AIエンジニアリング　(4)

線形分類器(おさらい)・非線形分類器

機械学習とはなんだろう

機械学習とは

- 実は機械学習の定義もあまりはっきりしたものはない
- ただ、一般的に機械学習は次の3種類からなるので、この総称といっていい
  1. 教師あり学習
  2. 教師なし学習
  3. 強化学習

1.教師あり学習(supervised learning)

1. 人間がお手本(教師データ)を用意してあげる
2. 教師データをもとにして、人間と同じように判断できるまで学習する

- 例
  - 迷惑メール(スパムメール)フィルタ
  - 文字認識
  - 画像認識
  - 音声認識

2.教師なし学習(unsupervised learning)

1. 人間がデータを用意する
2. データからなにかしらの法則などを見つけ出す

- 例
  - データマイニング系
  - クラスタリングなど
- 機械学習の中では下火だったが、最近アツい分野

3.強化学習(reinforcement learning)

1. 直接お手本は用意しないが、人間がゴールだけ決める
2. あとは試行錯誤して勝手にうまくなっていく

- 例
    - ゲームAI
        - 囲碁、将棋、チェスなど
    - ロボット制御
- 一見便利そうに思えるが、学習がすごく難しいので教師あり学習のほうが便利なことが多い

教師あり学習を体験してみよう

Boston Dataset

- ボストンの住宅価格を予想する問題
- 犯罪発生数や部屋の平均数など14の項目から予想する

項目名	説明
CRIM	人口 1 人当たりの犯罪発生数
ZN	25,000 平方フィート以上の住居区画の占める割
INDUS	小売業以外の商業が占める面積の割合
CHAS	1:チャールズ川が近い 0:川が近くない
NOX	NOx の濃度
RM	住居の平均部屋数
AGE	1940 年より前に建てられた物件の割合
DIS	5つのボストン市の雇用施設からの距離
RAD	環状高速道路へのアクセスしやすさ
TAX	$10,000 ドルあたりの不動産税率の総計
PTRATIO	町毎の児童と教師の比率
B	町毎の黒人 (Bk) の比率を次の式で表したもの。
LSTAT	給与の低い職業に従事する人口の割合 (%)

どうやって予想するか？

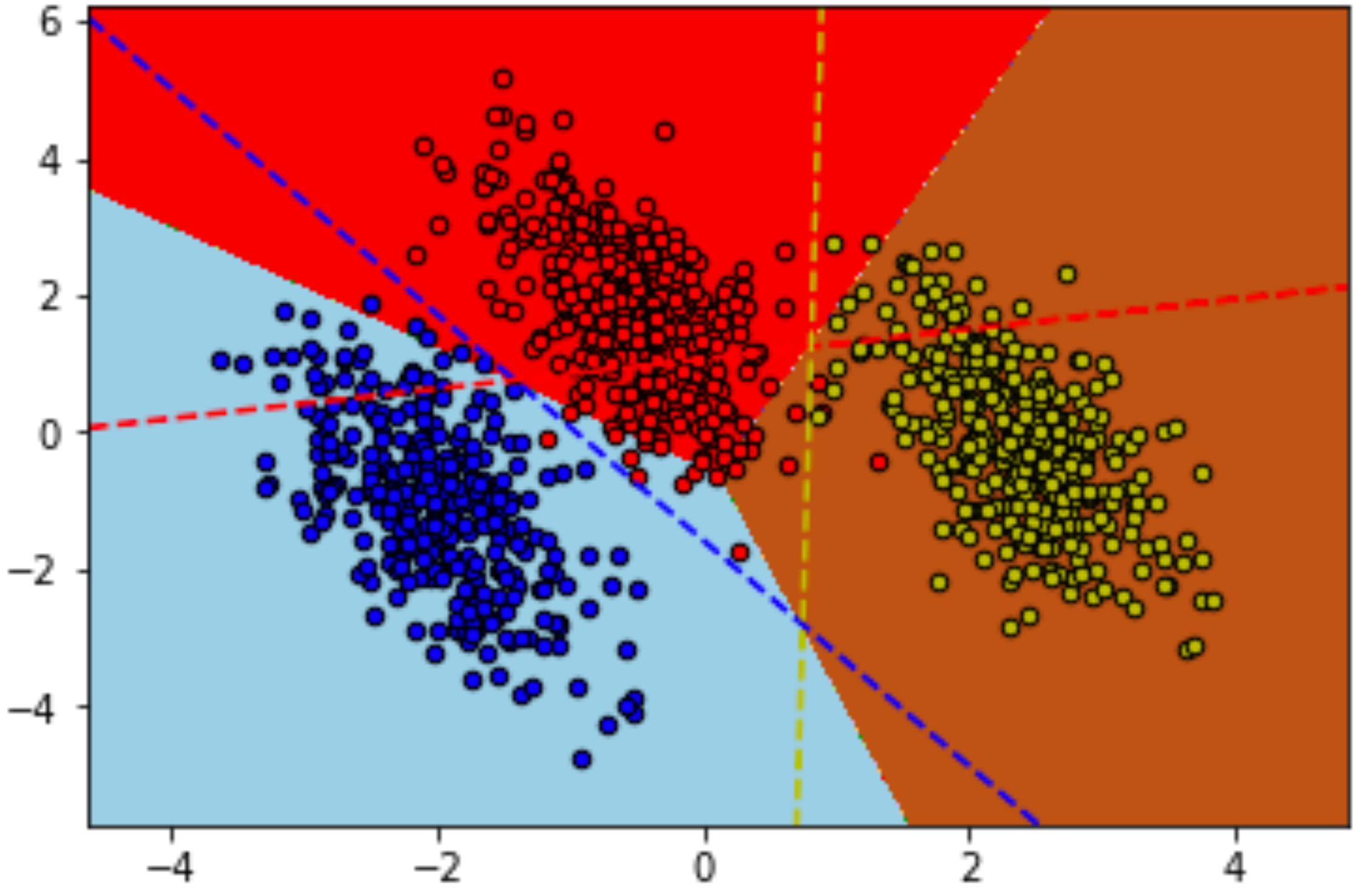
- とりあえず次のような式を考えて、そこから予想しよう

- A * CRIM + B * ZN+ … + M * LSTAT + N = 住宅価格
- みたいなのを考えて、A〜Nまでの値をいい感じに調整すれば、住宅価格を予想する式ができそう
- こういうのをモデルという
  - たとえば「今回は住宅価格を予測するためにこのようなモデルを考えます」のように言う

どうやって予想するか？

- 今はとりあえず各属性にそれぞれ1個ずつ数字を掛け算するようなモデルを考えた
- 特に正解はないので好きなモデルを考えれば良い
- ただ、この形式のモデルは一番単純なのに一番奥が深いのでよく使われる
- こいつはよく使うので特別に名前がついていて、線形モデル(線形回帰モデル)という

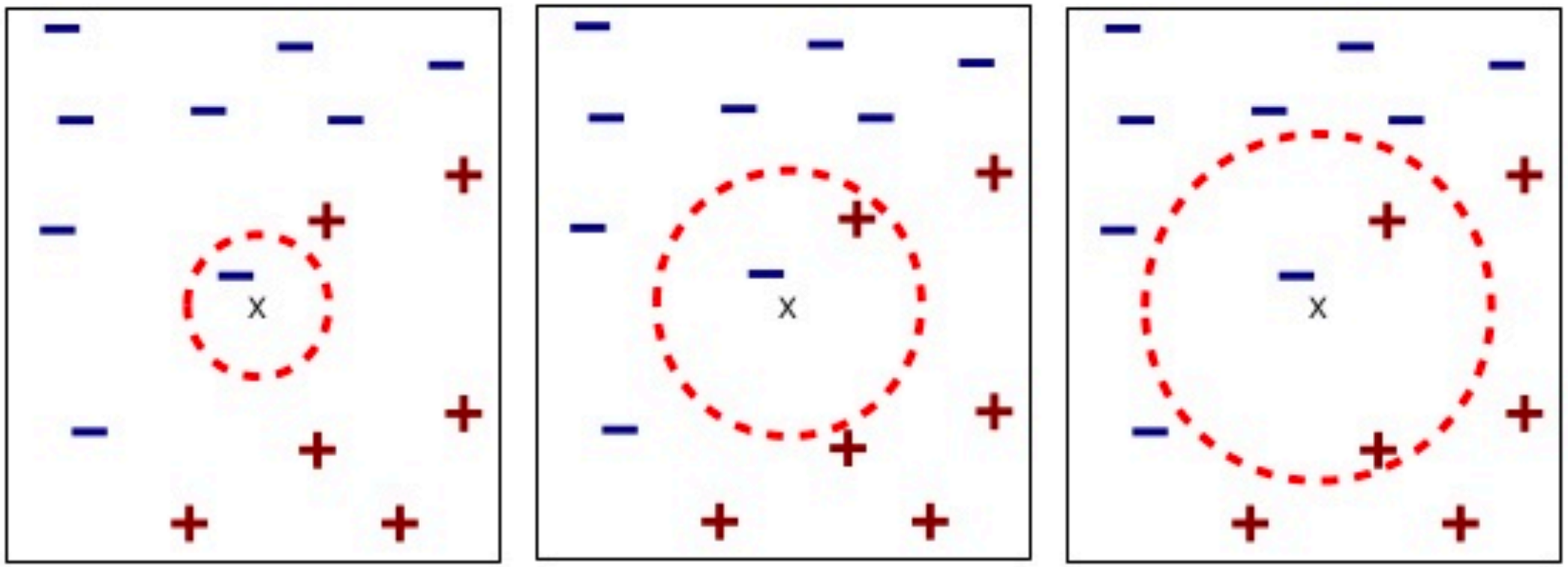
分類問題(線形分類)



非線形のものも扱いたい

- 世の中のデータは非線形なものが多い
- Random Forest 系を使うことが多い

k-NN (k=1,2,3)



(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

普通はkは奇数にする

Random Forest 系

- 2001年登場→最近の主流
- scikit-learn に実装済みかつ有名なものは3つ
  - Random Forest (元祖)
  - Extremely Randomized Trees (Extra Trees)
  - Gradient Boosting Decision Trees (GBDT)
- その他色々

Wine dataset

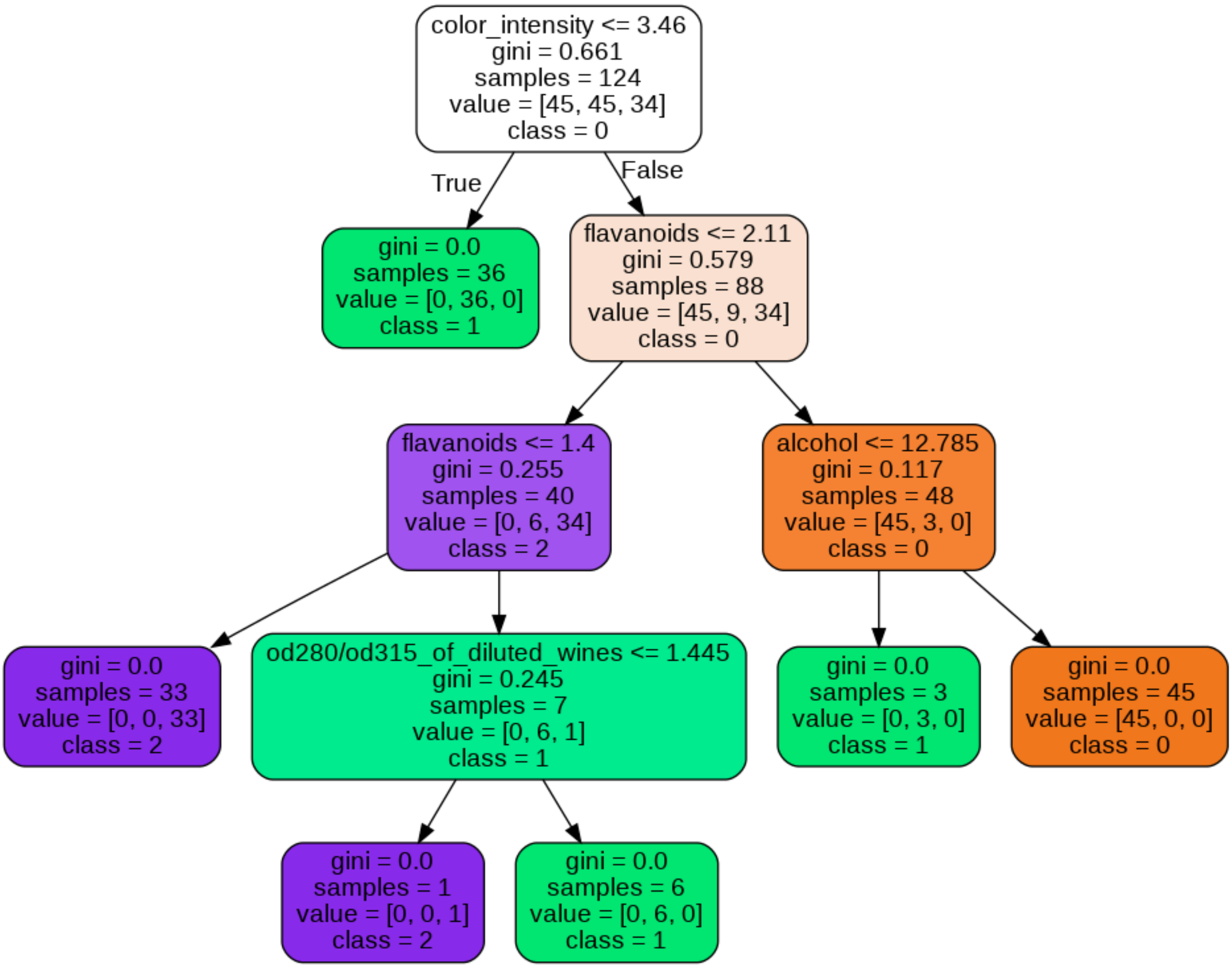
説明変数

1. alcohol アルコール濃度
2. malic_acid リンゴ酸
3. ash 灰（？
4. alcalinity_of_ash 灰のアルカリ成分（？
5. magnesium マグネシウム
6. total_phenols 総フェノール類量
7. flavanoids フラボノイド（ポリフェノールらしい）
8. nonflavanoid_phenols 非フラボノイドフェノール類
9. proanthocyanins プロアントシアニジン（ポリフェノールの一種らしい）
10. color_intensity 色の強さ
11. hue 色合い
12. od280/od315_of_diluted_wines ワインの希釈度合い
13. proline プロリン（アミノ酸の一種らしい）

目的変数

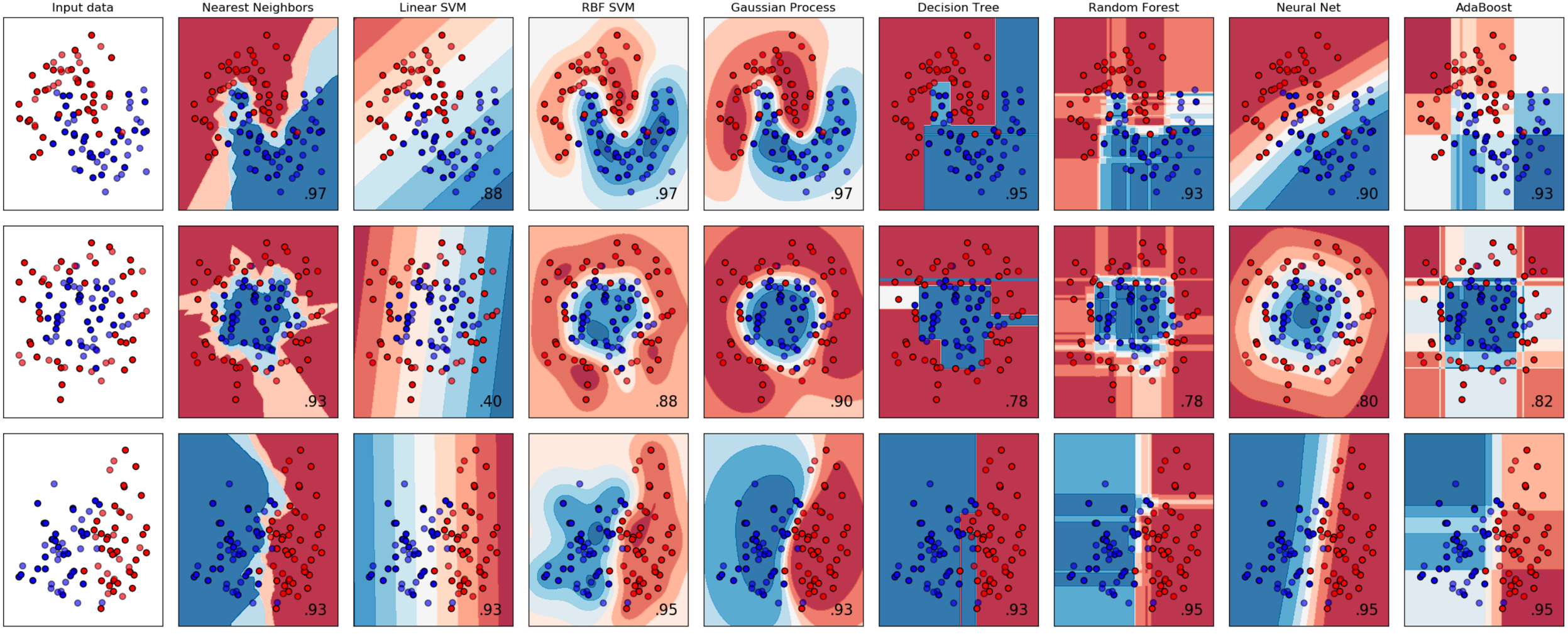
14. ワインの品種

	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavanoids	nonflavanoid_phenols	proanthocyanins	color_intensity	hue	od280/od315
0	14.23	1.71	2.43	15.6	127.0	2.80	3.06	0.28	2.29	5.64	1.04	
1	13.20	1.78	2.14	11.2	100.0	2.65	2.76	0.26	1.28	4.38	1.05	
2	13.16	2.36	2.67	18.6	101.0	2.80	3.24	0.30	2.81	5.68	1.03	
3	14.37	1.95	2.50	16.8	113.0	3.85	3.49	0.24	2.18	7.80	0.86	
4	13.24	2.59	2.87	21.0	118.0	2.80	2.69	0.39	1.82	4.32	1.04	


color_intensity <= 3.46
gini = 0.661
samples = 124
value = [45, 45, 34]
class = 0
True
False
gini = 0.0
samples = 36
value = [0, 36, 0]
class = 1
flavanoids <= 2.11
gini = 0.579
samples = 88
value = [45, 9, 34]
class = 0
flavanoids <= 1.4
gini = 0.255
samples = 40
value = [0, 6, 34]
class = 2
alcohol <= 12.785
gini = 0.117
samples = 48
value = [45, 3, 0]
class = 0
gini = 0.0
samples = 33
value = [0, 0, 33]
class = 2
od280/od315_of_diluted_wines <= 1.445
gini = 0.245
samples = 7
value = [0, 6, 1]
class = 1
gini = 0.0
samples = 3
value = [0, 3, 0]
class = 1
gini = 0.0
samples = 45
value = [45, 0, 0]
class = 0
gini = 0.0
samples = 1
value = [0, 0, 1]
class = 2
gini = 0.0
samples = 6
value = [0, 6, 0]
class = 1

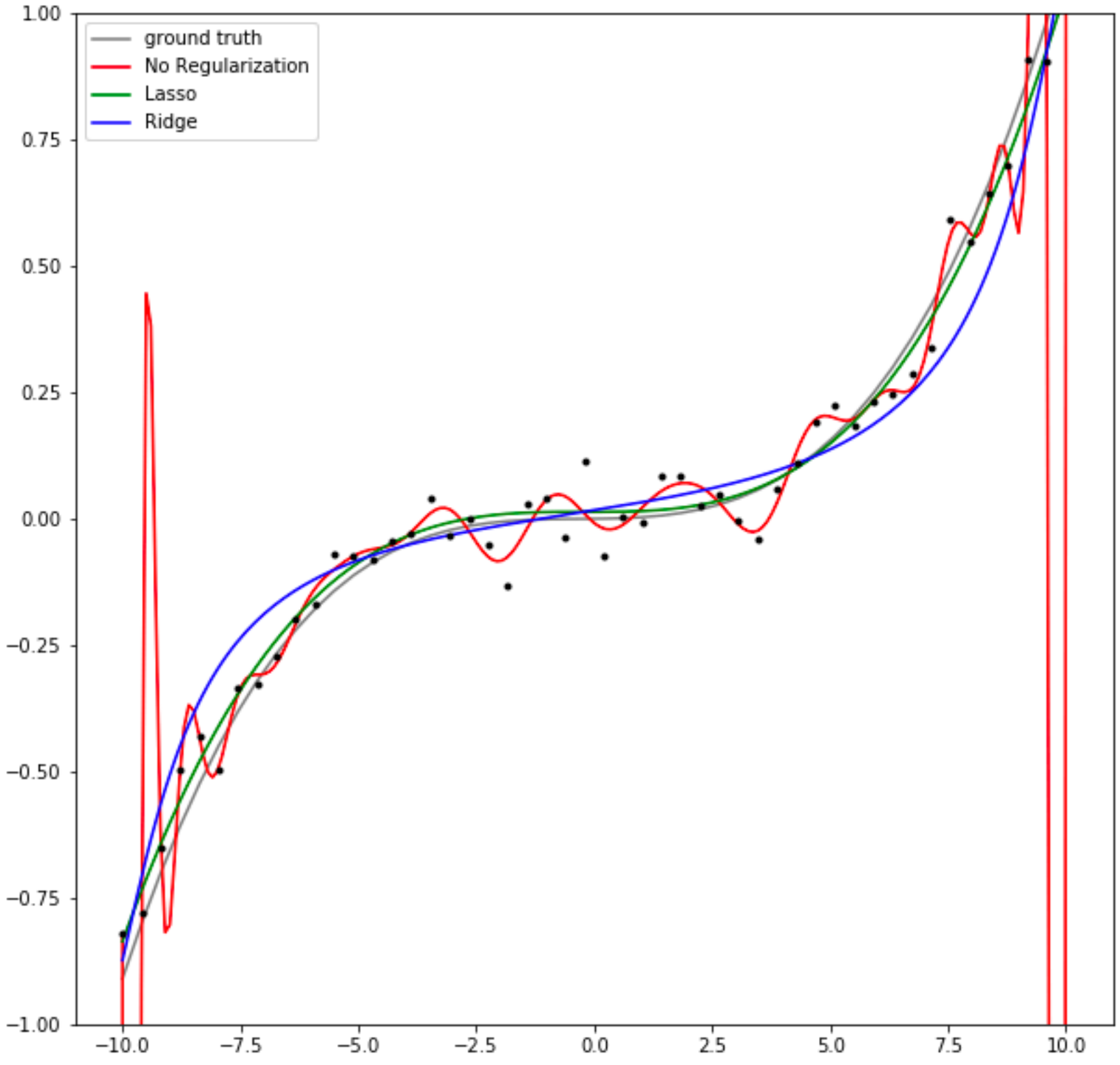
分類器いろいろ

kNN SVM SVM DT RF NN Boost



気をつけること

- 過学習


ground truth
No Regularization
Lasso
Ridge
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0