解釈可能な機械学習

第２章　解釈可能性の重要な概念

この章のトピック

・Feature Importanceと決定領域

解釈方法の種類とその解釈ができる範囲について学ぶ

・機械学習の解釈可能性を妨げる要素

妨げる要素を理解する

ミッション

心血管疾患（CVD）の流行を想定する。

・どのような危険要因に対処できるか

・予測できる場合は、予測を解釈する

CVDに関する詳細を得る

- 問題のコンテキストと関連性を理解
- データ分析とモデル解釈に役立つドメイン知識を得る
- 専門家の情報に基づいた背景をデータセットの特徴量に関連付ける

解釈をモデル化するためにドメイン知識がいかに重要かはちょとむずいとこですよねと。

アプローチ

- ロジスティック回帰は、医療ユースケースの危険因子をランク付けする一般的な方法の１つ。
- Ｘ：各患者のデータ　Ｙ：０から１の間の心血管疾患を患っている確率

解釈方法の種類と範囲について

学習時の概要から解釈する

- 学習を行うと図のような学習の概要を見れる
- モデル係数Coef.を見ると最も貢献した特徴量を見れる
- 線形結合指数はロジスティック関数である

→解釈が困難

- Coef.の絶対値が高いのは、コレステロールと活動である

→これが何を意味するか直感的でない

→解釈可能性のある方法はこれらの係数の指数を計算すればよい

```
Optimization terminated successfully.
         Current function value: 0.561557
         Iterations 6
                         Results: Logit
================================================================
Model:              Logit            Pseudo R-squared: 0.190
Dependent Variable: cardio           AIC:              65618.3485
Date:               2020-06-10 09:10 BIC:              65726.0502
No. Observations:   58404            Log-Likelihood:   -32797.
Df Model:           11               LL-Null:          -40481.
Df Residuals:       58392            LLR p-value:      0.0000
Converged:          1.0000           Scale:            1.0000
No. Iterations:     6.0000
```

解釈方法の種類と範囲について

学習時の概要から解釈する

- 学習を行うと図のような学習の概要を見れる
- モデル係数Coef.を見ると最も貢献した特徴量を見れる
- 線形結合指数はロジスティック関数である

→解釈が困難

- Coef.の絶対値が高いのは、コレステロールと活動である

→これが何を意味するか直感的でない

→解釈可能性のある方法はこれらの係数の指数を計算すればよい

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
const	-11.1730	0.2504	-44.6182	0.0000	-11.6638	-10.6822
age	0.0510	0.0015	34.7971	0.0000	0.0482	0.0539
gender	-0.0227	0.0238	-0.9568	0.3387	-0.0693	0.0238
height	-0.0036	0.0014	-2.6028	0.0092	-0.0063	-0.0009
weight	0.0111	0.0007	14.8567	0.0000	0.0096	0.0125
ap_hi	0.0561	0.0010	56.2824	0.0000	0.0541	0.0580
ap_lo	0.0105	0.0016	6.7670	0.0000	0.0075	0.0136
cholesterol	0.4931	0.0169	29.1612	0.0000	0.4600	0.5262
gluc	-0.1155	0.0192	-6.0138	0.0000	-0.1532	-0.0779
smoke	-0.1306	0.0376	-3.4717	0.0005	-0.2043	-0.0569
alco	-0.2050	0.0457	-4.4907	0.0000	-0.2945	-0.1155
active	-0.2151	0.0237	-9.0574	0.0000	-0.2616	-0.1685

解釈方法の種類と範囲について

Coef.の指数（オッズ）から解釈する

- Coef.の指数を計算したのが右図
- 係数の対数オッズである（オッズは陰性の確率を陽性の確率が上回るということ）
- オッズはよく比率として表される

cholesterol	1.637374
ap_hi	1.057676
age	1.052357
weight	1.011129
ap_lo	1.010573
height	0.996389
gender	0.977519
gluc	0.890913
smoke	0.877576
alco	0.814627
active	0.806471
const	0.000014
dtype: float64	

オッズ

- オッズは比率
- 今日の雨の確率が６０％とすると、雨のオッズは３：２である（６０％：４０％）
- 対数オッズ形式では0.176である（1.5の対数は0.176）
- 指数関数は対数の逆関数であるから、同じ値を受け取れば同じ値を返す

解釈方法の種類と範囲について

Coef.の指数（オッズ）から解釈する

- コレステロールの場合、CVDのオッズが追加ごとに1.64倍に追加することを意味する。
- 他の特徴量は変更されない
- 具体的な用語でモデルに対する特徴量の影響を説明できることは、ロジスティック回帰などの本質的に解釈可能なモデルの利点である

```
cholesterol    1.637374
ap_hi          1.057676
age            1.052357
weight         1.011129
ap_lo          1.010573
height         0.996389
gender         0.977519
gluc           0.890913
smoke          0.877576
alco           0.814627
active         0.806471
const          0.000014
dtype: float64
```

オッズの弱点1

- オッズは有用な情報を提供するが、何が最も重要であるかは教えてくれないために、オッズでFeature Importanceを測ることはできない
- 何かがどれだけ増加するかを測定する場合、コンテキストを理解して、通常どれだけ増加するのかを知る必要がある

→オッズのコンテキストを提供するために、特徴量の標準偏差を計算する

解釈方法の種類と範囲について

特徴量の標準偏差から解釈する

- オッズのコンテキストを理解するには標準偏差を見ることで、通常どれくらい変化するかわかる
- バイナリや順序のある特徴量は通常最大で１つしか変化していない
- 連続値をとる特徴量（weight, ap_hi）は１０～２０倍変化する可能性がある

```
age            6.757537
gender         0.476697
height         8.186987
weight        14.335173
ap_hi         16.703572
ap_lo          9.547583
cholesterol    0.678878
gluc           0.571231
smoke          0.283629
alco           0.225483
active         0.397215
dtype: float64
```

オッズの弱点２

・オッズを使ってFeature Importanceを測定できないもう１つの理由は、オッズが良好（数字が大きい）にも関わらず、特徴量が統計的に有意でない場合がある

→モデルの概要表のP>|z|を見ると良い。

P>|z|（＝ｐ値）について

- Ｐ値が０．０５未満の場合、仮説検定によりそれが有意であるという強力な証拠になる
- ０．０５より大幅に値をうわ待っている場合、予測スコアに影響を与えるという統計的証拠はない

（→このデータセットでは性別などに言える）

解釈方法の種類と範囲について

特徴量の標準偏差とCoef.で解釈する

- Coef.に特徴量の標準偏差をかけることで、最も重要な特徴量を取得できる
- 標準偏差を組み込むと、特徴量間の差異の違いが考慮される
- 性別はP値によると有意でないので削除した
- 右図は、**モデルに応じた危険因子の近似値**として解釈できる

→これを**グローバルモデル解釈法**という

ap_hi	0.936632
age	0.344855
cholesterol	0.334750
weight	0.158651
ap_lo	0.100419
active	0.085436
gluc	0.065982
alco	0.046230
smoke	0.037040
height	0.029620

dtype: float64

モデル解釈可能性メソッドタイプ

- モデル固有

このメソッドは特定のモデルクラスにのみ使用でき、モデル固有である。前の例で説明した手法は、係数を使用するためにロジスティック回帰でのみ機能する。

- モデルにとらわれない

任意のモデルクラスで機能する。

モデル解釈可能性の範囲

大域的全体論的解釈

- モデルがどのように予測を行うかを説明する。データを完全に理解した上でモデル全体を一度に理解できるからで、訓練されたモデルだからである
- 線形回帰みたいな単純なモデルであるからこれができるけど、一般的には無の理

モデル解釈可能性の範囲

グローバルモジュラー解釈

- 部品の役割を説明できるのと同じように、モデルも可能である。
- 特徴量がモデルにどのように影響を与えているのかを考える
- Feature Importanceなど

モデル解釈可能性の範囲

局所的単一予測解釈

- なぜその1つの予測がされたのかを説明する

モデル解釈可能性の範囲

ローカルグループ予測解釈

- 局所的単一予測解釈と同じで、それをグループに適用している

次のミッション

- グローバルモデル解釈法で危険因子を決定した
- 次はモデルを使用して個々の症例を解釈できるかどうかを見ていく

ロジスティック回帰による各予測の解釈

- 陽性と予測されたテストケース2872の詳細が右図
- 高いap_hi（収縮期血圧）
- 通常のap_loである（拡張期血圧）。高い収縮期血圧と正常な拡張期血圧をもつことは、孤立性収縮期高血圧として知られている。これがポジティブケースと予測された可能性があるが、ap_hiはボーダー（130mmHgがボーダー）であり、孤立性収縮期高血圧の状態がボーダーである。

```
age            60.521849
gender          1.000000
height        158.000000
weight         62.000000
ap_hi         130.000000
ap_lo          80.000000
cholesterol     1.000000
gluc            1.000000
smoke           0.000000
alco            0.000000
active          1.000000
Name: 46965, dtype: float64
```

ロジスティック回帰による各予測の解釈

- 高齢ではないが、データセット内では最も高齢である
- コレストロールは正常
- 体重は健康な範囲内

年齢やボーダーの孤立性収縮期血圧だけで陽性になっているのか？陽性になっている理由が明確でない。

→すべての予測をコンテキストに入れずに個別の予測の理由を理解するのは難しい

```
age            60.521849
gender          1.000000
height        158.000000
weight         62.000000
ap_hi         130.000000
ap_lo          80.000000
cholesterol     1.000000
gluc            1.000000
smoke           0.000000
alco            0.000000
active          1.000000
Name: 46965, dtype: float64
```

ロジスティック回帰による各予測の解釈

決定領域/決定境界

- 1つの予測と他1万の予測を比較して可視化するのは、10次元とかの可視化になるから表現できないし、できても理解できん。
- 2つの特徴量については可視化が行えるので、モデルの決定境界がどのへんにあるか見れる
- テスト用データセットの予測を上から描画もできる

→2つの特徴量と他11の特徴量の効果の不一致を可視化するため

ロジスティック回帰による各予測の解釈

決定領域/決定境界

- この可視化解釈方法を決定境界という
- クラスに属する領域を描画することは決定領域という

ロジスティック回帰による各予測の解釈

ケテリスパリブスの仮定

- 他の全ての特徴量が一定に保たれているとき、2つだけ分離して観察できるという大きな前提のもとで、一度に2つの意思決定ベースの特徴量を可視化できた

→これをケテリスパリブスの仮定という

ロジスティック回帰による各予測の解釈

ケテリスパリブスの仮定を行う方法

- 結果に影響を与えないような値で他の特徴量を埋めることで、これを行える
- 例えばオッズの表を使うと、CVDのオッズが増加するため、特徴量が増加するか判断できる

→全体として値が低いほどCVDのリスクは低くなる

ロジスティック回帰による各予測の解釈

可視化を行うために値を置換する

- 年齢では30歳が最もリスクの低い値
- アクティブ値は1のほうがリスクが低い

残りの特徴量は最頻値から考える

- **height**=165.
- **weight**=57 (optimal for that **height**).
- **ap_hi**=110.
- **ap_lo**=70.
- **smoke**=0.
- **cholesterol**=1 (this means normal).
- **gender** can be coded for male or female, which doesn't matter because the odds for gender (**0.977519**) are so close to 1.

ロジスティック回帰による各予測の解釈

決定領域/決定境界

- テストケース2872は57%で陽性判定だった
- 右図の黒丸を見るとだいたい境界あたりにいる
- ap_hiとコレステロールは陰性の範囲内
- y軸が増加するにつれ、三角形が陽性側によってる

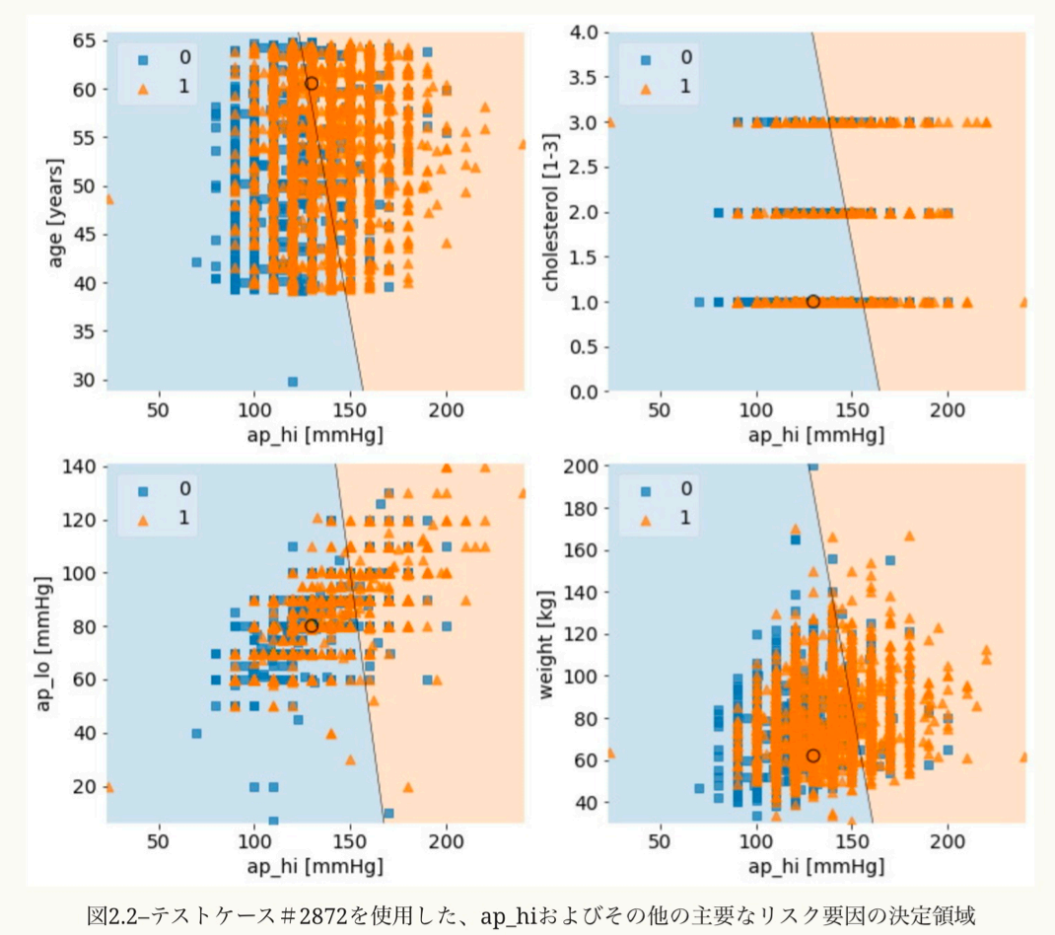

図2.2–テストケース＃2872を使用した、ap_hiおよびその他の主要なリスク要因の決定領域

ロジスティック回帰による各予測の解釈

決定領域/決定境界

- ap_hiとweihgtグラフでは、weightが増えるにつれ「三角形が陽性側による」というパターンが当てはまらない

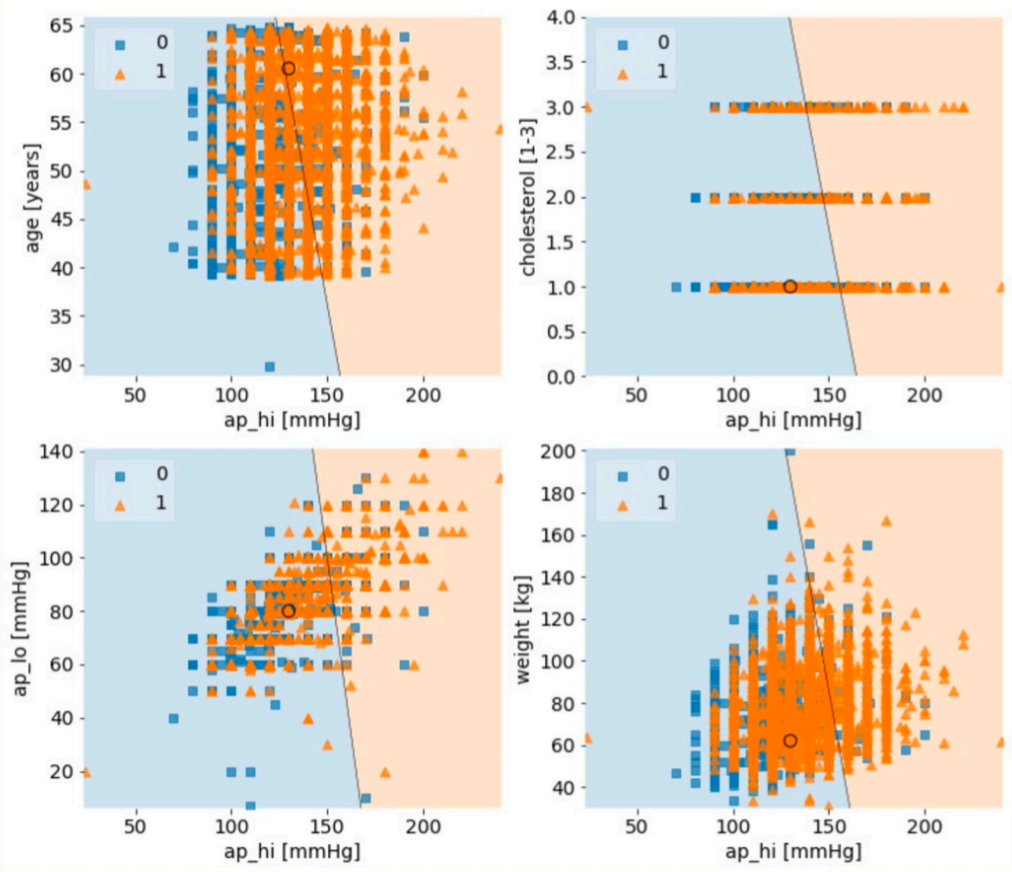

図2.2–テストケース＃2872を使用した、ap_hiおよびその他の主要なリスク要因の決定領域

ロジスティック回帰による各予測の解釈

決定領域/決定境界

- 決定領域/決定境界のプロットは個々の症例予測を解釈するためのツールであった
- どのような違いになるか見るために一部の埋めた変数を変更してみる
- 例えば年齢の穴埋め値を中央値の54歳にテストケース2872の年齢あげるとどうなるか
- ボーダーのap_hiやコレステロールは陽性にする原因になりうるか

解釈可能性を妨げる要素

体重と身長の関係

- 前の章でap_hiとweightのプロットがパターンに合わなかった
- CVDのリスク増加を説明できる他の重要な媒介変数がある可能性がある
- 媒介変数は独立変数とターゲット変数のつながりに影響する
- weightとheightに対して、人間の健康については体重と身長両方見る必要があるので、プロットしてみていく

解釈可能性を妨げる要素

体重と身長の関係

- 陽性(オレンジ)にはパターンがあり、密集した箇所がある

→身長があがると体重が増えると予想しても、本質的に健康な体重の概念は身長に比例して増えるものではない

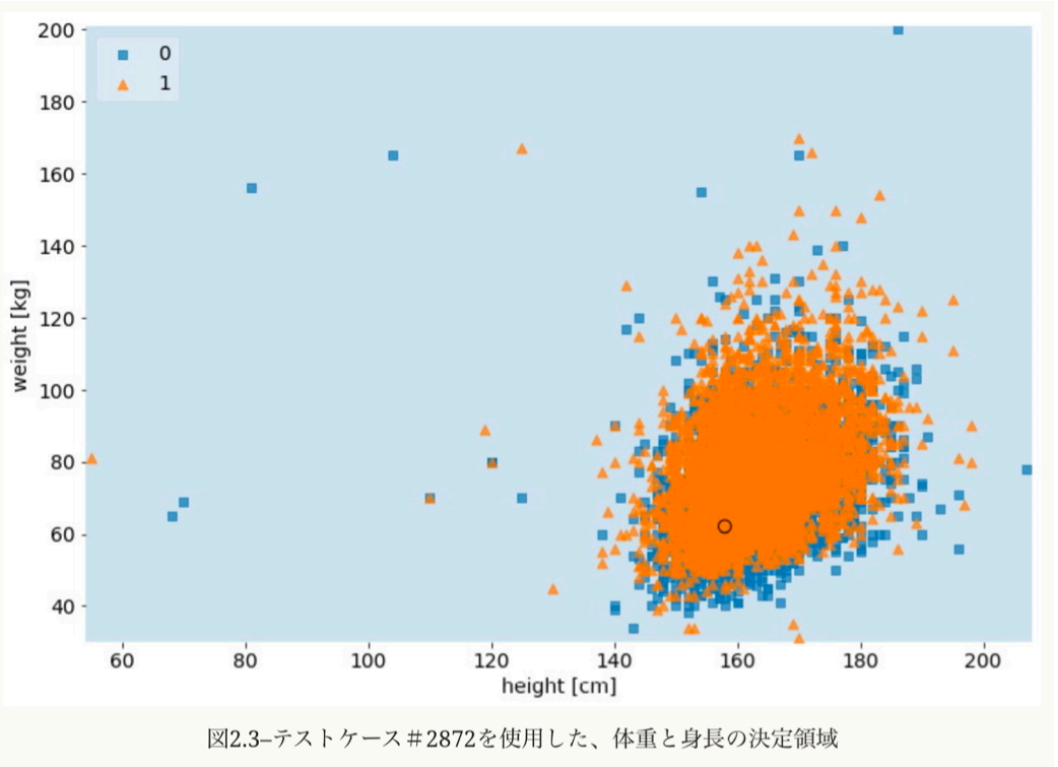

図2.3–テストケース＃2872を使用した、体重と身長の決定領域

解釈可能性を妨げる要素

体重と身長の関係

- ボディマス指数(BMI)とよばれる数式で理解できる

$$\mathrm{BMI} = \frac{\mathrm{weight}_{kg}}{\mathrm{height}_{m}^{2}}$$

解釈可能性を妨げる要素

解釈を困難にする複雑さをもたらす3つのこと

- 非線形性
- 双方向性
- 非単調性

解釈可能性を妨げる要素

非線形性

- y=ax+bなどの一次方程式は理解しやすく、相加的であるために定量化が簡単
- 多くのモデルクラスには線型方程式が埋め込まれていて、これらは学習やモデルの説明のために使う
- 学習に非線形性を導入して本質的に非線形であるモデルがある
- ただし、ロジスティック回帰は相加的であるために一般化線形モデル(GLM)とみなされる(結果は加重入力とパラメータの合計になる)

解釈可能性を妨げる要素

非線形性

- モデルが線形であっても変数間の関係が線形でないことがある

→パフォーマンスと解釈可能性が低下する可能性がある

→そのときは非線形モデルクラスを使うか、ドメイン知識を使ってそれを線形にする特徴量を設計するかで対処できる

解釈可能性を妨げる要素

非線形性モデルクラス

- 非線形特徴量の関係をより反映させられる
- その結果モデルの精度があがる可能性がある
- 使用することで解釈が難しくなる可能性がある

解釈可能性を妨げる要素

ドメイン知識を使って線形の特徴量を作成

- ドメイン知識を使うことで、非線形の特徴量を線形に変換できるような特徴量を設計する
- 例えば、ある特徴量が別の特徴量に対して指数関数的に増加した場合、その特徴量の対数を使用して新しい変数を設計できる

→CVD予測では、BMIが身長と体重を理解するためによい方法とわかってるためにこれを使うのが良い

→任意の特徴量ではないために、解釈が容易である

解釈可能性を妨げる要素

ドメイン知識を使って線形の特徴量を作成

- BMIを追加の特徴量とし、プロットしてみると右図の通り。
- BMIとweightは、heightとweightやBMIとheightの関係よりも明確な線形関係がある

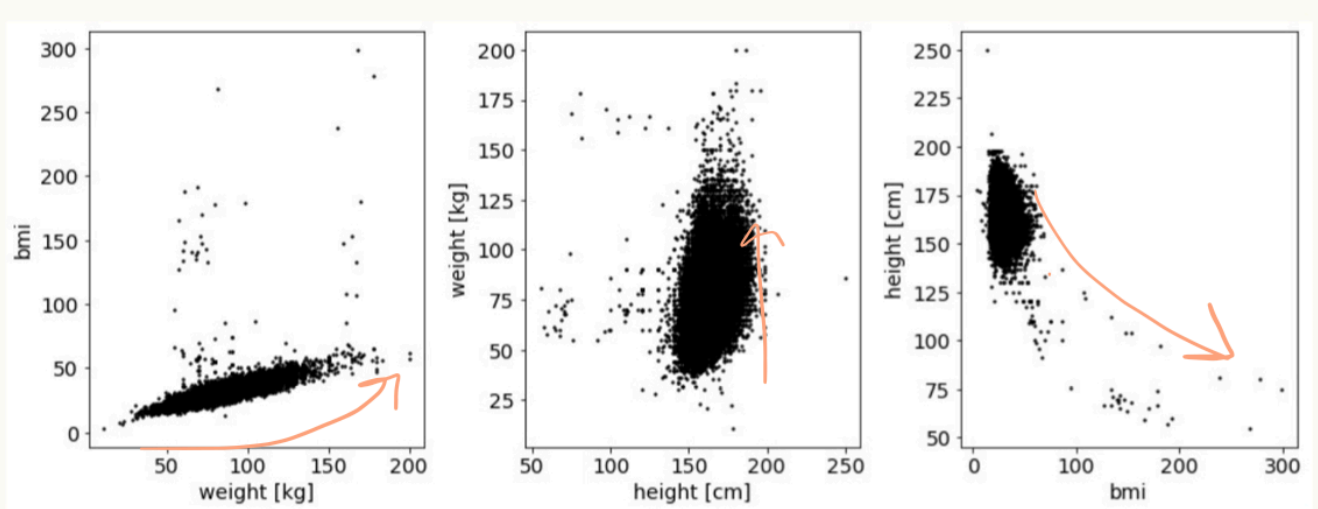

図2.4–体重、身長、BMIの二変量比較

解釈可能性を妨げる要素

ドメイン知識を使って線形の特徴量を作成

- 決定領域をプロットすると右図のようになる
- 黒丸が陽性の領域にあることから、ap_hiとBMIがCVDの陽性予測を説明するのに役立っていることがわかる

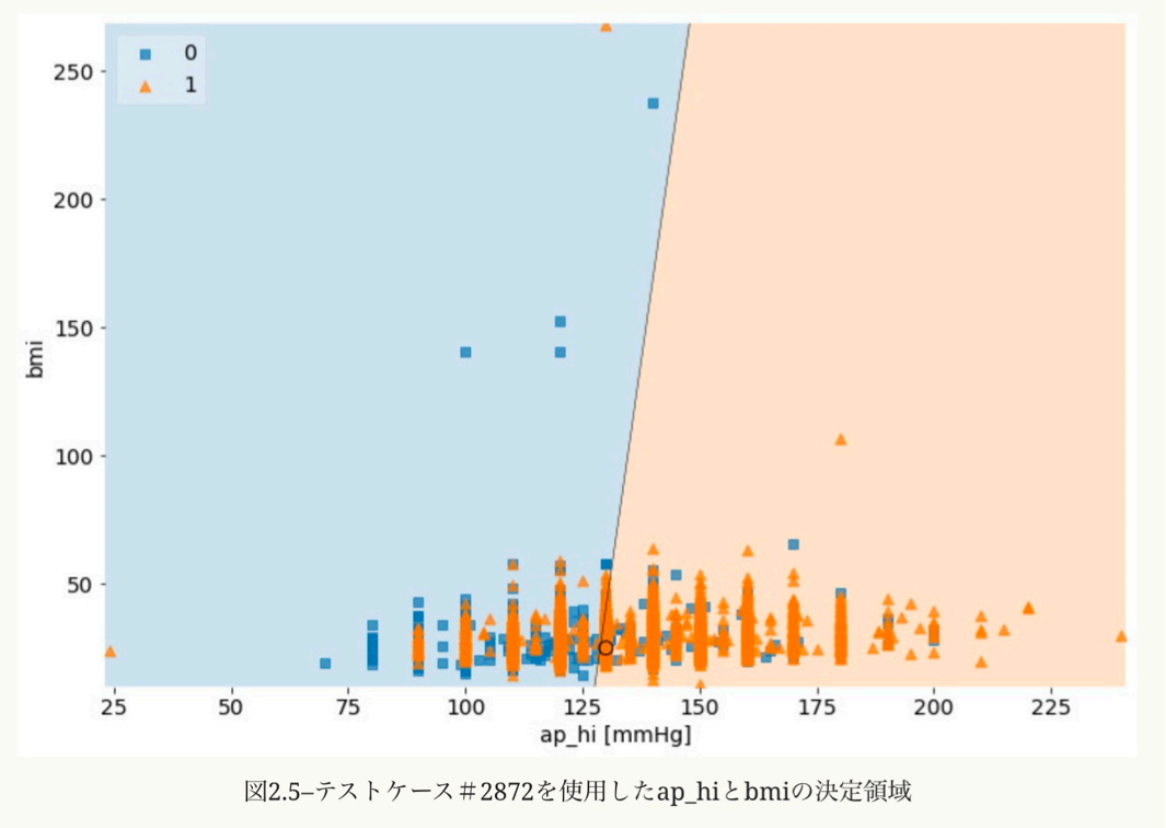

図2.5–テストケース＃2872を使用したap_hiとbmiの決定領域

解釈可能性を妨げる要素

ドメイン知識を使って線形の特徴量を作成

- しかし、異常なBMI値(外れ値)がある
- これによってデータセット内にいくつか誤ったweightやheightがあるとわかる

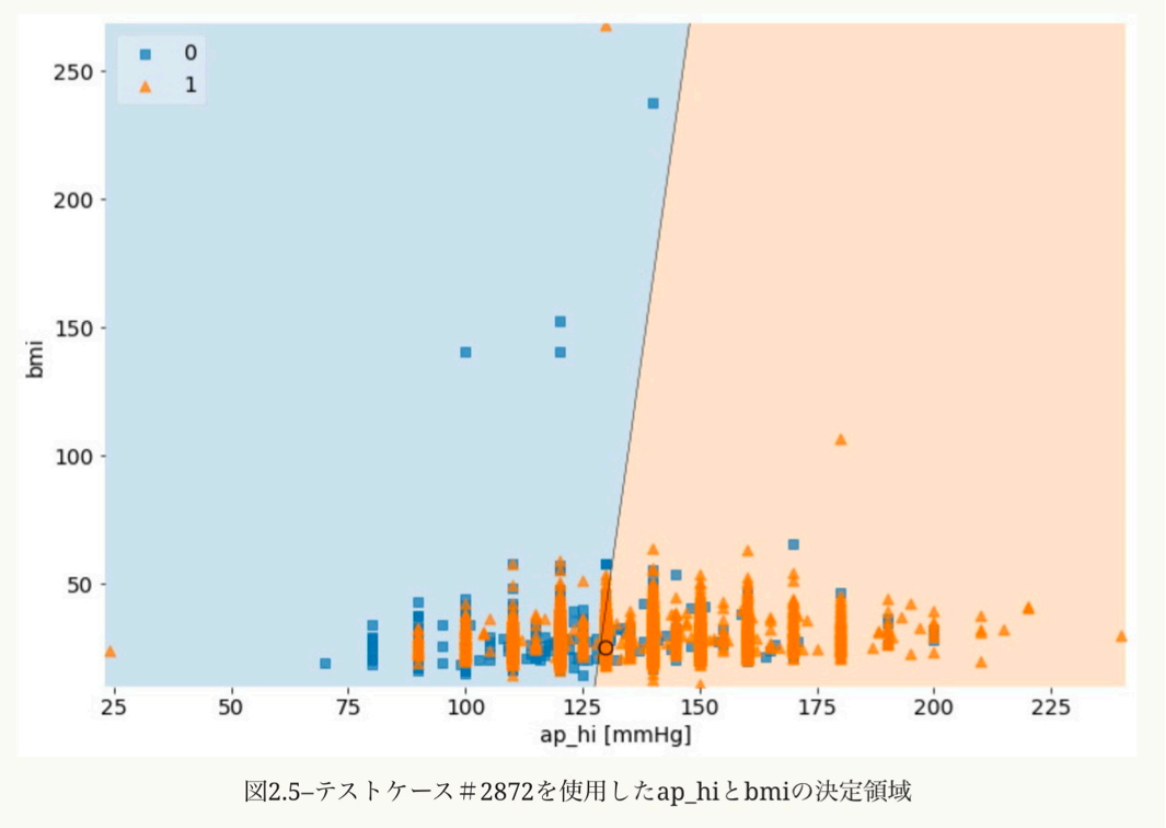

図2.5–テストケース＃2872を使用したap_hiとbmiの決定領域

解釈可能性を妨げる要素

外れ値

- 外れ値は影響力があるか、レバレッジが高い可能性があるために、これを含んだまま学習するとモデルに影響がでる
- そうでない場合でも解釈可能性を下げる可能性がある
- 外れ値は削除する必要がある
- 他の特徴量のコンテキストで外れ値が隠れる場合がある
- 前図をズームアウトして重要な決定境界を理解せずに、モデルの学習を行ってしまうなどの問題がある

解釈可能性を妨げる要素

双方向性

- BMIを作成したとき、非線形関係を線形化するだけでなく、2変数間の相互作用を作っている
- よって、BMIは相互特徴量であるが、これはドメイン知識に基づくものである
- 多くのモデルクラスは特徴量間で全ての種類の操作を並べてこれを自動で行っている
- つまり特徴量にはheightとwidth、ap_hiとap_loのように潜在的な相互作用関係がある

解釈可能性を妨げる要素

双方向性

- ただし構造化データの場合、相互作用はモデルの精度に重要な可能性もあるが、潜在的に不要な複雑さを追加することで解釈可能性を下げる可能性がある
- それに加えて意味のない潜在的な関係（これをスプリアス関係または相関という）を見つけることによって解釈可能性を損ねる可能性がある

解釈可能性を妨げる要素

非単調性

- 多くの場合、特徴量とターゲット変数の間に意味のある一貫した関係がある
- これによって、年齢があがるにつれてCVDのリスクが高まるはずだとわかる
- ある年齢に達するとリスクが低下することはない(緩やかにはなるかも)
- これを単調性といい、単調な関数はドメイン全体を通して常に増加か減少かしている

解釈可能性を妨げる要素

非単調性

- 多くの場合、特徴量とターゲット変数の間に意味のある一貫した関係がある
- これによって、年齢があがるにつれてCVDのリスクが高まるはずだとわかる
- ある年齢に達するとリスクが低下することはない(緩やかにはなるかも)
- これを単調性といい、単調な関数はドメイン全体を通して常に増加か減少かしている

解釈可能性を妨げる要素

非単調性

- 全ての線形関係は単調であるが、全ての単調関係が必ずしも線形であるとは限らない
- これは直線である必要がないから
- 機械学習でよくある問題は、領域の専門性から期待する単調な関係を、モデルが知らないこと
- データにノイズや抜けがあるため、想定していないところで浮き沈みがあるようなモデルを学習してしまう

解釈可能性を妨げる要素

非単調性

- 例えばCVD予測で、57~60歳のデータが全然入手できなくて、しかも入手できたサンプルは陰性ばかりだったときを考える
- モデルはこの範囲ではCVDリスクが低下すると学習できるはず

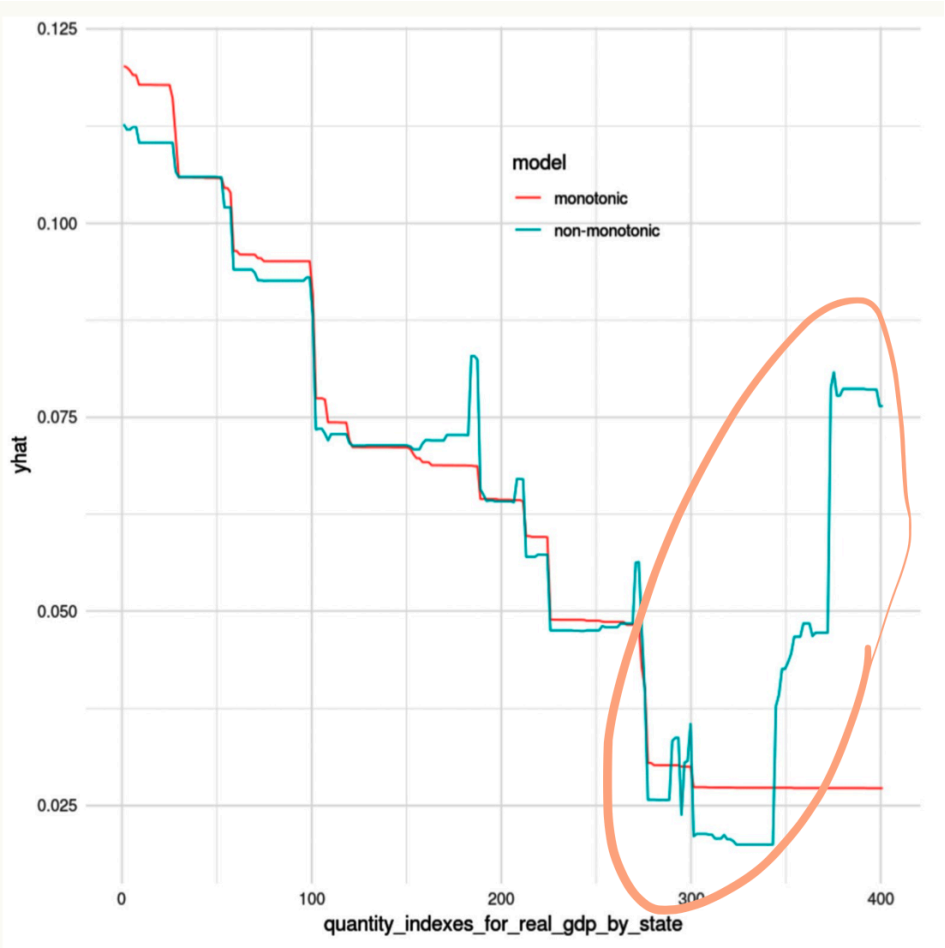

図2.6–単調モデルと非単調モデルを使用したターゲット変数（yhat）と予測子の間の部分依存プロット

解釈可能性を妨げる要素

非単調性

- 右図は部分依存プロット（PDP）と呼ばれる
- 非単調な関数ではギザギザしながら減少し、最後に増加してしまう

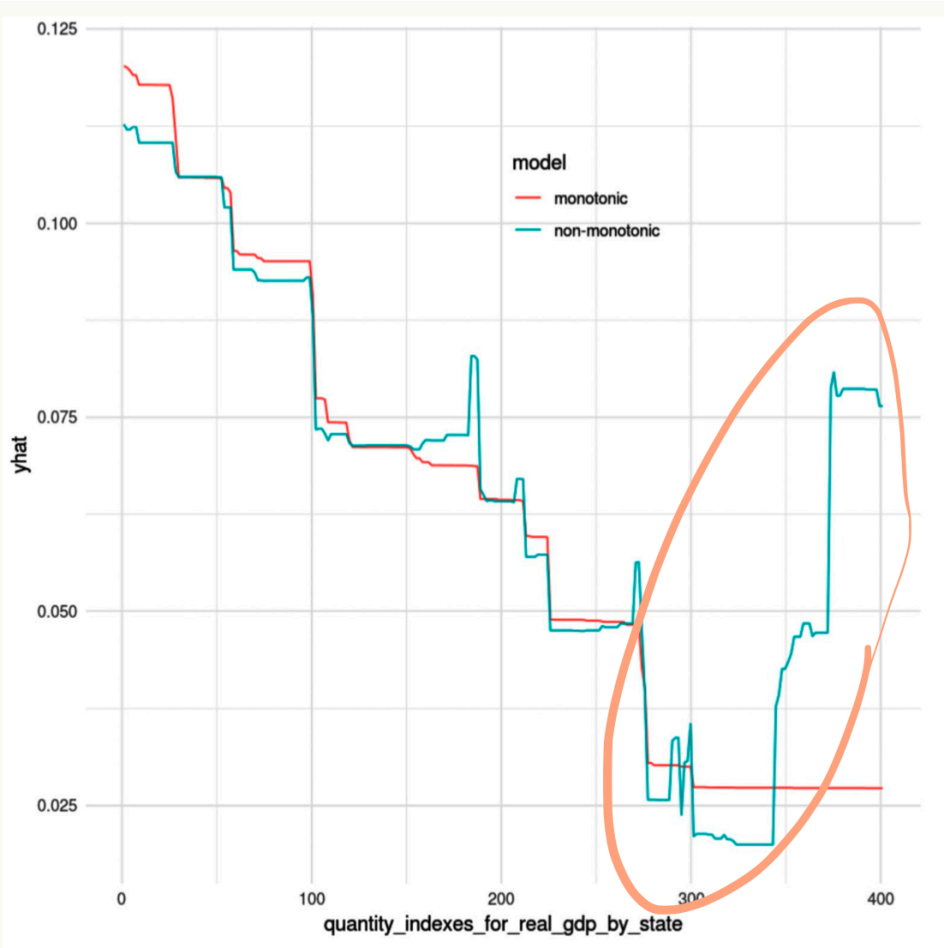

図2.6–単調モデルと非単調モデルを使用したターゲット変数（yhat）と予測子の間の部分依存プロット

まとめ

- 心血管疾患の危険因子はap_hi、年齢、コレステロール、体重である。
- しかし、収縮期血圧（ap_hi）は拡張期血圧（ap_lo）に依存して解釈されるため、単体ではあまり意味がない
- 体重と身長も同様
- 解釈には特徴量の相互作用が重要な役割を果たし、線形か単調かといった互いの関係や対象変数との関係も同様である

まとめ

- データはあくまで真実の表現であり、間違っていることもある
- 外れ値が見つかり、それを放置しておくと、モデルに偏りが生じる可能性がある

まとめ

- もう一つのバイアスの原因は、データの収集方法
- CVDの例で、モデルの上位特徴量がすべて客観的で検査機能であった
- なぜ喫煙や飲酒がより大きな要因にならないのか

→サンプルの偏りが関係しているかどうかを検証するには、より信頼できる他のデータセットと比較して、データセットが飲酒者や喫煙者を過小に表現していないかどうかをチェックする必要がある

まとめ

- あるいは、長期間喫煙したことがあるかどうかではなく、現在喫煙しているかどうかを問う質問によって、バイアスが生じたのかも
- モデルが描こうとしている真実を説明する情報がデータに欠けている可能性がある
- 例えば、CVDのリスクを高める孤立性収縮期高血圧などの血圧の問題は、糖尿病、甲状腺機能亢進症、動脈硬化、肥満などの基礎疾患によって引き起こされることが分かっている

まとめ

- 医学的な原因のうち、データから導き出せるのは肥満だけで、他の条件はない
- モデルの予測をうまく解釈しようと思えば、関連する特徴をすべて持っている必要があり、そうでないと予測にギャップが生じる

まとめ

- 決定領域をプロットすることで個々のモデルの予測値を解釈できる
- これは単純な方法であるが、多くの制限がある
- それらが互いに多くの相互作用をする傾向があるような状況では、制限がある