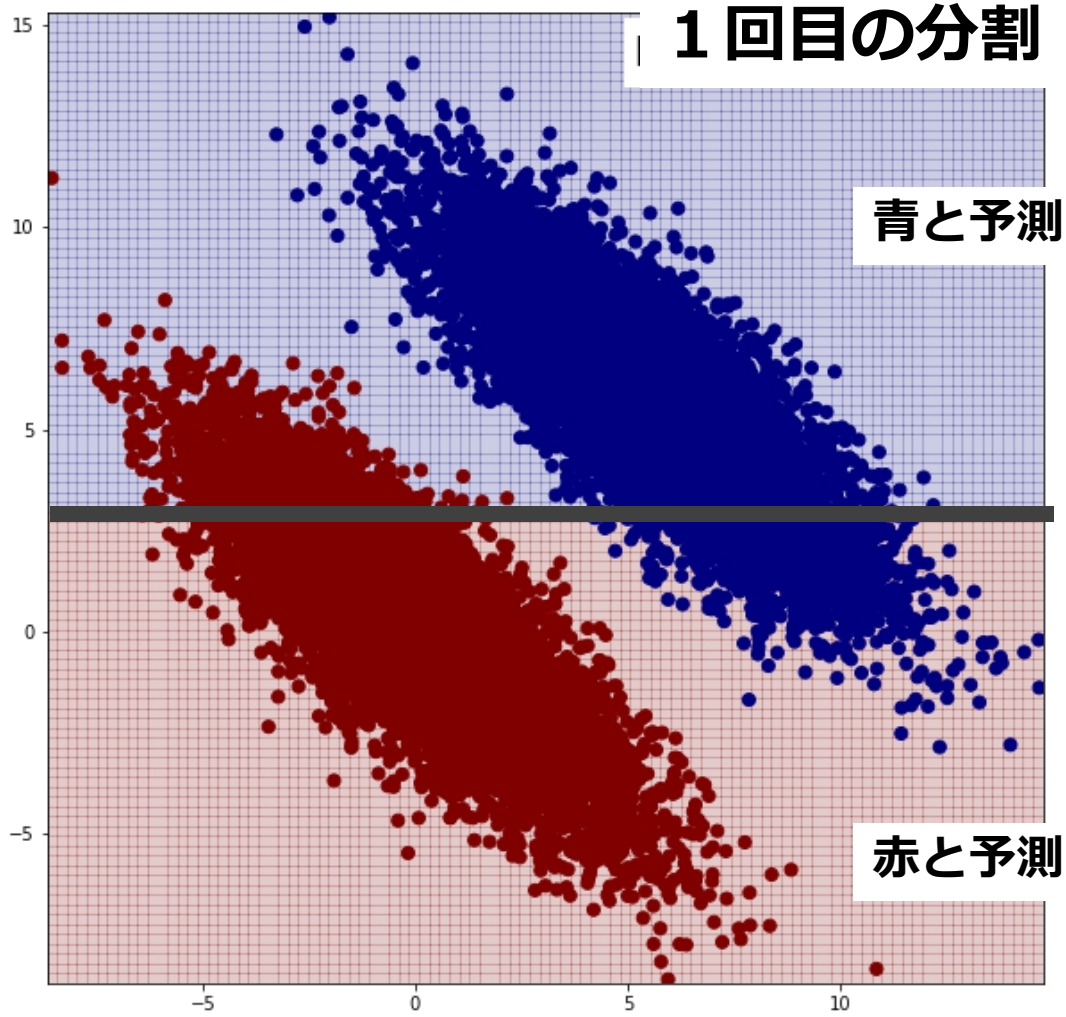
人間の場合は、、、



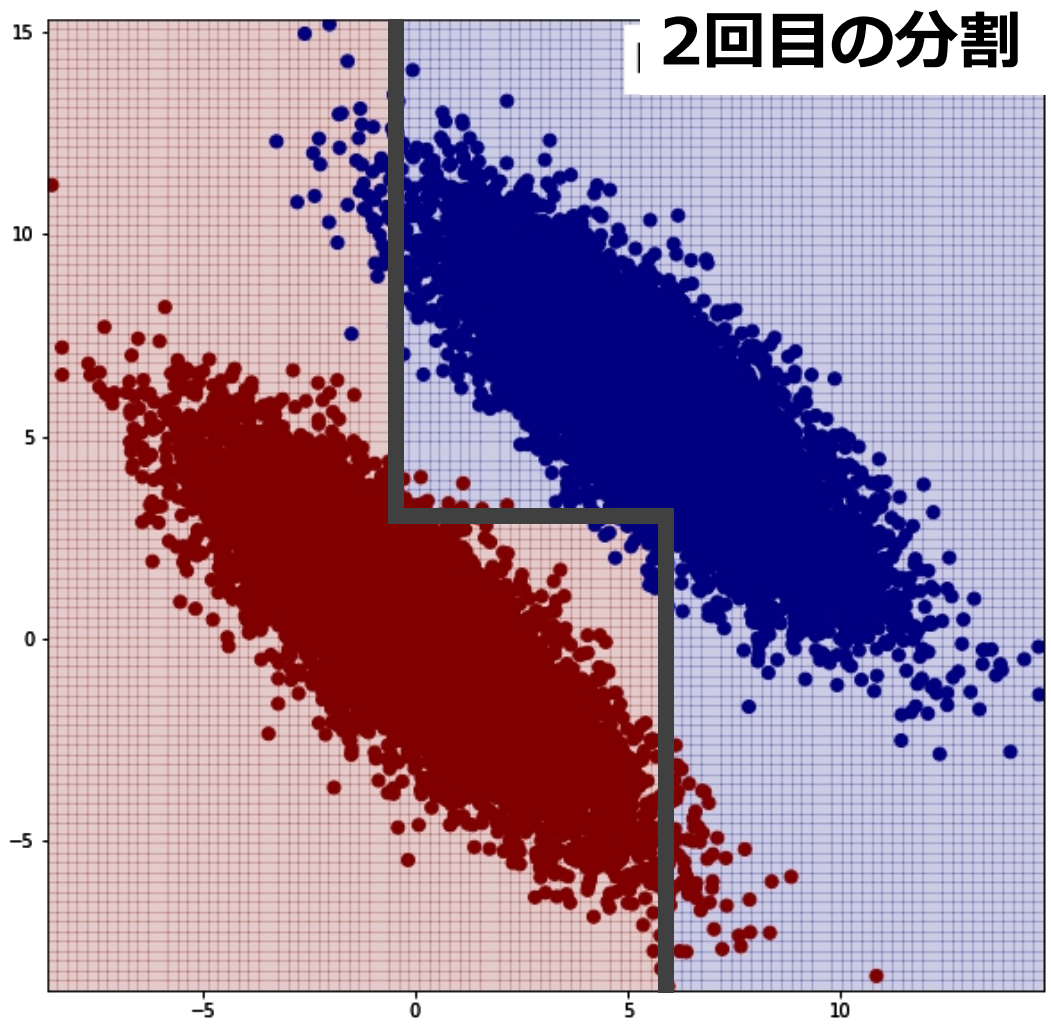
決定木だと、、、



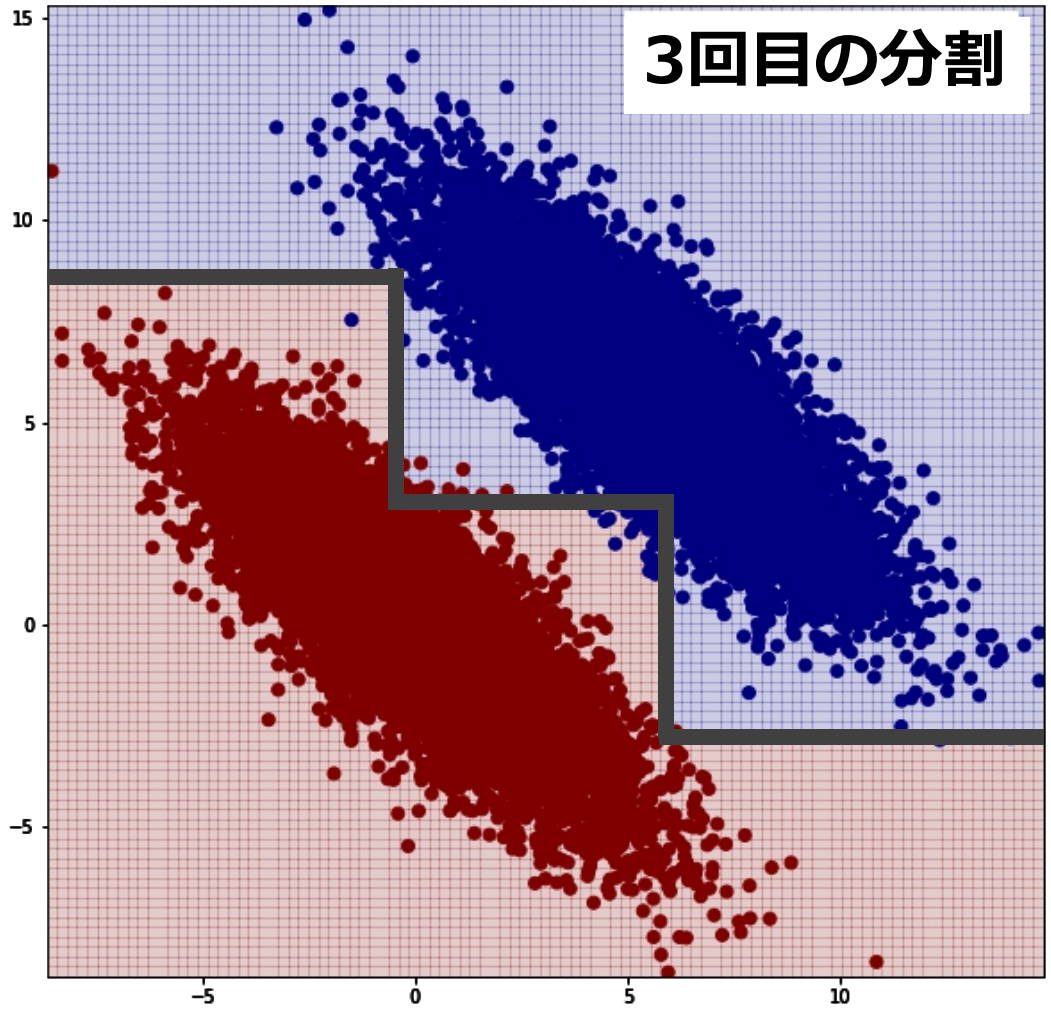
決定木だと、、、



決定木だと、、、

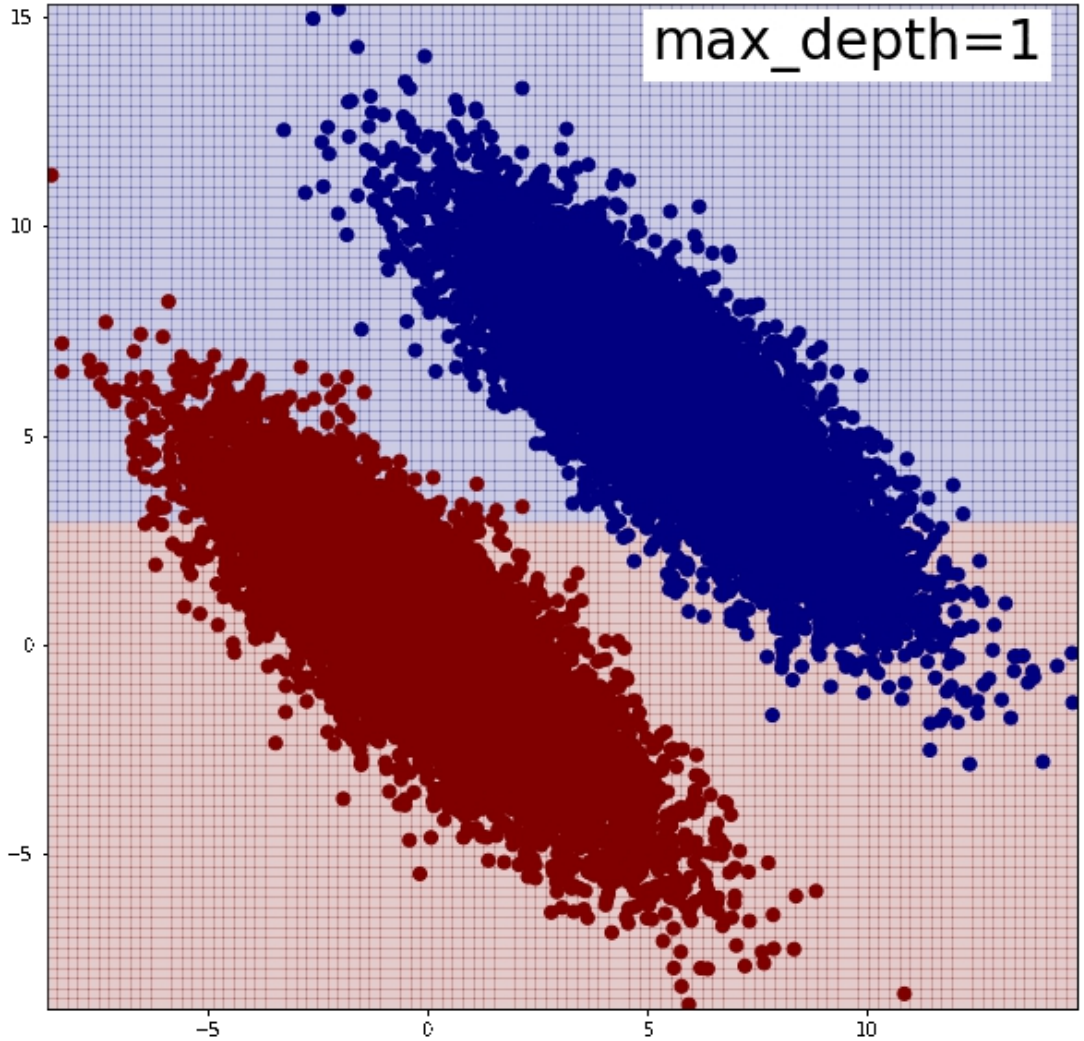


決定木だと、、、

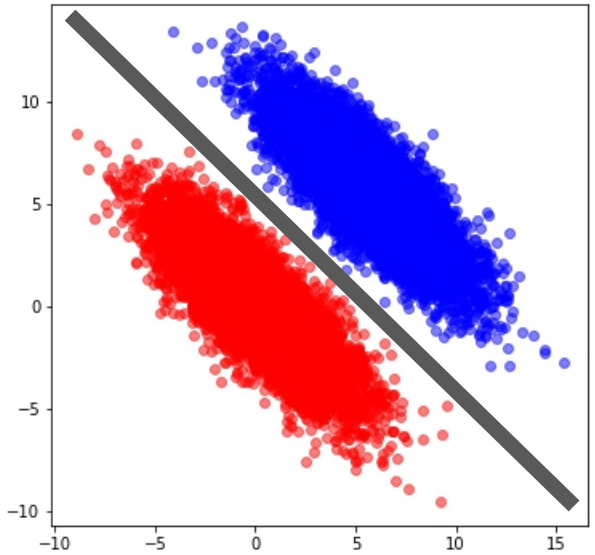


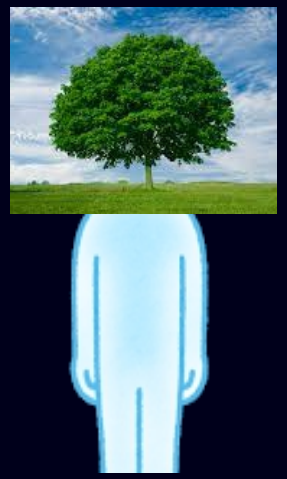
決定木だと、、、

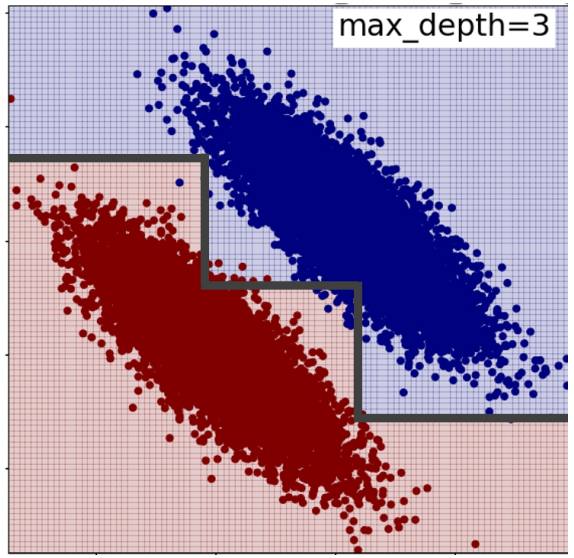
分割の進み方



結局、人間と決定木とではみている世界が違う







人間のように斜めの線を引くことはできない

→明示してあげる必要がある（気持ちを察する）

考察とまとめ

じゃあ全ての特徴量の組み合わせを追加すればいいのでは？

現実的ではない、、、（非推奨）

元の特徴量数	追加される特徴量数
50	5,000
100	20,000
1,000	2,000,000

・「2万個の特徴量からいい感じの組み合わせを見つけといて」
と言われる決定木の気持ちになって考える
（実際、特徴量が増えすぎると訓練が進まない）
・もし仮に効いたとして、理解不能な特徴量を施策に活かせるとは限らない
（例）コールセンターへの電話回数÷入会した月

四則演算が効果的になりそうなケース

■足し算
- 深夜のWebサイト回遊行動（０時台の行動＋1時台の行動＋2時台の行動）

■引き算
- サービス入会から現在までの経過日数（現在の日付-入会日）

■割り算
- 一回あたりの購買金額（合計購買金額/購買回数）

■掛け算
- 合計購買金額（1回あたりの購買金額×購買回数）

四則演算が効果的になりそうなケース

■足し算
- 深夜のWebサイト回遊行動（0時台の行動＋1時台の行動＋2時台の行動）

■引き算
- サービ

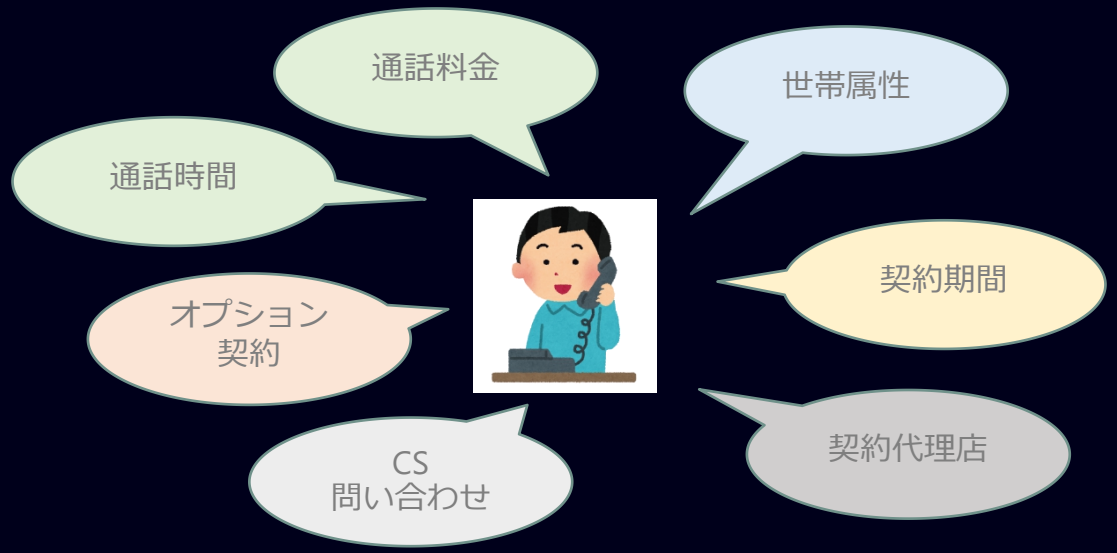
■割り算
- 一回あ

■掛け算
- 合計購買金額（1回あたりの購買金額×購買回数）

結局、現場の人の仮説を元に追加していくのが近道だったりする

【ユーザに関するドメイン知識】特徴量エンジニアリングの考え方①

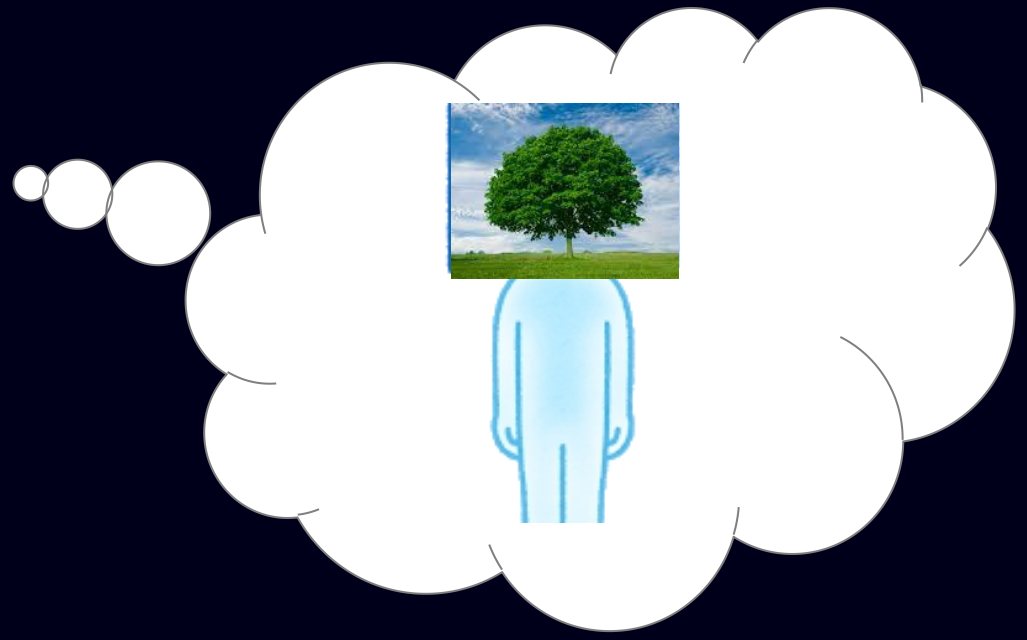
- 特徴量エンジニアリングでは、例えば顧客になりきり、顧客の気持ちを想像しながら、「ユーザはこんな気持ちになるはずだから、こんな人はこんな行動をするはずだから、この特徴量を入れよう」と進めていきます
- チェックする基準
  - ユーザ属性：　性別、年齢
  - ユーザ行動：　利用料、アクセス数



継続／解約を上手く説明できる要因は何だろうか？

まとめ

- 決定木並びにその進化系であるLightGBMやランダムフォレストは、十分賢いアルゴリズムとはいえ、苦手なパターンがある
- 特徴量の組み合わせなどがその例で、新たな特徴量として明示する必要
- かといって自動的に入れると、訓練時間が長くなりすぎる、解釈不能になる、などの問題が生じる可能性
- 分析官は決定木のお気持ちとユーザーのお気持ちを考えることが大事



今回話さなかったこと

- カテゴリ変数の組み合わせ
- 特徴量を増やして行った時の弊害
- 特徴量の削減