PWSCUP2025 Team16

大門ステテ子

明治大学 菊池研究室 ：関口智大 小泉海斗 三浦晃暉 菊池浩明

明治大学

合成データ

CTGAN

1. カテゴリ変数をOne-hotエンコーディング後にasthma_flagを目的変数として
ロジスティック回帰を実行、その後回帰変数をその変数の重みとして設定し、
特徴量を調整する

	p値	重み計算	重み	重み付き特徴量
AGE	0.001	$-\log10(0.001)=3.0$	3/6	46*3/6
encounter	0.01	$-\log10(0.01)=2.0$	2/6	381*2/6
BMI	0.1	$-\log10(0.1)=1.0$	1/6	29*1/6

2．調整特徴量で各変数の変数の平均との差を取り、マハラノビス距離を計算する

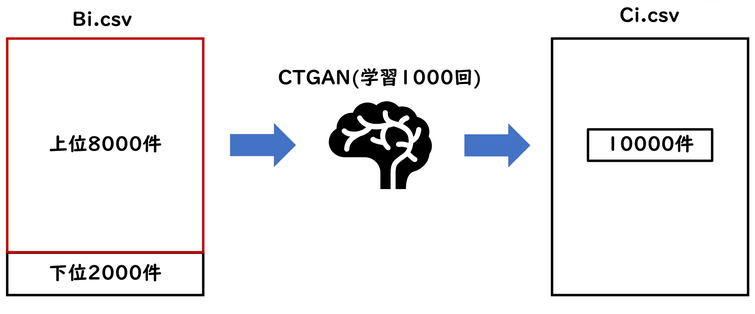
マハラノビス距離の公式

$$d=\sqrt{(\vec{x}-\vec{\mu})^T\Sigma^{-1}(\vec{x}-\vec{\mu})}$$

$\sum$:分散・共分散行列
x:変数
μ:xの平均

	重み付き マハラノビス距離	ランキング	採用
関口	2.1	3000	採用
小泉	0.4	1	採用
三浦	10.3	9000	不採用

3．ユーザー間のマハラノビス距離によって距離の小さいもの学習させて出力する

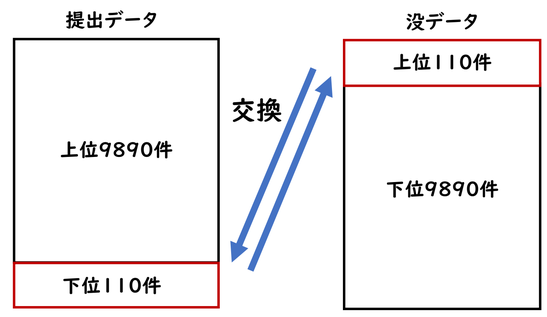


有用性向上

有用性
10.44UP↗

Cデータ

1．それぞれのデータ内でマハラノビス距離を取り、下位と上位を交換する



2．asthma_flagを目的変数として
線形回帰予測確率が0.5付近のものを反転
※カテゴリ変数除くすべての変数を
説明変数とする

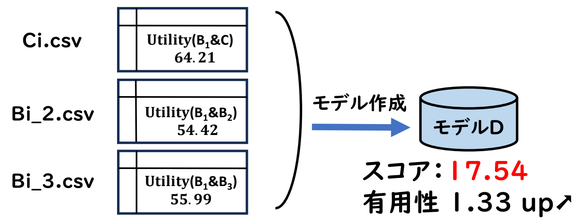
	AGE	…	asthma_flag	予測確率
関口	45	…	0	0.7
小泉	27	…	1→0	0.53
三浦	38	…	0	0.15

3．Bデータ内とCデータ内でマハラノビス距離を
計算し距離の遠いユーザーと近いユーザーの
変数を中央値±10%に変更または反転を行う

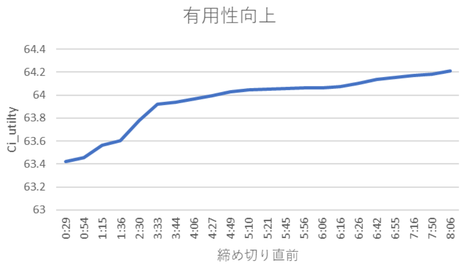
	AGE	…	stroke_flag	mean_bmi
関口	45	…	0	29.57
小泉	27	…	1→0	17.8
三浦	38	…	0	47.2→29.3

モデルD

Cデータで学習させたモデルのスコアは16.41
→ AデータとBデータに対して性能が低い
Cと使わないBを結合してモデルの汎化性能を上げる



スコア：17.54
有用性 1.33 up↗



Gower距離攻撃

Gower距離…クラスタリングに用いられる量質混在型データの差異を測るための指標

Gower距離計算方法

数値データ… $d_{ij}^{(k)}=\frac{|x_i-x_j|}{R_k}$

カテゴリデータ…一致→ $d_{ij}^{(k)}=0$, 不一致→ $d_{ij}^{(k)}=1$

Gower距離：$D_{ij}=\frac{\sum_{k=1}^{c}w_k d_{ij}^{(k)}}{\sum_{k=1}^{c}w_k}$

頻度重み

偶然一致しやすい列を軽く、一致しにくい列を重くする
項目kにおける値vの相対頻度：$p_{k,v}=\frac{n_{k,v}}{N_k}$

正規化後重み：$\widetilde{w}_k=\frac{1}{\sum_v p_{k,v}^2}$

重み：$w_k=\frac{\widetilde{w}_k}{\sum_l \widetilde{w}_l}$

i,j	レコード(患者)
c	項目数(列数)
k	項目(GENDER, AGE, …)
R_k	項目kの範囲
$d_{ij}^{(k)}$	項目kにおける距離
D_{ij}	患者間のGower距離
w_k	項目kの重み
$\widetilde{w}_k$	項目kの正規化前の重み
v	連続値やカテゴリの値
$n_{k,v}$	項目kにおける値vの数
N_k	項目の値域

計算例

患者	GENDER	AGE	RACE	stroke	weight
小泉	M	27	asian	0	60.0
A_1	M	45	asian	0	78.0
A_2	F	28	asian	1	60.5
A_3	M	34	white	1	70.0
A_4	M	35	other	0	80.0
範囲	2	18	3	2	20.0

項目	頻度重み
GENDER	0.165
AGE	0.238
RACE	0.195
stroke	0.165
weight	0.238

項目	単純平均の場合
D_{k,A_1}	0.380
D_{k,A_2}	0.416
D_{k,A_3}	0.778
D_{k,A_3}	0.489

$$D_{k,A_1}=\frac{|27-45|}{18}\times w_{AGE}+\frac{|60.0-78.0|}{20.0}\times w_{weight}=0.452$$

$$D_{k,A_2}=1\times w_{GENDER}+\frac{|27-28|}{18}\times w_{AGE}+1\times w_{stroke}+\frac{|60.0-60.5|}{20.0}\times w_{weight}=0.348$$

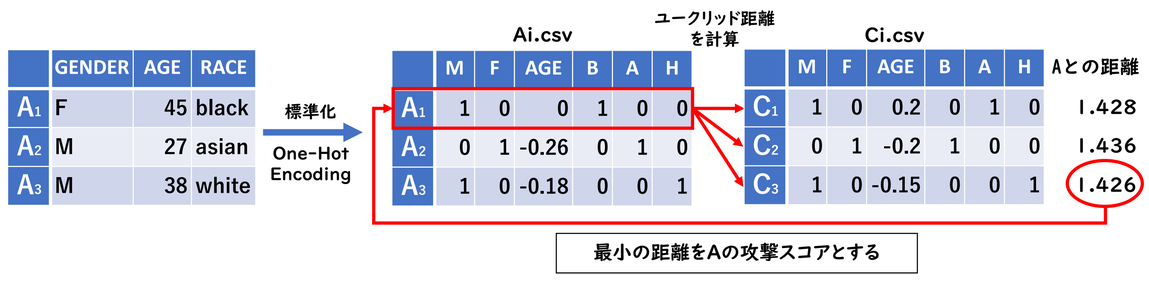
$$D_{k,A_3}=\frac{|27-34|}{18}\times w_{AGE}+1\times w_{RACE}+1\times w_{stroke}+\frac{|60.0-70.0|}{20.0}\times w_{weight}=0.736$$

$$D_{k,A_3}=\frac{|27-35|}{18}\times w_{AGE}+1\times w_{RACE}+\frac{|60.0-80.0|}{20.0}\times w_{weight}=0.538$$

最近傍 & 絞り込み攻撃

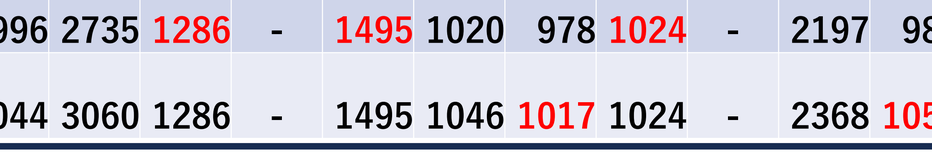
最近傍攻撃

1．数値データはCデータから統計情報を標準化、
カテゴリデータはOne-Hotエンコーディングを行う
2．AとCのレコードのユークリッド距離をすべての組み合わせで計算し、
最小距離をAのレコードの攻撃スコアとする
3．Aを攻撃スコアの小さい順に1万件取り出す



絞り込み攻撃

1．メンバーと推定したF1のファイルを提出し、BかつF1(攻撃成功数)が
1000を下回る場合にはF1を除く
2．A-F1(F1を除いたA)の9万件からメンバーだと推定したF2を提出し、
1と同じ処理を繰り返しメンバーの絞り込みを行う



本戦攻撃結果

チーム	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Gower	987	1020	1176	433	1097	998	990	586	973	996	974	1019	1023	3060	1114	-	1282	963	969	1005	-	2368	967	2432
Gower2回目	948	989	986	1039	1042	982	995	1137	952	1014	1024	974	982	1024	994	-	983	1008	952	987	-	1023	960	1018
近傍1+Gower3	1031	1002	1077	615	1064	959	1027	1088	997	980	1027	1025	968	1584	1093	-	1180	939	1000	1011	-	1389	974	1105
近傍2+Gower4	957	976	1176	433	1120	1049	990	586	973	1002	974	1019	1027	3060	1114	-	1282	1046	969	991	-	2368	967	2432
近傍3+Gower5	973	1089	1176	433	1107	1016	990	1039	973	983	974	1019	991	3060	1114	-	1282	943	969	977	-	2368	967	2432
近傍4+Gower6	990	965	1176	433	990	983	990	1062	973	970	974	1019	1044	3060	1114	-	1282	957	969	980	-	2368	967	2432
期待値以下を外してランダム	1033	1089	1134	1106	1231	1049	1023	1137	1024	1038	1005	1070	996	2735	1286	-	1495	1020	978	1024	-	2197	981	2934
期待値以下を外してランダム2	1033	1089	1176	1106	1227	1049	1027	1137	1024	1038	1027	1070	1044	3060	1286	-	1495	1046	1017	1024	-	2368	1051	2934