鷲尾研・ERATO 合同セミナ
September 13, 2010

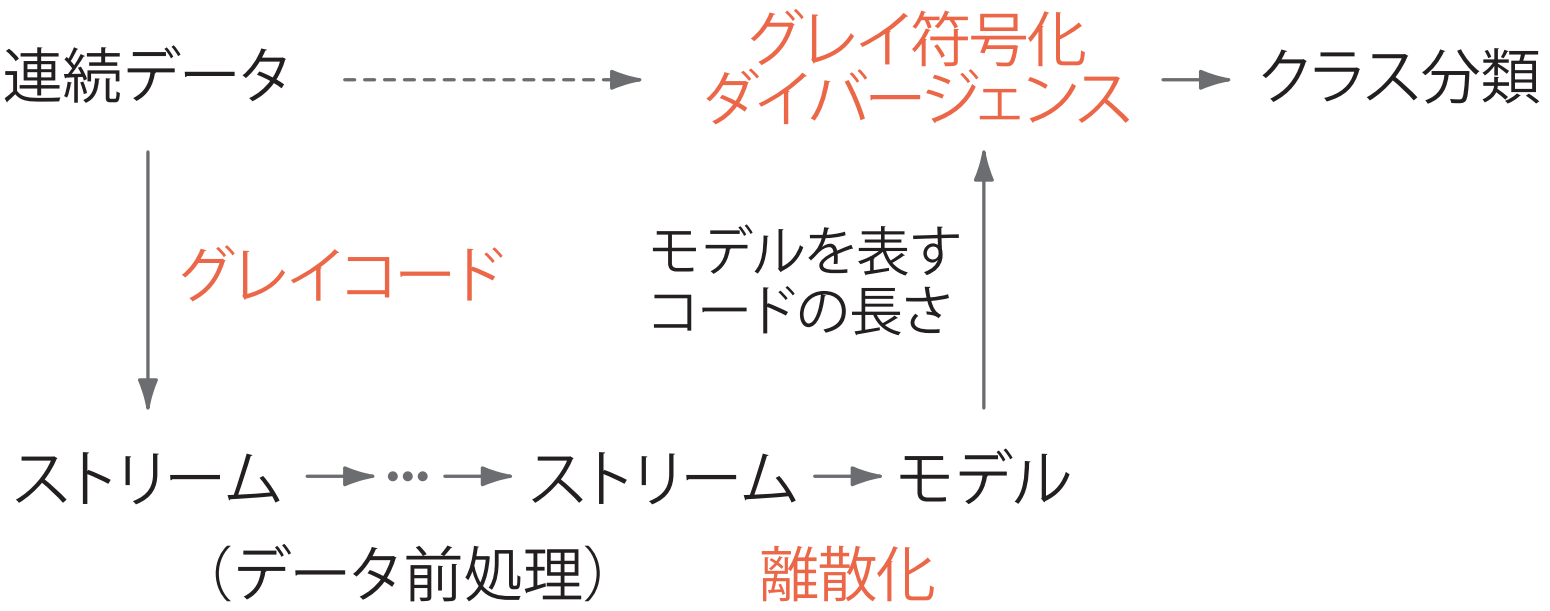
グレイ符号化ダイバージェンスによる連続データからの計算論的知識発見

杉山 麿人[†,‡]，山本 章博[†]
[†] 京都大学情報学研究科
[‡] 日本学術振興会特別研究員 DC2

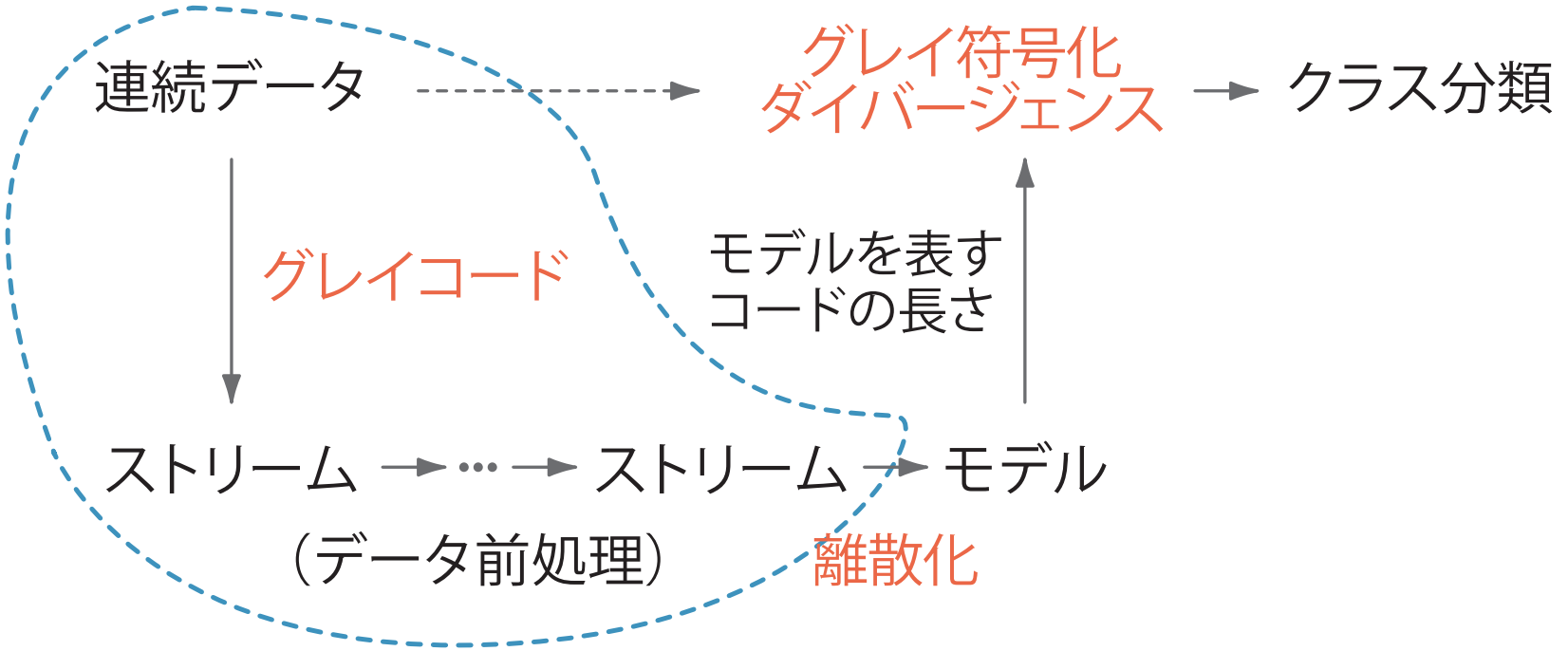
研究の概要

- 連続データからの知識発見・機械学習では，ほとんどの場合，離散化された誤差を含む近似値を使う
  - 得られた結果は重大な誤差を含む可能性がある
- 目標：数値誤差ゼロの知識発見を計算論的な枠組みで実現し，得られた結果（知識）の正当性を保証する
- 成果：
  1. 2 つの連続データ集合間の異なり具合を測るグレイ符号化ダイバージェンスという新規の尺度を定式化した
     - クラス分類やクラスタリングの基礎となる
  2. 実データ実験でグレイ符号化ダイバージェンスを用いた分類精度を評価し，頑健かつ優れていることを示した

研究の概要（図解）

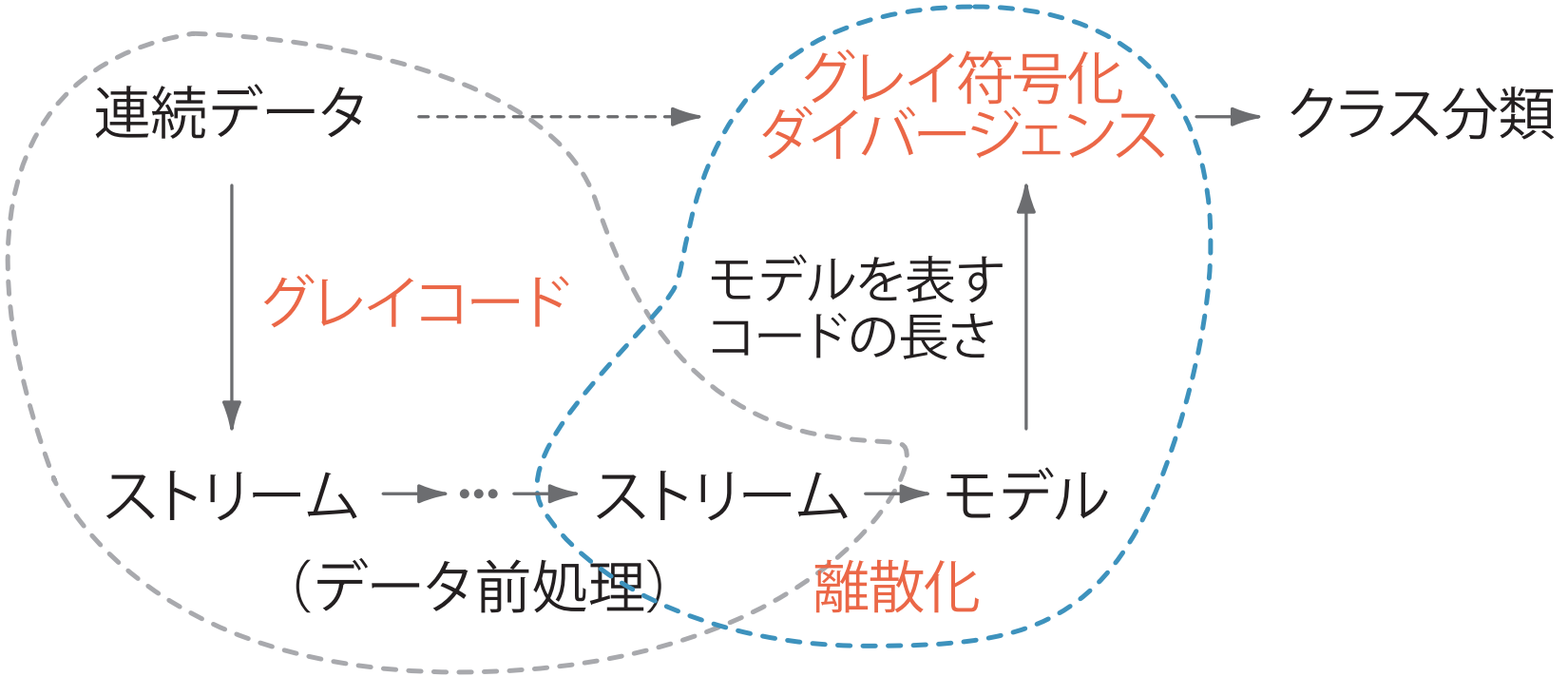

連続データ
グレイ符号化
ダイバージェンス
クラス分類
グレイコード
モデルを表す
コードの長さ
ストリーム
ストリーム
モデル
（データ前処理）
離散化

研究の概要（図解）



計算可能性解析学を用いて
ストリーム計算を保証

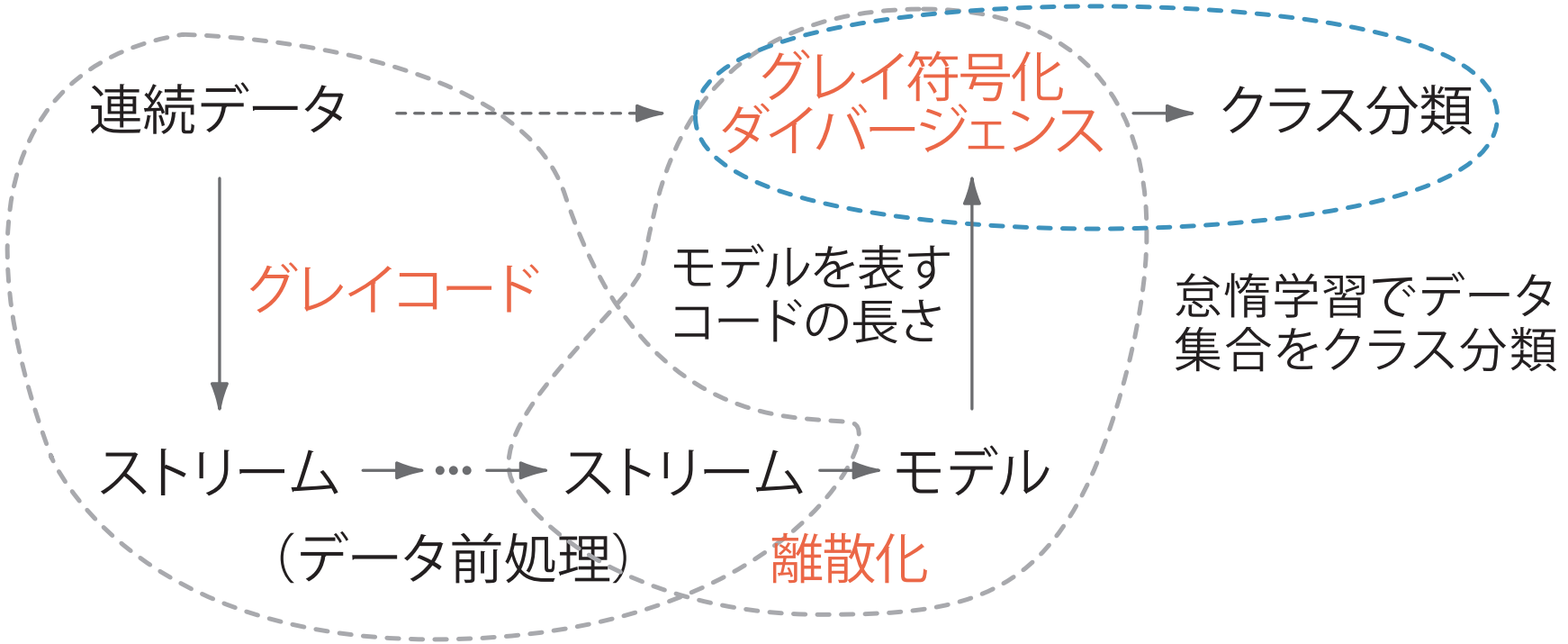
研究の概要（図解）



計算可能性解析学を用いて
ストリーム計算を保証

2つのストリーム集合をそれぞれ正例，負例の集合と捉えて，計算論的にグレイ符号化ダイバージェンスを学習

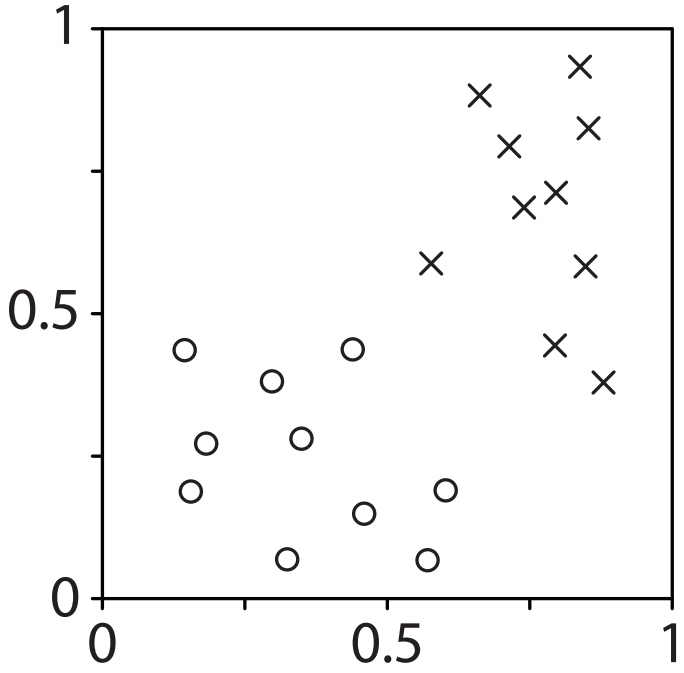
研究の概要（図解）

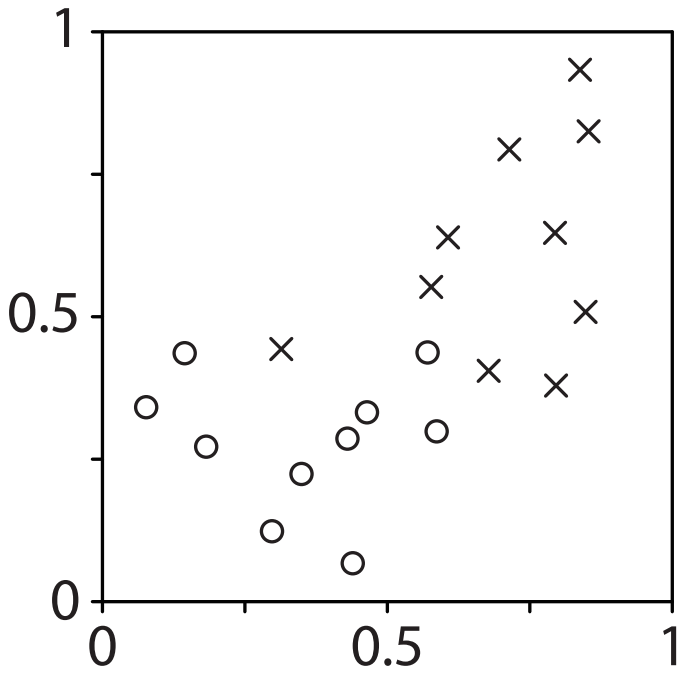


計算可能性解析学を用いて
ストリーム計算を保証

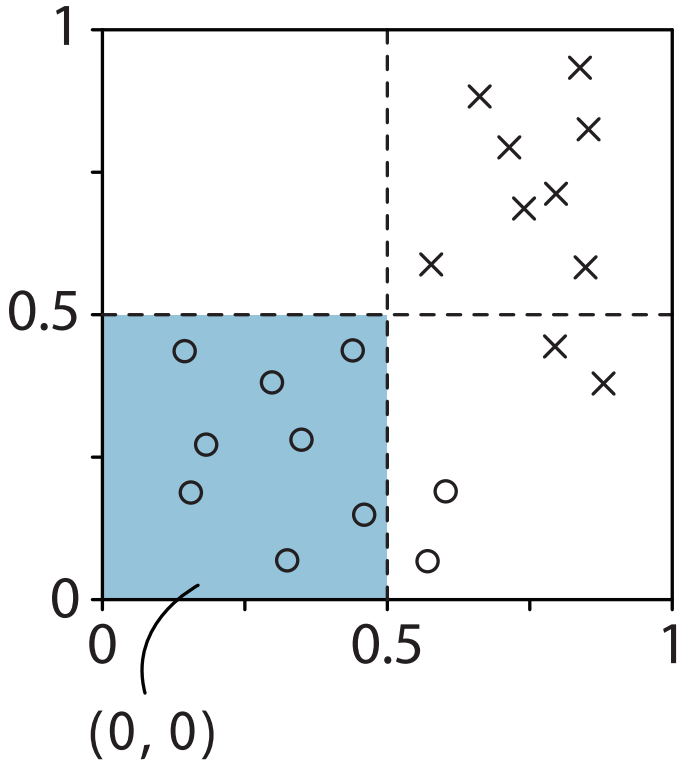
2つのストリーム集合をそれぞれ正例, 負例の集合と捉えて, 計算論的にグレイ符号化ダイバージェンスを学習

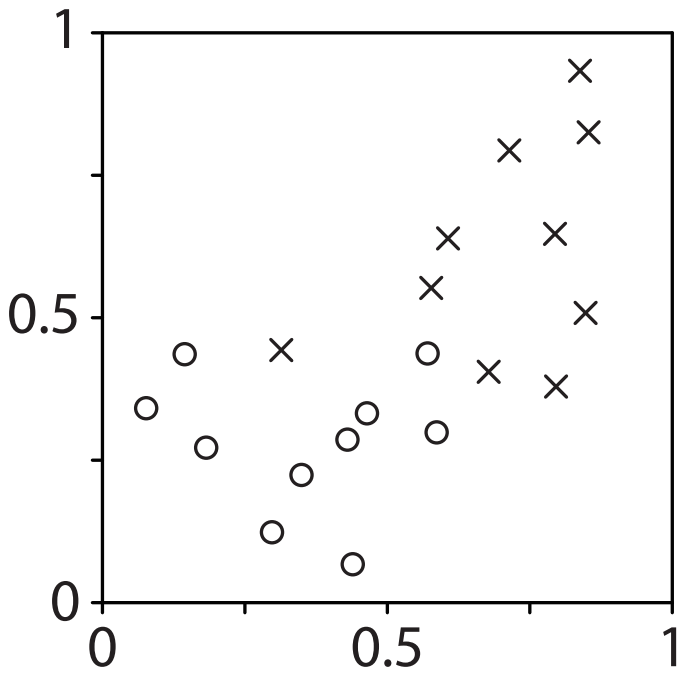
グレイ符号化ダイバージェンスの例



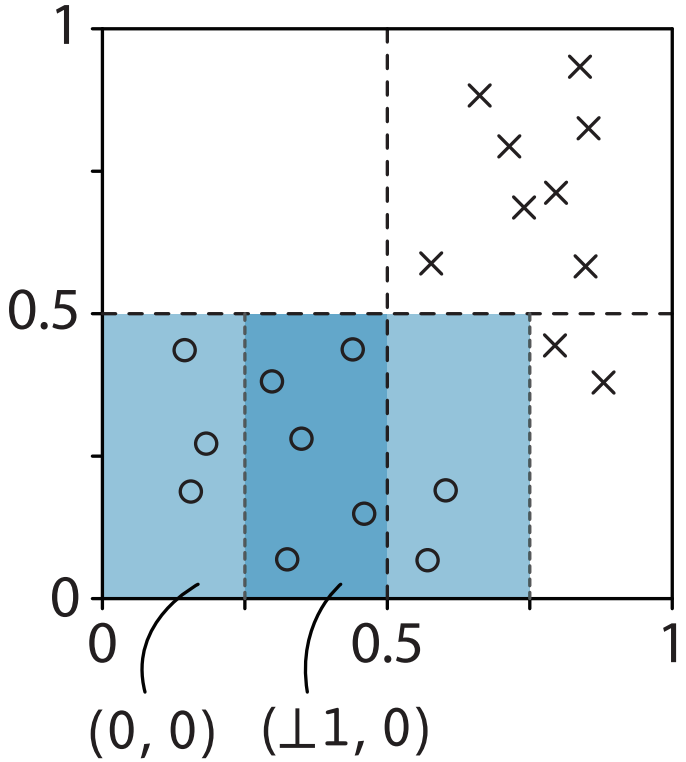


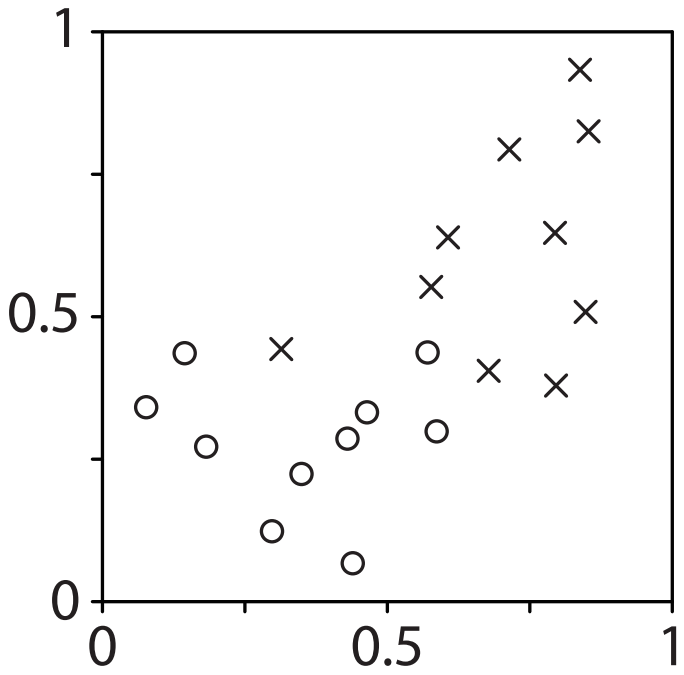
グレイ符号化ダイバージェンスの例



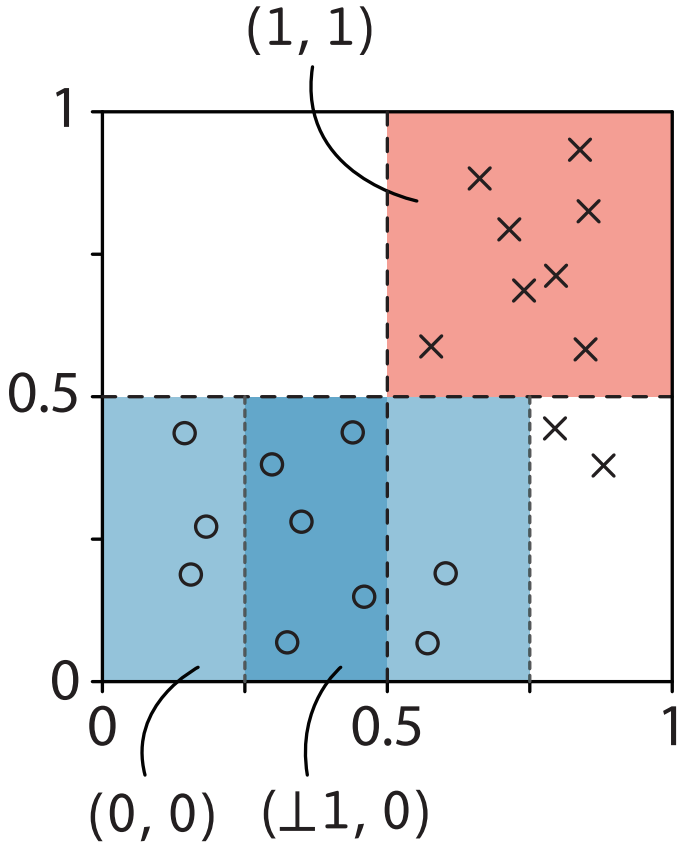


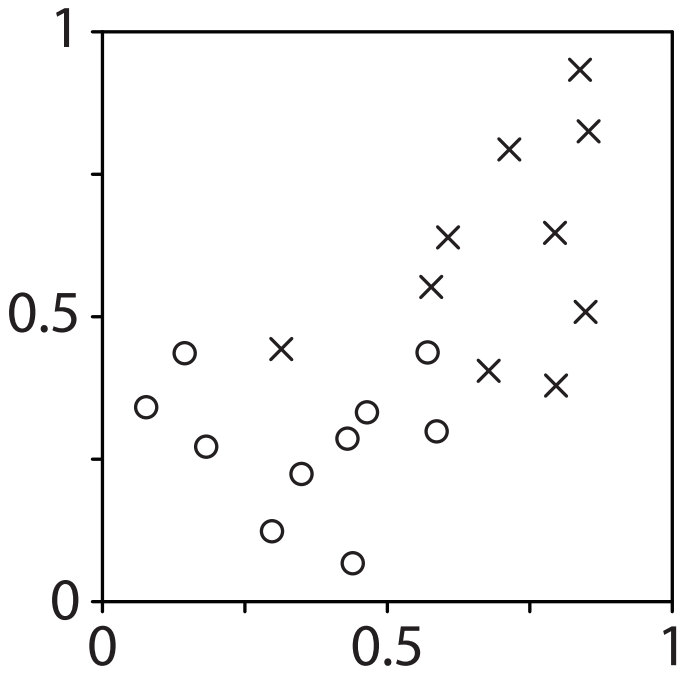
グレイ符号化ダイバージェンスの例



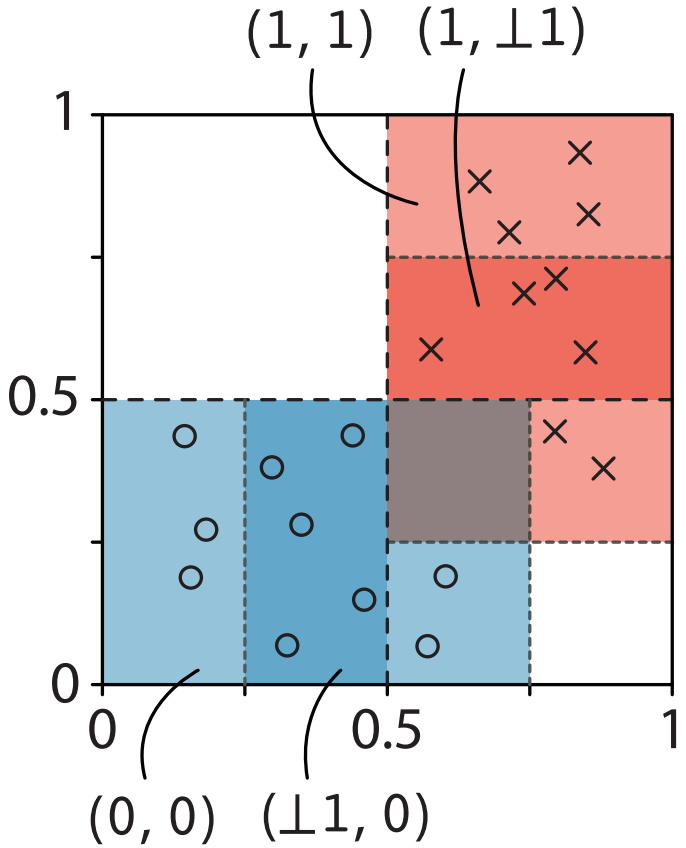


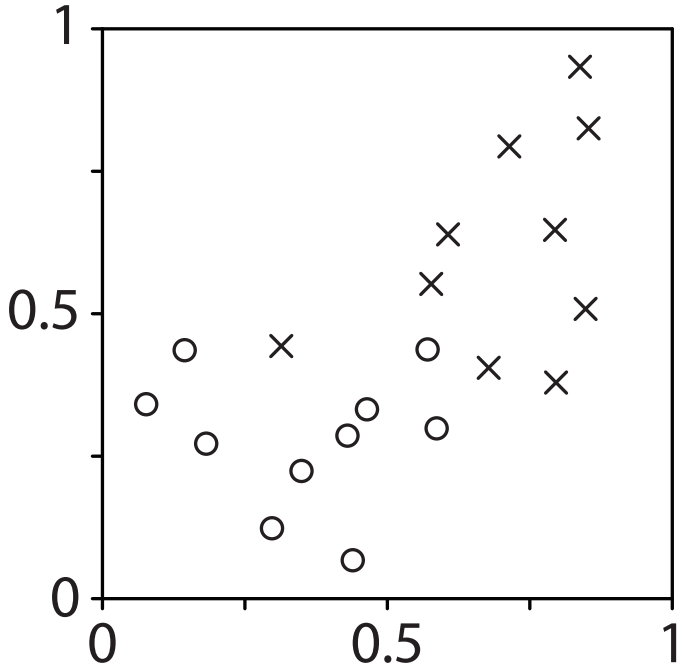
グレイ符号化ダイバージェンスの例



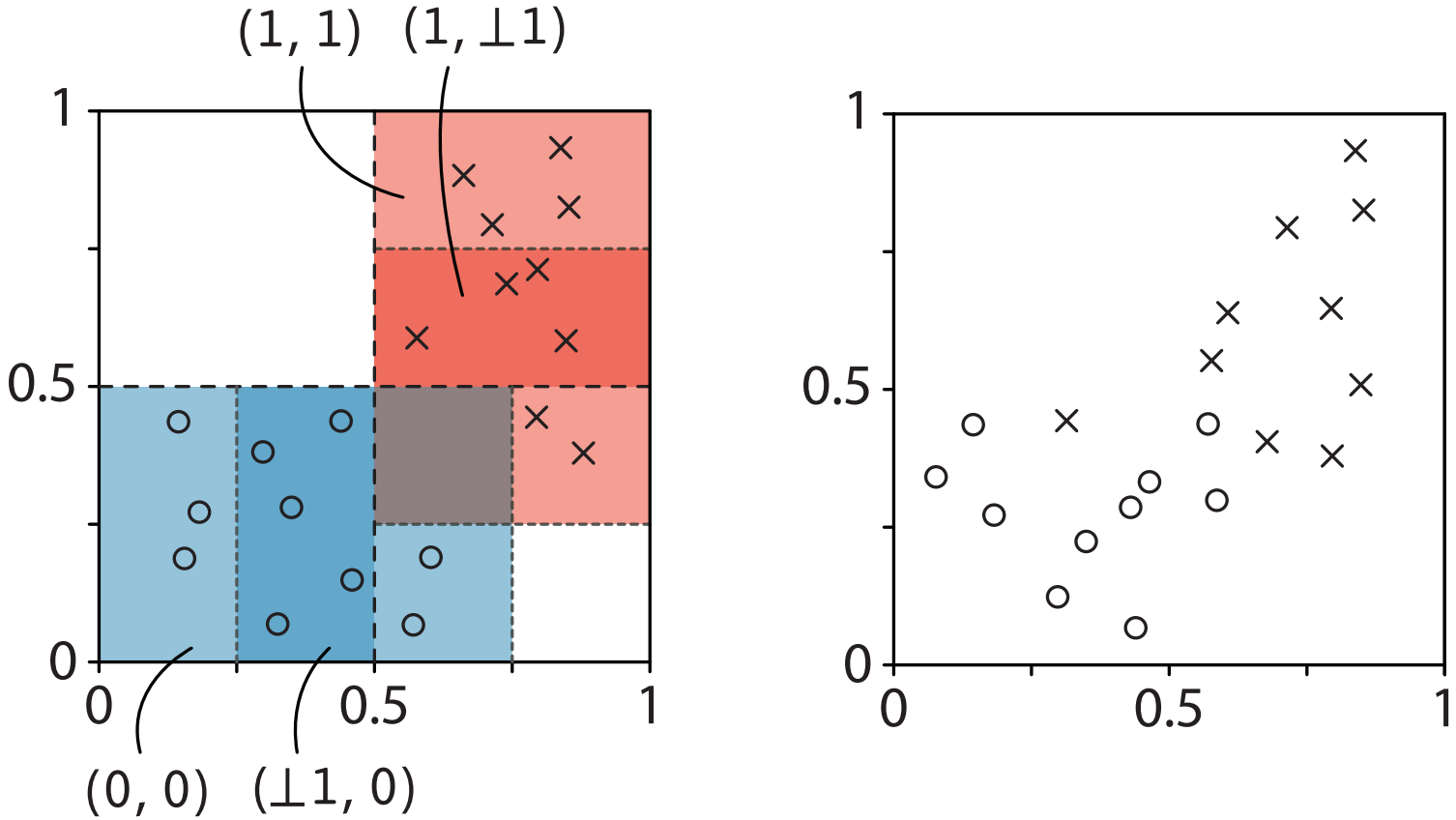


グレイ符号化ダイバージェンスの例





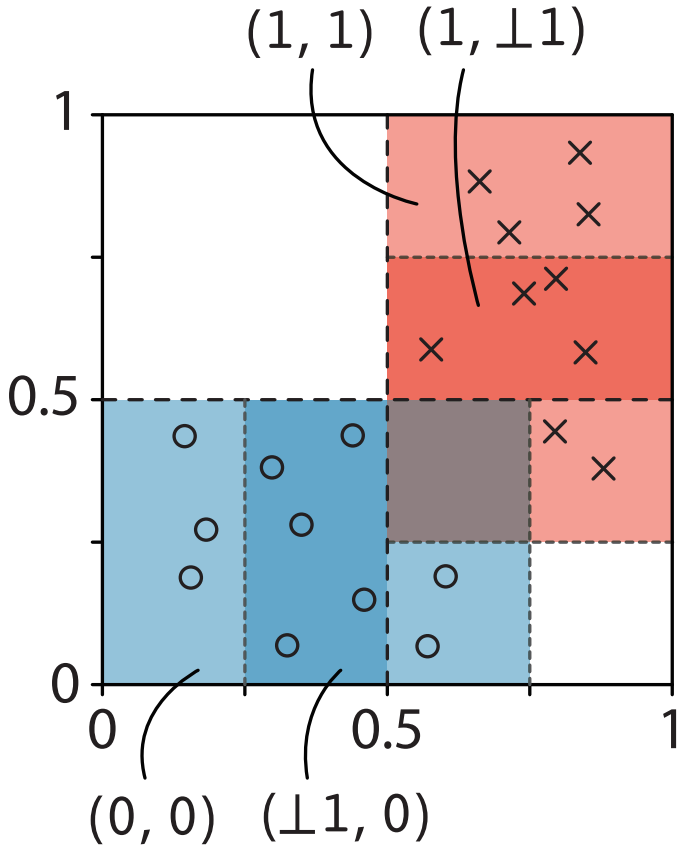
グレイ符号化ダイバージェンスの例

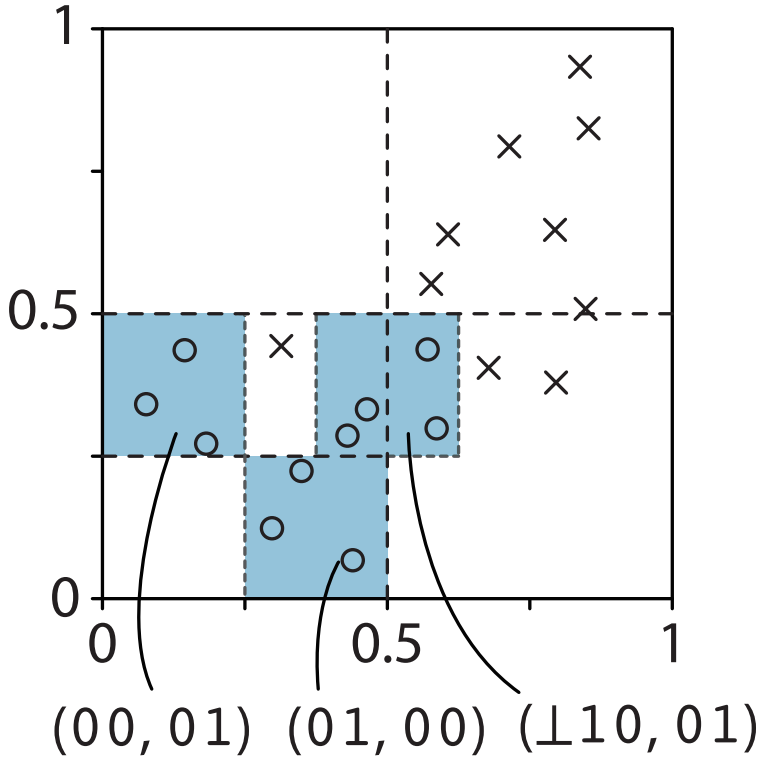


グレイ符号化ダイバージェンス：

4/10 + 4/10 = 0.8

グレイ符号化ダイバージェンスの例

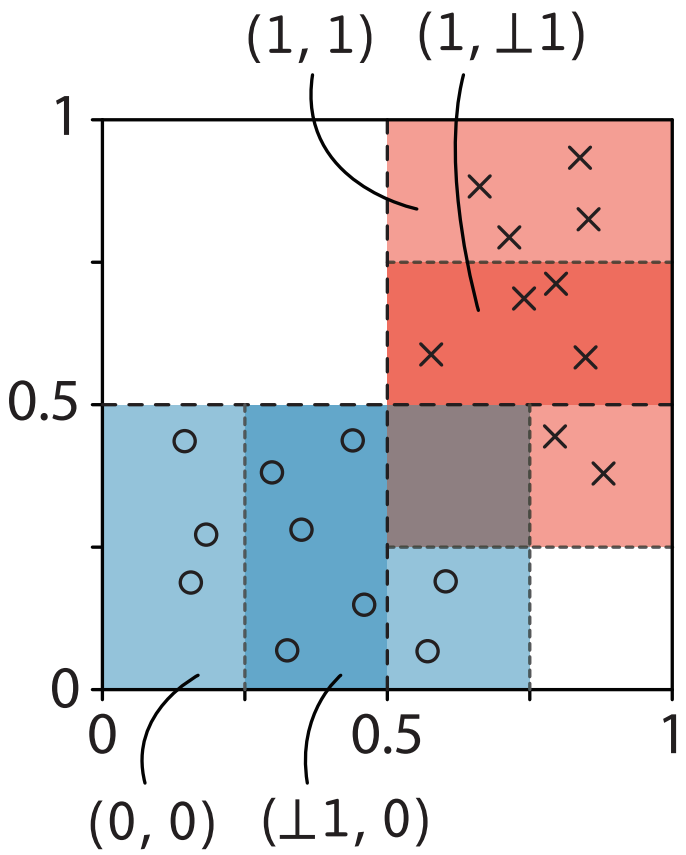


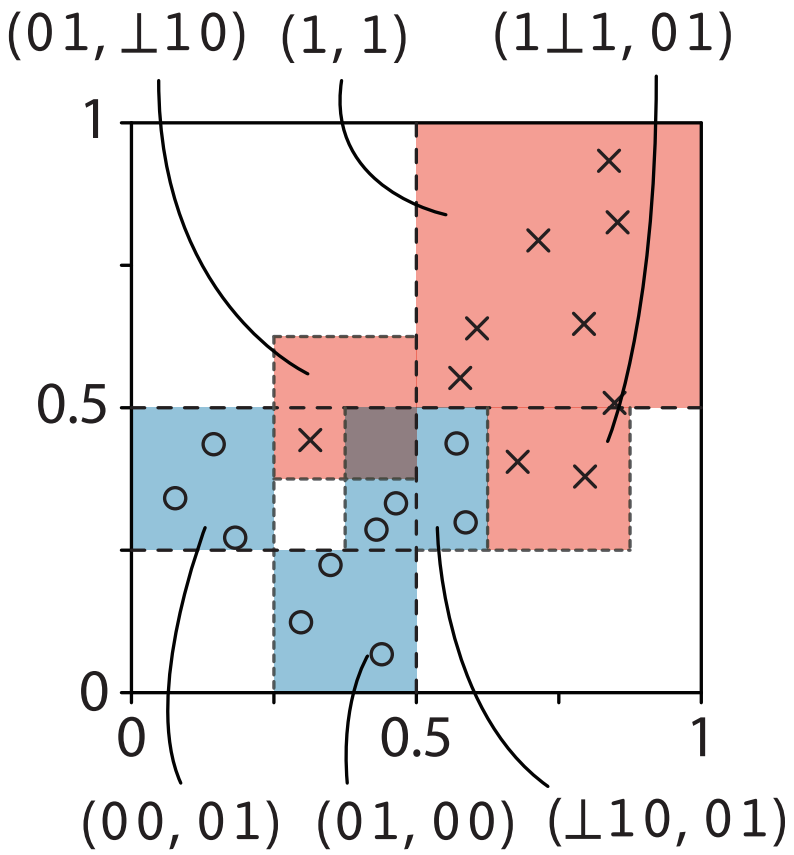


グレイ符号化ダイバージェンス：

4/10 + 4/10 = 0.8

グレイ符号化ダイバージェンスの例

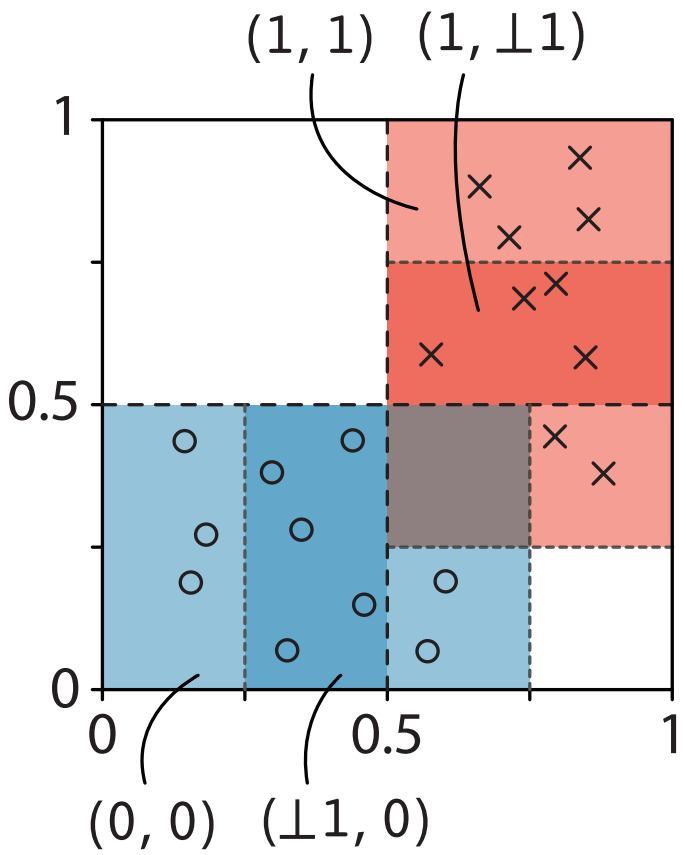


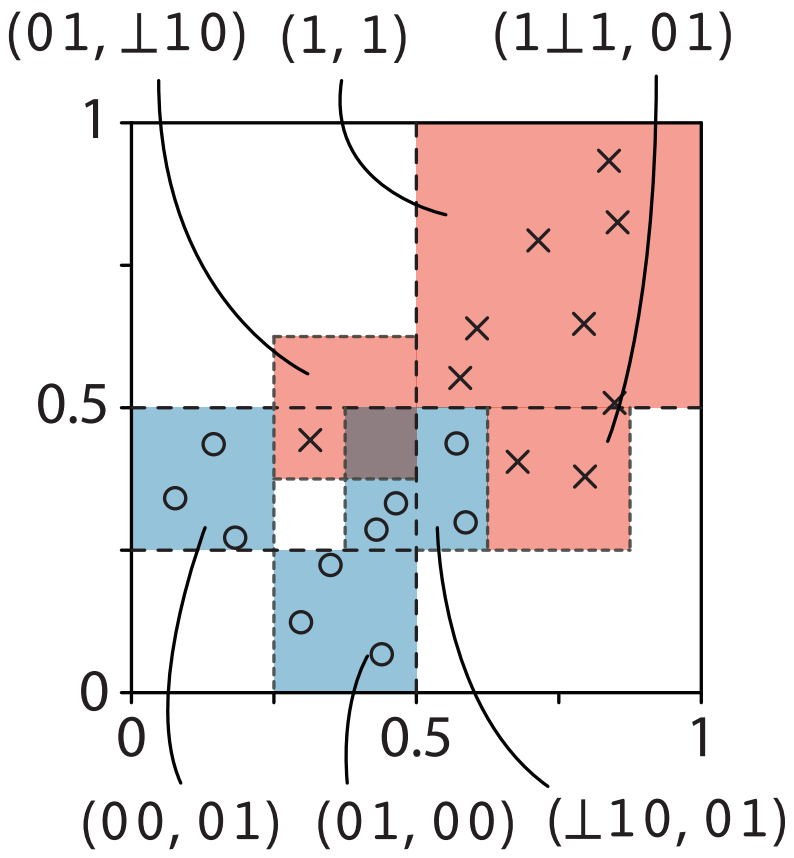


グレイ符号化ダイバージェンス：

4/10 + 4/10 = 0.8

グレイ符号化ダイバージェンスの例





グレイ符号化ダイバージェンス：

4/10 + 4/10 = 0.8

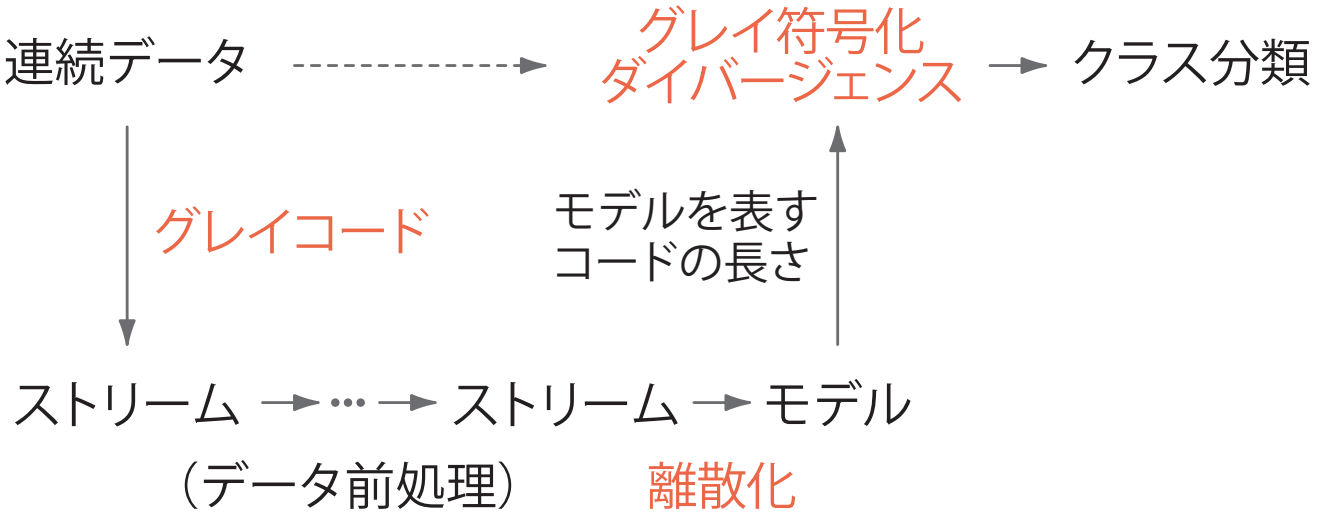
12/10 + 10/10 = 2.2

何の役に立つのか：実験科学への貢献

- 実験科学では，多くの場合コントロール群と処理群を比較する実験によって仮説を検証する
  - 例：新薬の効果を検証するために，偽薬を適用したコントロール群と新薬を適用した処理群を比較する
- 普通は統計的仮説検定（t 検定など）を使うが，実際の適用に際して問題点が多い [Johnson, 99]
  - データに対する前提条件や， p 値の任意性など
- 2 群を比較すればよいのだから，知識発見（機械学習）の文脈で扱うことができる問題
- グレイ符号化ダイバージェンスはそのための尺度となりうる
  - 1 つの解決策を前回の FPAI で提案

目次

- 研究背景
- グレイコードを用いたストリーム計算
- グレイ符号化ダイバージェンス
- クラス分類
- 実験
- まとめ

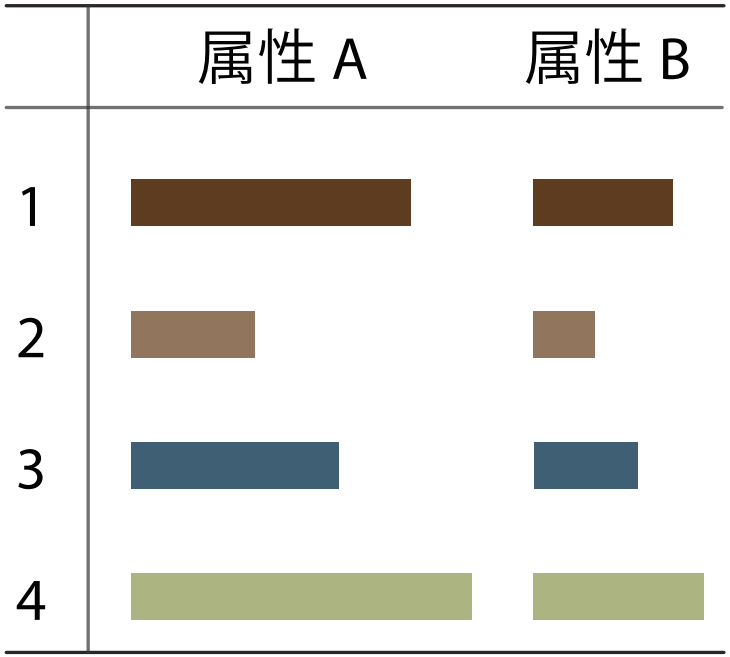


研究背景

- 近年，多くの知識が連続データ（例：実験・観測によって得られる実数値データ）から獲得されている
- しかし，連続データを対象とした手法は，実際にはアナログ形式からデジタル形式に離散化されたデータを用いる
  →理論的に仮定するデータと実際に用いるデータに差がある
- 得られた結果（獲得された知識）は，何らかの重大なエラーを含んでいる可能性がある
- このようなエラーを考慮し，知識に対して理論的正当性を与える必要がある

研究背景（図解）

連続データ（実数）



ストリーム（無限列）で符号化される

研究背景（図解）

連続データ（実数）

	属性 A	属性 B
1	1.239582…	0.6469…
2	0.426711…	0.2655…
3	1.111577…	0.4998…
4	1.801501…	0.7569…

ストリーム（無限列）で符号化される

研究背景（図解）

連続データ（実数）

	属性 A	属性 B
1	1.239582…	0.6469…
2	0.426711…	0.2655…
3	1.111577…	0.4998…
4	1.801501…	0.7569…

ストリーム（無限列）で符号化される

離散化

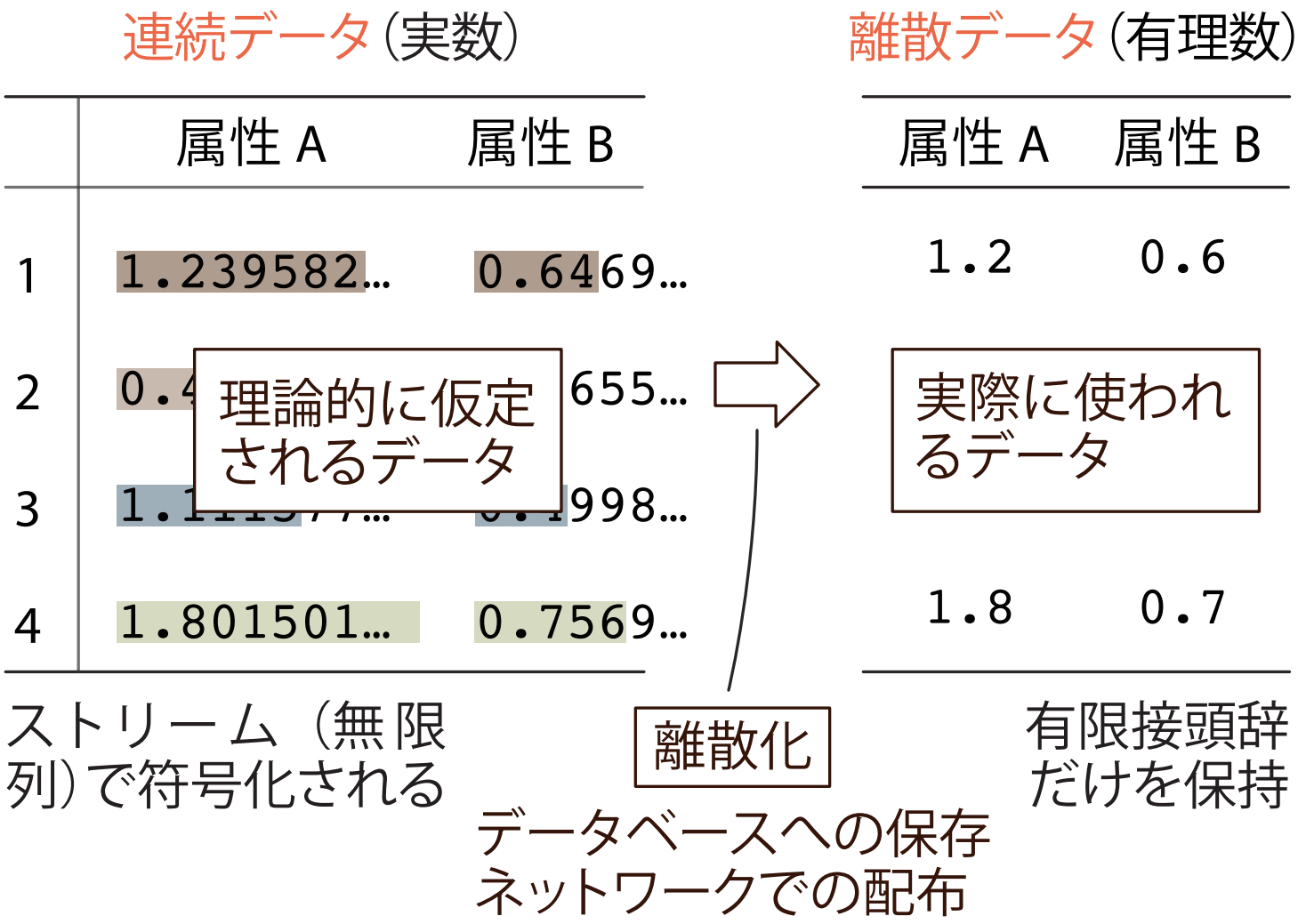
データベースへの保存
ネットワークでの配布

離散データ（有理数）

属性 A	属性 B
1.2	0.6
0.4	0.2
1.1	0.4
1.8	0.7

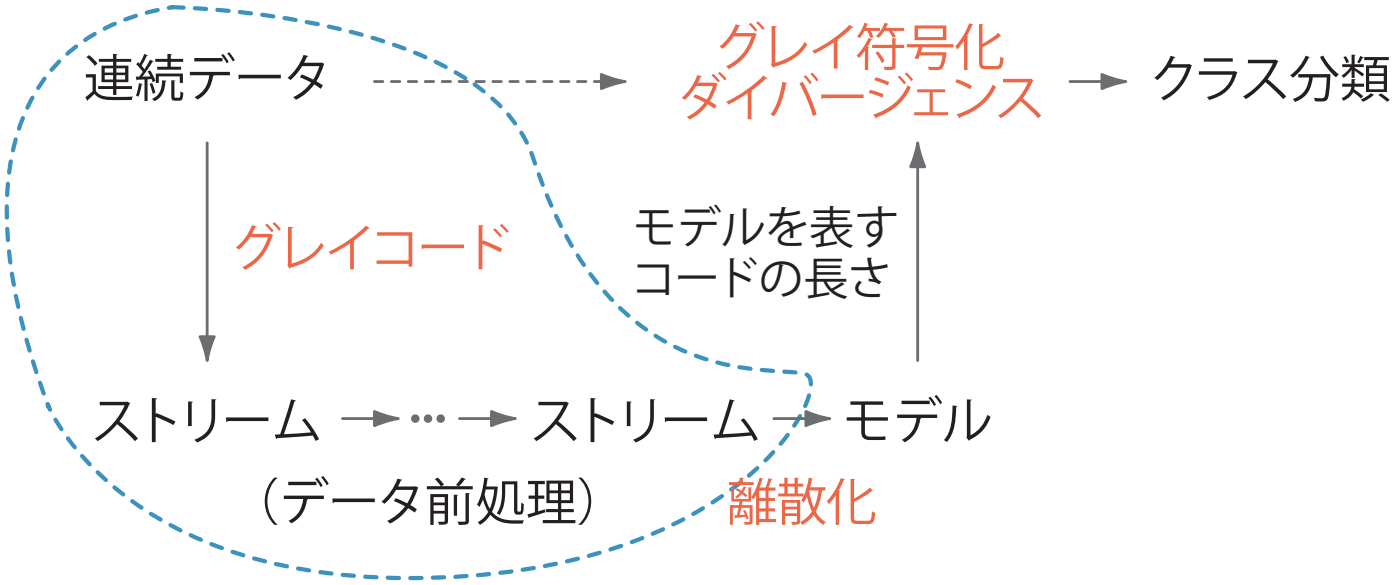
有限接頭辞だけを保持

研究背景（図解）


連続データ（実数）
離散データ（有理数）
属性 A
属性 B
属性 A
属性 B
1
1.239582…
0.6469…
1.2
0.6
2
655…
3
998…
4
1.801501…
0.7569…
1.8
0.7
理論的に仮定されるデータ
実際に使われるデータ
ストリーム（無限列）で符号化される
離散化
有限接頭辞だけを保持
データベースへの保存
ネットワークでの配布

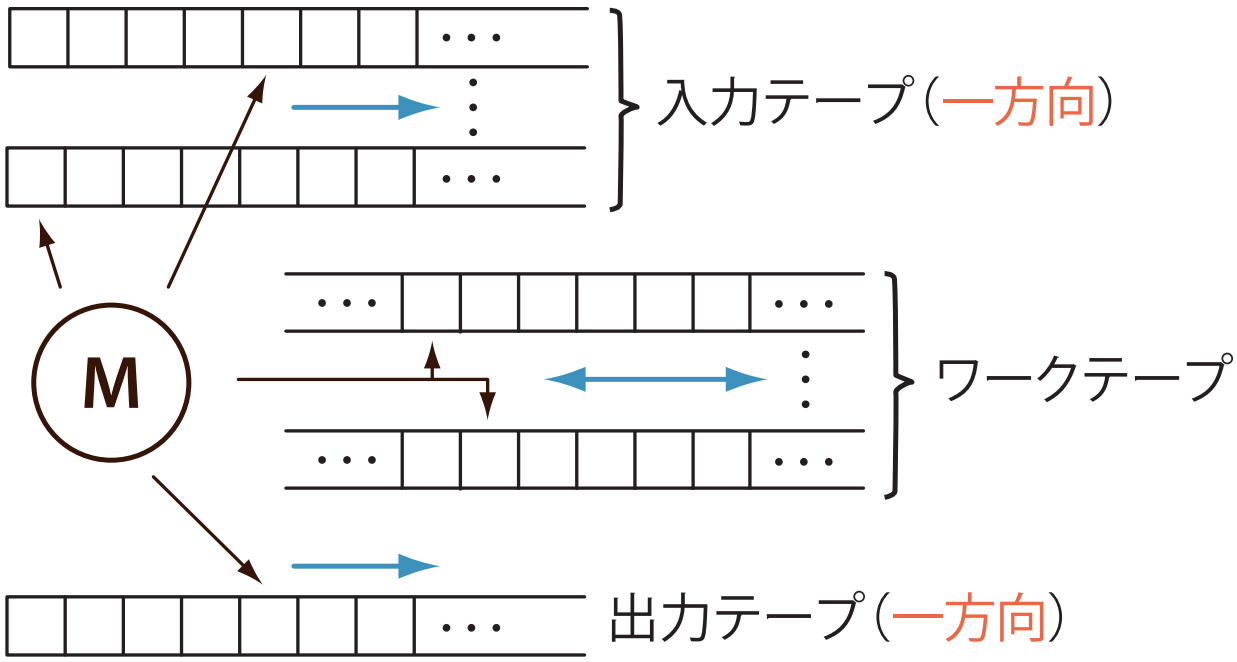
目次

- 研究背景
- グレイコードを用いたストリーム計算
- グレイ符号化ダイバージェンス
- クラス分類
- 実験
- まとめ



タイプ 2 マシンによるストリーム計算

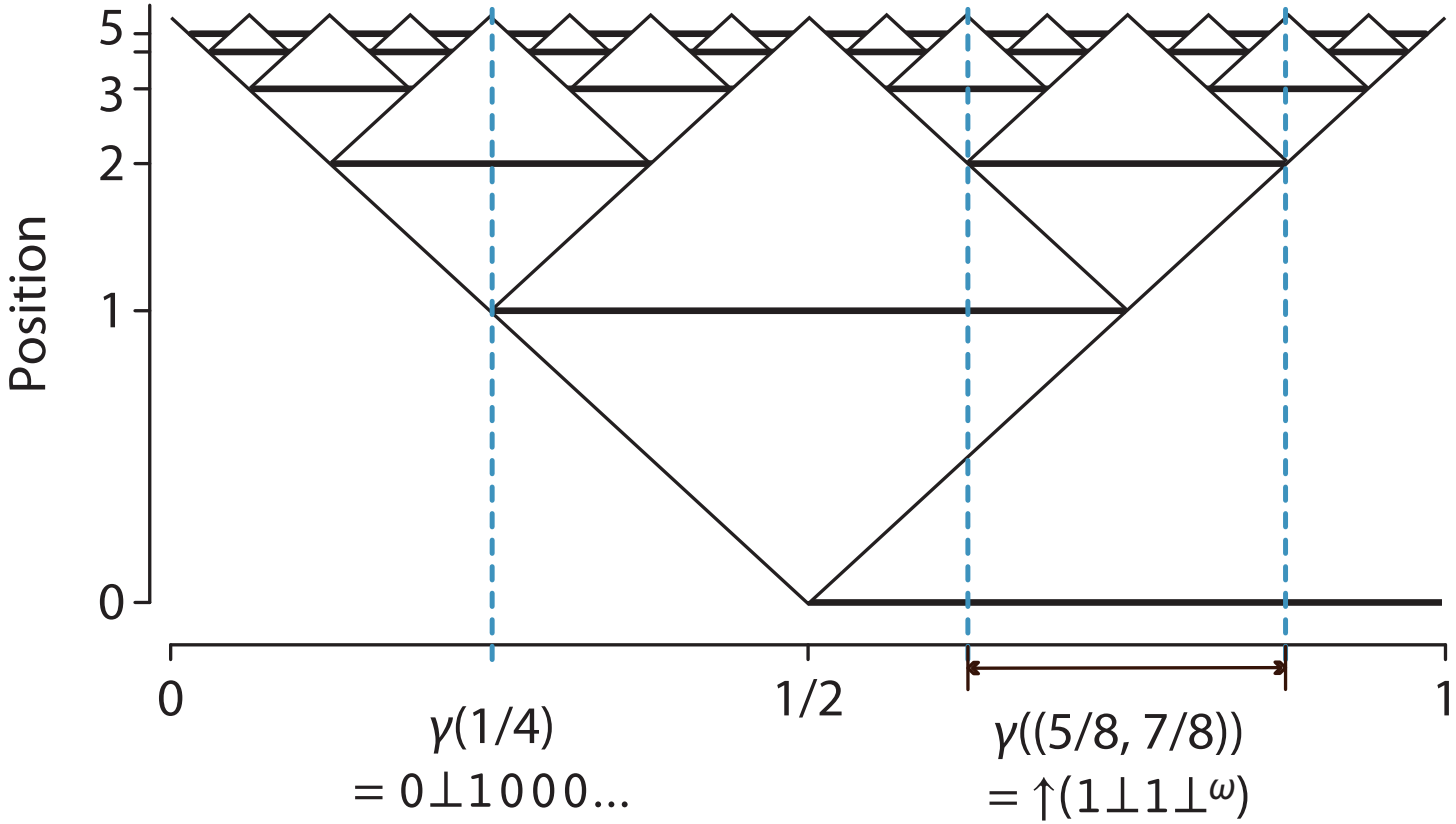
- ストリーム間の計算はタイプ 2 マシンによって実現される
- 計算可能性はコーディングの方法に依存する
  - 例：10 進表現や 2 進表現では $y = 3x$ が計算できない



なぜグレイコードなのか？

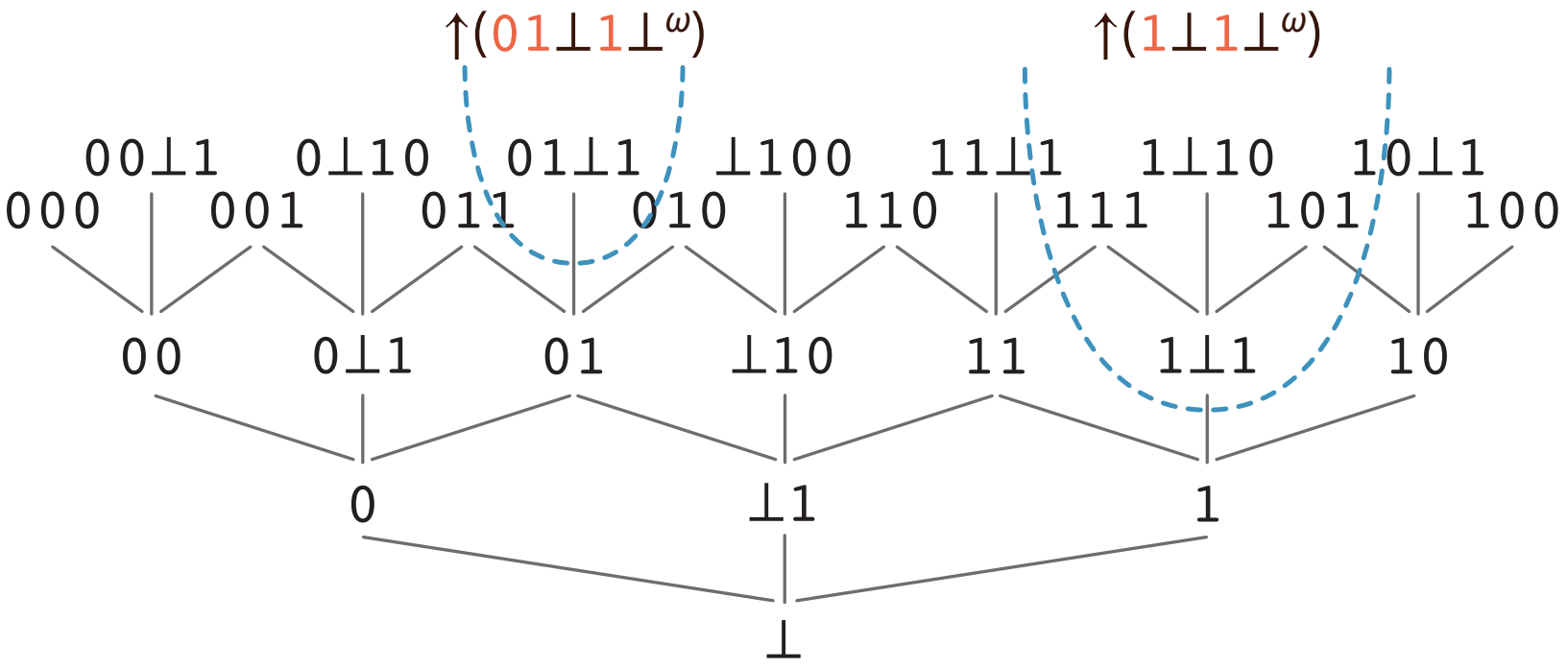
- 実数の小数点以下を 0, 1, ⊥ からなるストリームで符号化
  - ⊥ は高々1 回だけ出現し，⊥ の後は必ず 1000 …
- グレイコードは，実数の標準的な計算可能性を導く符号化の 1 つ [Tsuiki, 2002]
  - その他の代表的なものは，符号付き 2 進表現
- 単射になるのはグレイコードだけ
  - 1 つの実数に，1 つのストリームが対応する
  - 例：2 進表現では，0.25 は 00111 … と 01000 … があるが，グレイコードでは 0⊥1000 … だけ

グレイコードによる実数のストリーム表現



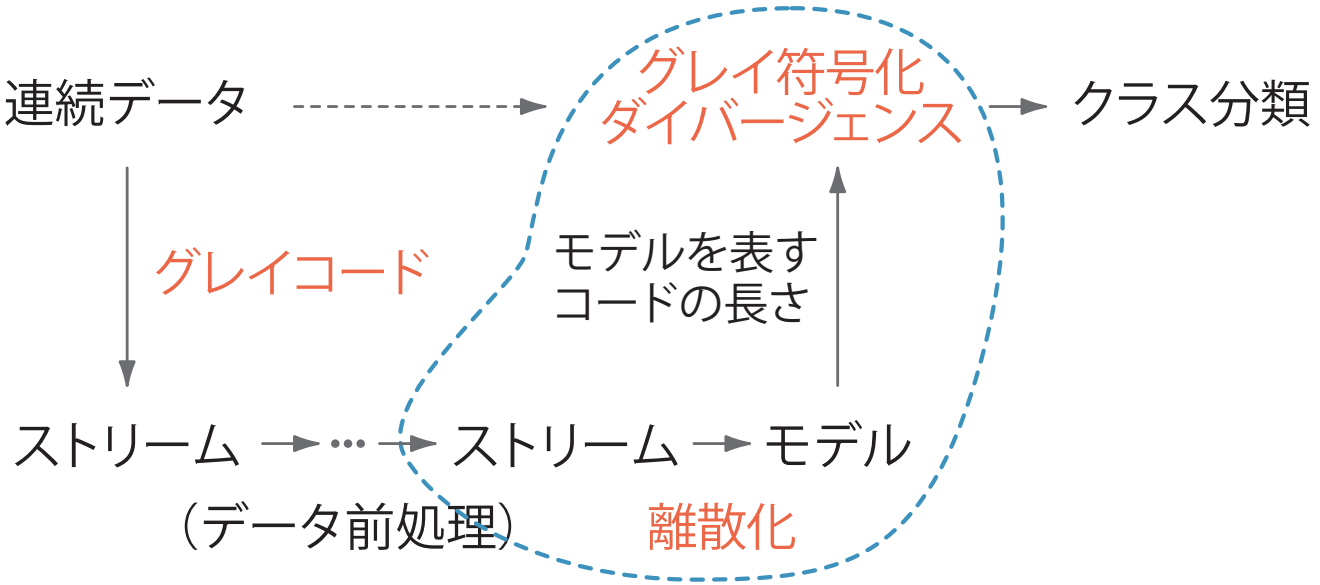
グレイコードの位相的性質

- グレイコードで符号化されたストリームがなす位相空間
- 接頭辞が決定されたストリーム（離散化された連続データ）は開集合の基底に対応する



目次

- 研究背景
- グレイコードを用いたストリーム計算
- グレイ符号化ダイバージェンス
- クラス分類
- 実験
- まとめ



定式化

- X と Y を単位区間 $\mathcal{I} = [0,1] \times \ldots \times [0,1] \subset \mathbb{R}^d$ の有限部分集合とする
  - それぞれ（正規化された）実数値データの集合に対応
- それらのデータ集合 X, Y をグレイコードによってストリームの集合 $\gamma(X)$, $\gamma(Y)$ として扱う
  - ユークリッド空間 $\mathbb{R}^d$ を γ で文字列空間に埋め込む
- ストリームの集合 $\gamma(X)$ を正例の集合，$\gamma(Y)$ を負例の集合とみなし，それらに無矛盾かつ最も単純なモデルを学習する
  - ここでモデルとは，文字列空間上の開集合
    - 開集合は接頭辞が決定されたストリームの集合なので，有限時間で計算論的に学習できる

グレイ符号化ダイバージェンス

- グレイ符号化ダイバージェンスを以下のように定義

$$C(X, Y) := \begin{cases} \infty & \text{if } X \cap Y \neq \emptyset, \\ D(X; Y) + D(Y; X) & \text{otherwise,} \end{cases}$$

  - ここで D は有向グレイ符号化ダイバージェンス

$$D(X; Y) := \frac{1}{\|X\|} \min \{ |O| \mid O \text{は開かつ} (\gamma(X), \gamma(Y)) \text{に無矛盾} \}$$

    - $\|X\|$ は X の要素数
    - R が (P, Q) に無矛盾 $\iff$ $R \supseteq P$ かつ $R \cap Q = \emptyset$

- グレイ符号化ダイバージェンスは文字列空間の位相的構造にのみ依存
  - 統計理論や確率モデルを全く使わない知識発見・機械学習

学習手続き

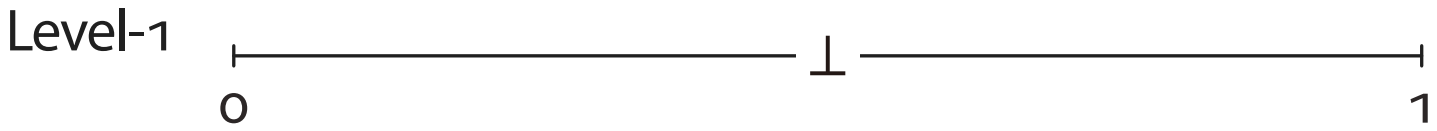
function MAIN($\gamma(X), \gamma(Y)$)
  LEARNING($\gamma(X), \gamma(Y), 0, 0, \|X\|, \|Y\|, 0$)

function LEARNING(P, Q, D_1, D_2, m, n, k)
  $V \leftarrow$ DISCRETIZE(P, k), $\quad W \leftarrow$ DISCRETIZE(Q, k)
  $V_{\mathrm{sep}} \leftarrow \{v \in V \mid v \notin W\}$, $\quad W_{\mathrm{sep}} \leftarrow \{w \in W \mid w \notin V\}$
  $D_1 \leftarrow D_1 + \min_{V \in \boldsymbol{V}} \sum_{v \in V} |v|$ $(\boldsymbol{V} = \{V' \subseteq V_{\mathrm{sep}} \mid f_P(V') = f_P(V_{\mathrm{sep}})\})$
  $D_2 \leftarrow D_2 + \min_{W \in \boldsymbol{W}} \sum_{w \in W} |w|$ $(\boldsymbol{W} = \{W' \subseteq W_{\mathrm{sep}} \mid f_P(W') = f_P(W_{\mathrm{sep}})\})$
  output $D_1/m + D_2/n$
  $P \leftarrow \{p \in P \mid p \notin f_P(V_{\mathrm{sep}})\}$, $\quad Q \leftarrow \{q \in Q \mid q \notin f_Q(W_{\mathrm{sep}})\}$
  if $P = \emptyset$ and $Q = \emptyset$ **then** halt
  else return LEARNING($P, Q, D_1, D_2, m, n, k+1$)

function DISCRETIZE(P, k)
  return $\{w \mid |w| = n,\ p \in \uparrow(w\perp^{\omega}),\ p \in P\}$ $\quad (n = (k+1)d - 1)$

（ただし，$f_P(V) = \{p \in P \mid \exists v \in V.\, p \in \uparrow(v\perp^{\omega})\}$, $P \subset \Sigma^{\omega}_{\perp,d}$, $V \subset \Sigma^{*}$）

学習例

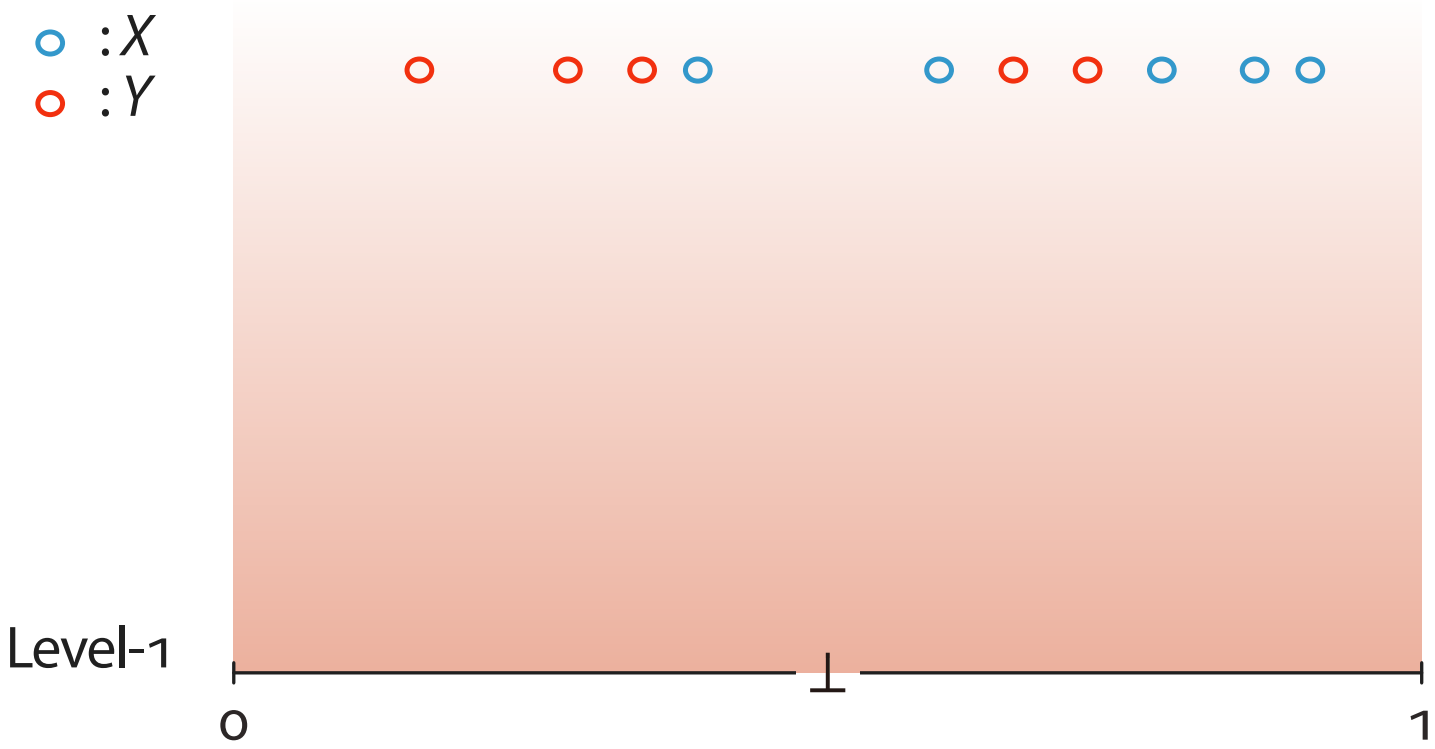


$D(X; Y) \rightarrow \{$ $\}$

$D(Y; X) \rightarrow \{$ $\}$

$C(X, Y) =$

学習例

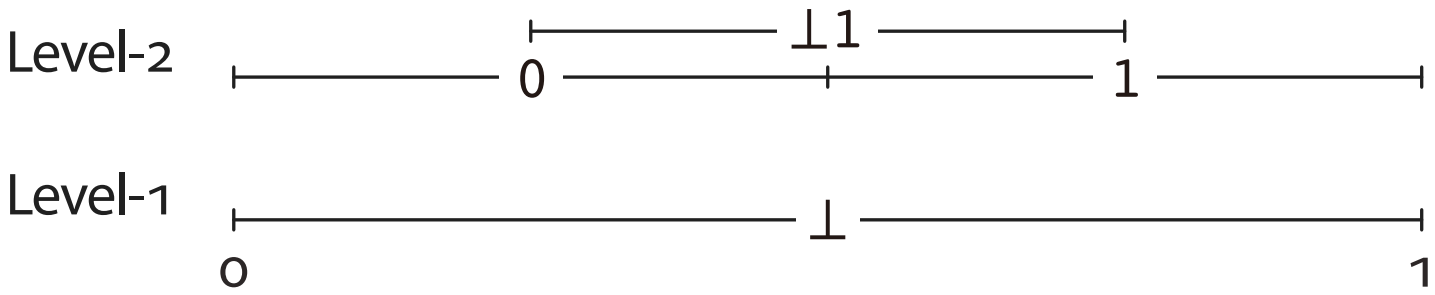


$D(X; Y) \rightarrow \{$ $\}$

$D(Y; X) \rightarrow \{$ $\}$

$C(X, Y) =$

学習例

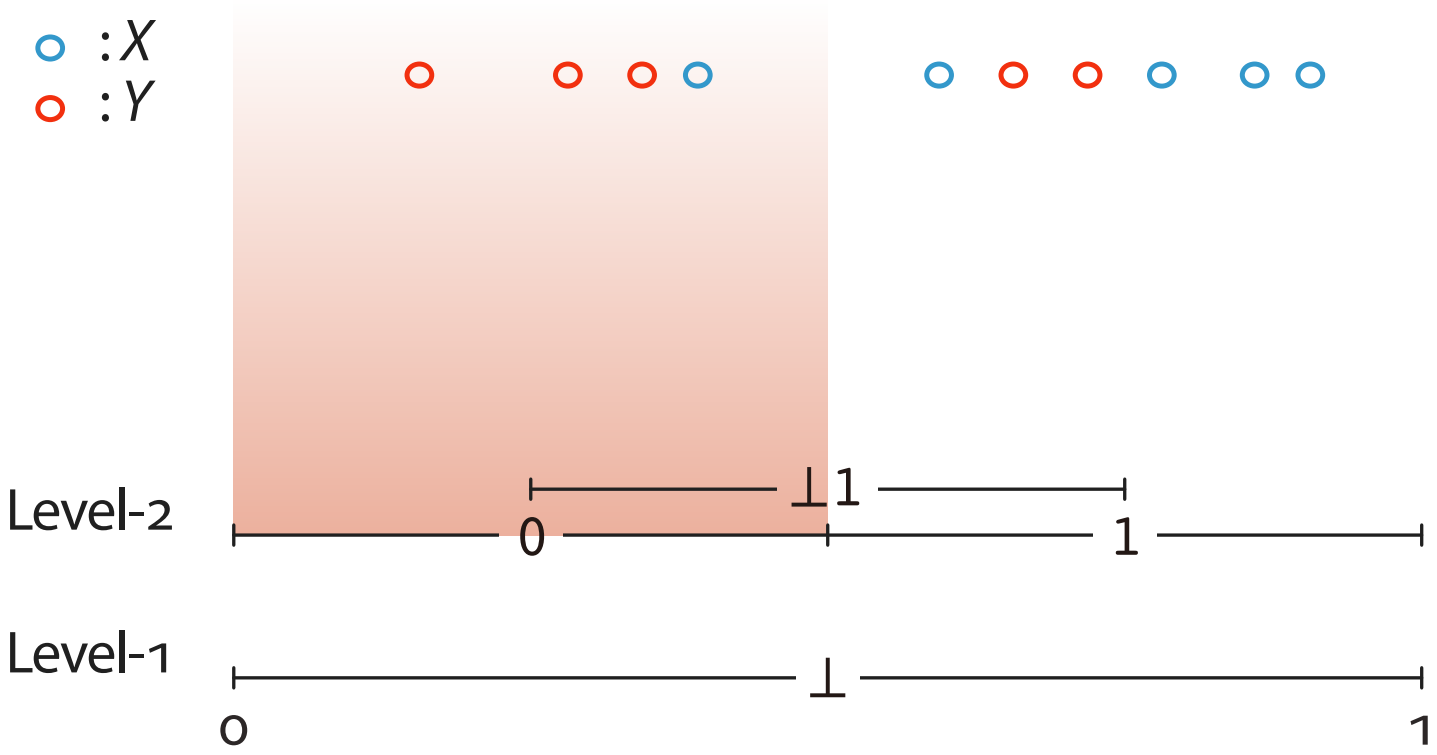


$D(X; Y) \rightarrow \{$ $\}$

$D(Y; X) \rightarrow \{$ $\}$

$C(X, Y) =$

学習例

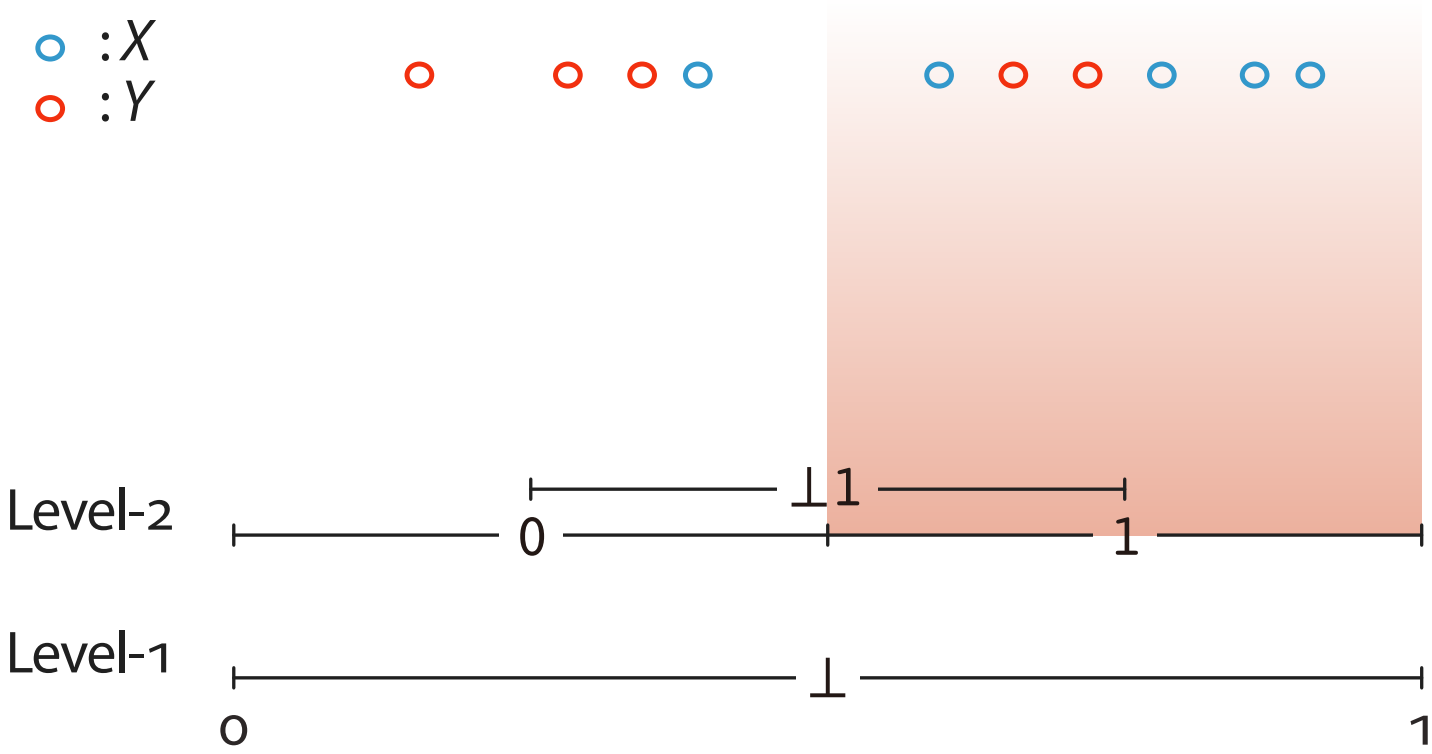


$D(X; Y) \rightarrow \{$ $\}$

$D(Y; X) \rightarrow \{$ $\}$

$C(X, Y) =$

学習例

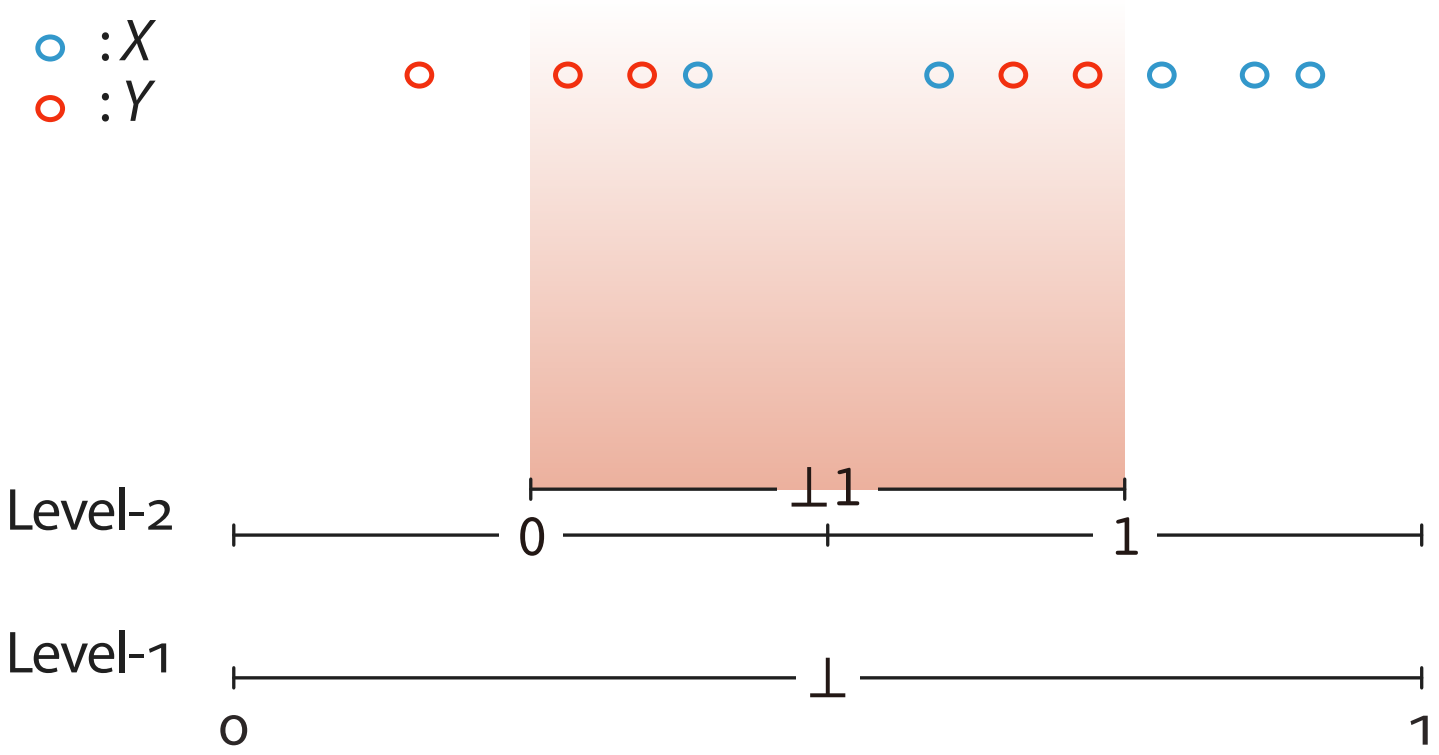


$D(X; Y) \rightarrow \{$ $\}$

$D(Y; X) \rightarrow \{$ $\}$

$C(X, Y) =$

学習例

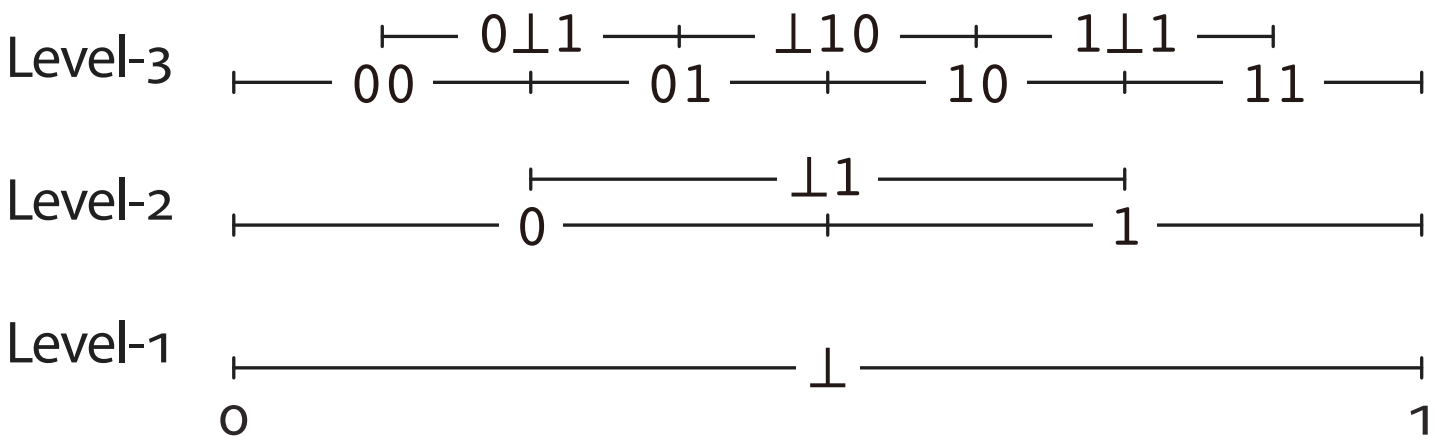


$D(X; Y) \rightarrow \{$ $\}$

$D(Y; X) \rightarrow \{$ $\}$

$C(X, Y) =$

学習例

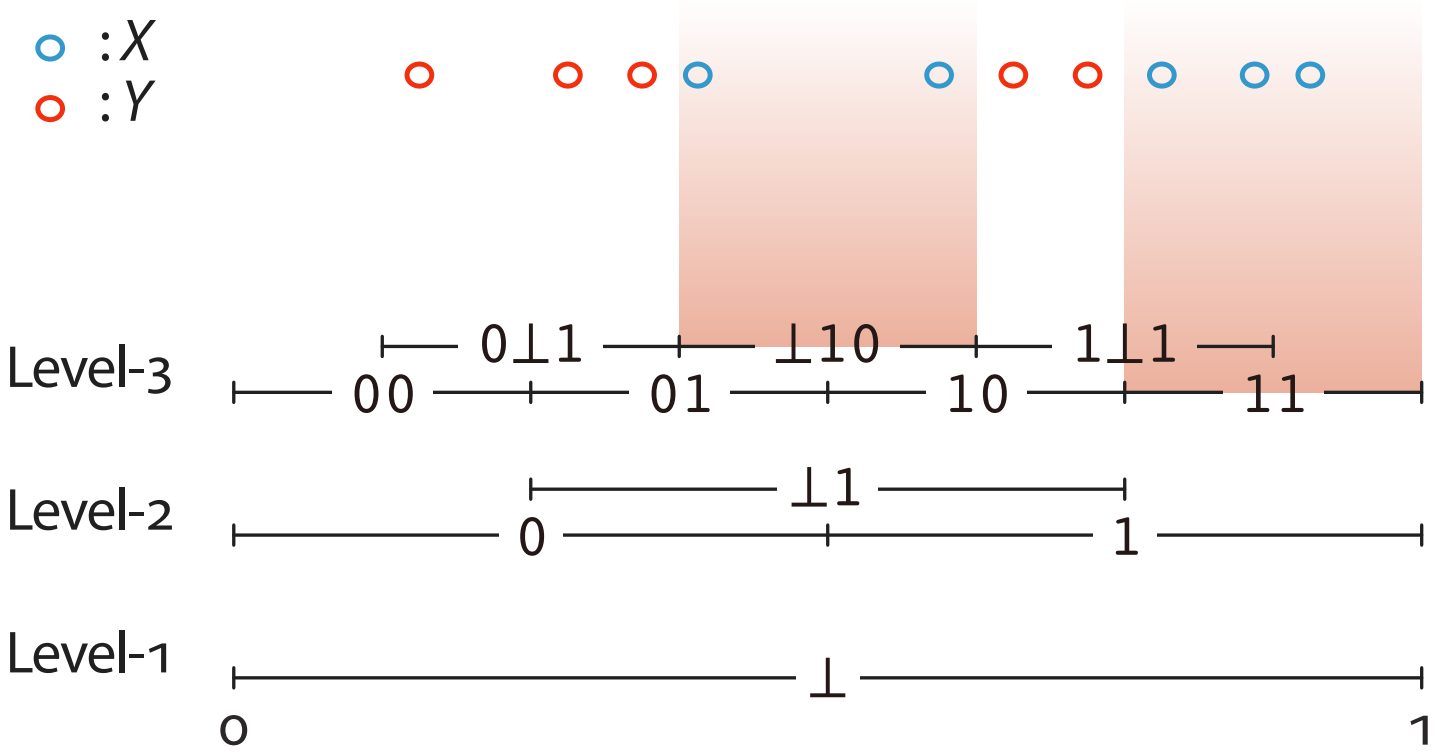


$D(X; Y) \rightarrow \{$ $\}$

$D(Y; X) \rightarrow \{$ $\}$

$C(X, Y) =$

学習例

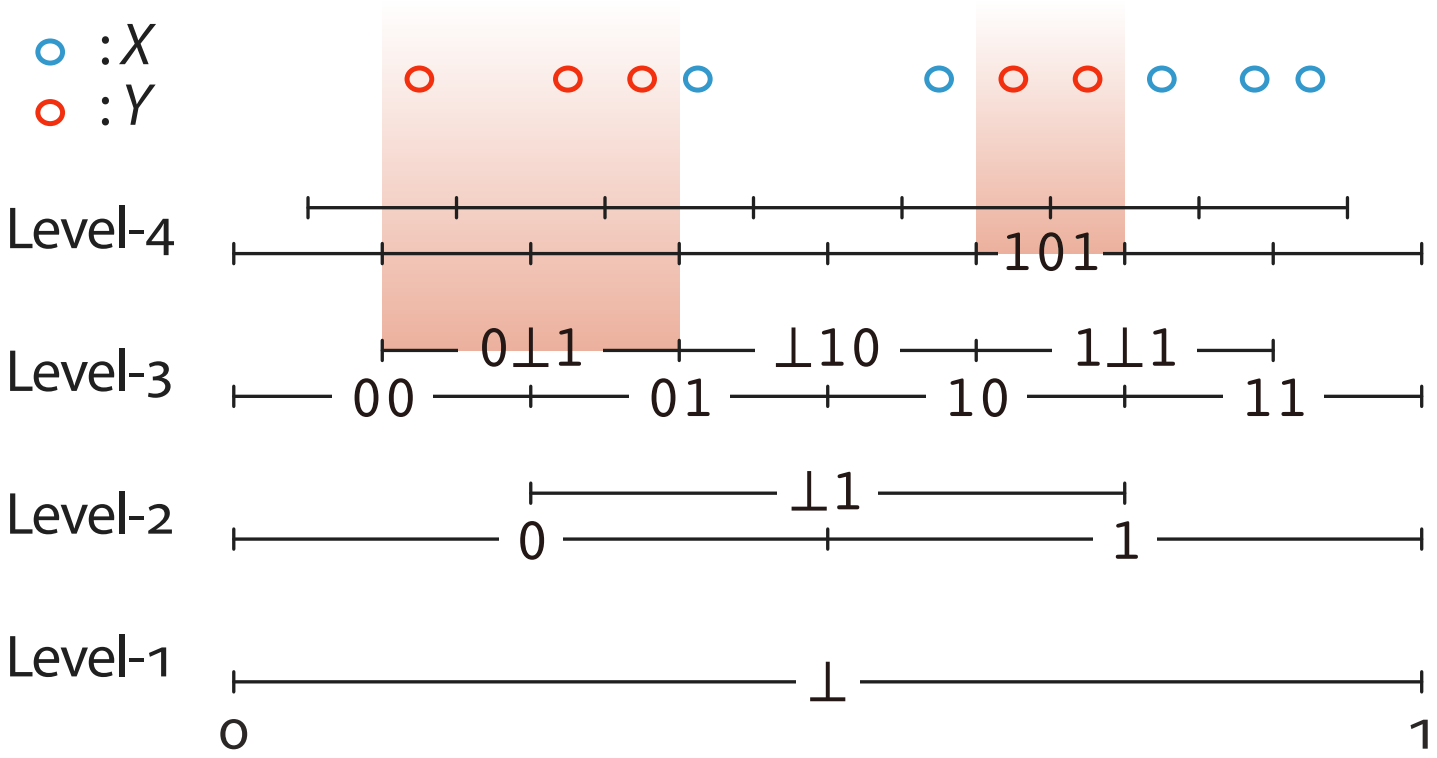


$D(X; Y) \rightarrow \{\perp 10, 11\}$

$D(Y; X) \rightarrow \{\quad\quad\quad\quad\quad\quad\}$

$C(X, Y) =$

学習例



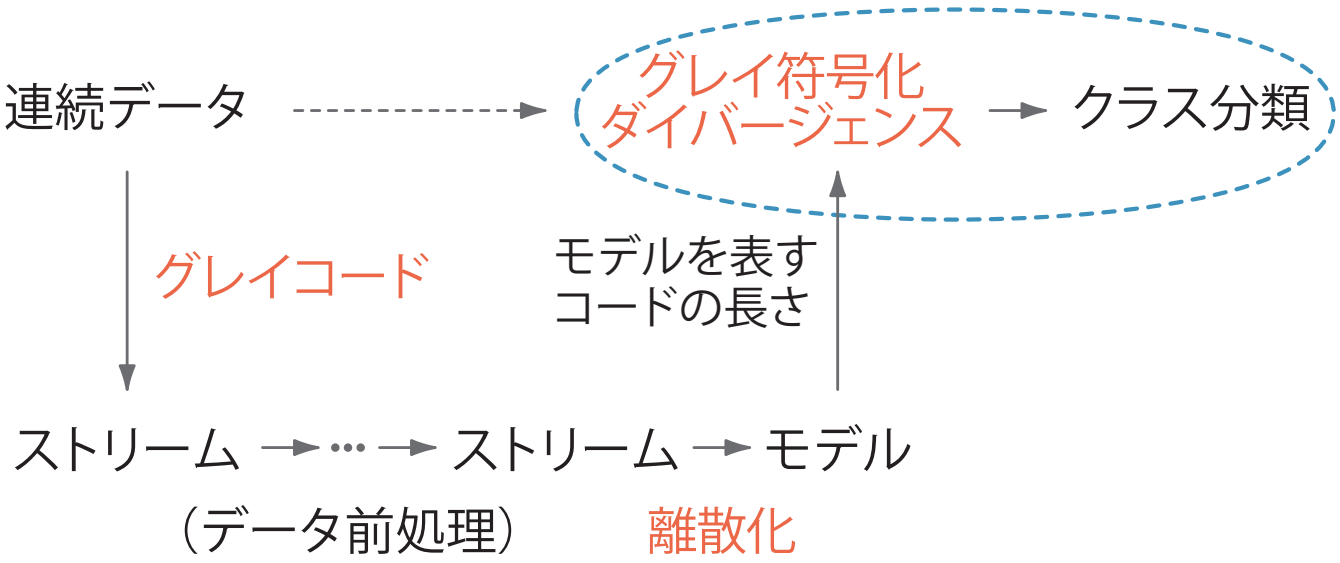
$D(X; Y) \rightarrow \{\perp 10, 11\}$

$D(Y; X) \rightarrow \{0 \perp 1, 101\}$

$C(X, Y) = 4/5 + 5/5 = 1.8$

目次

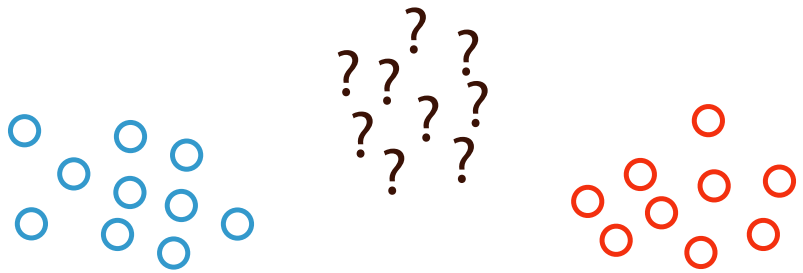
- 研究背景
- グレイコードを用いたストリーム計算
- グレイ符号化ダイバージェンス
- クラス分類
- 実験
- まとめ



グレイ符号化ダイバージェンスで分類

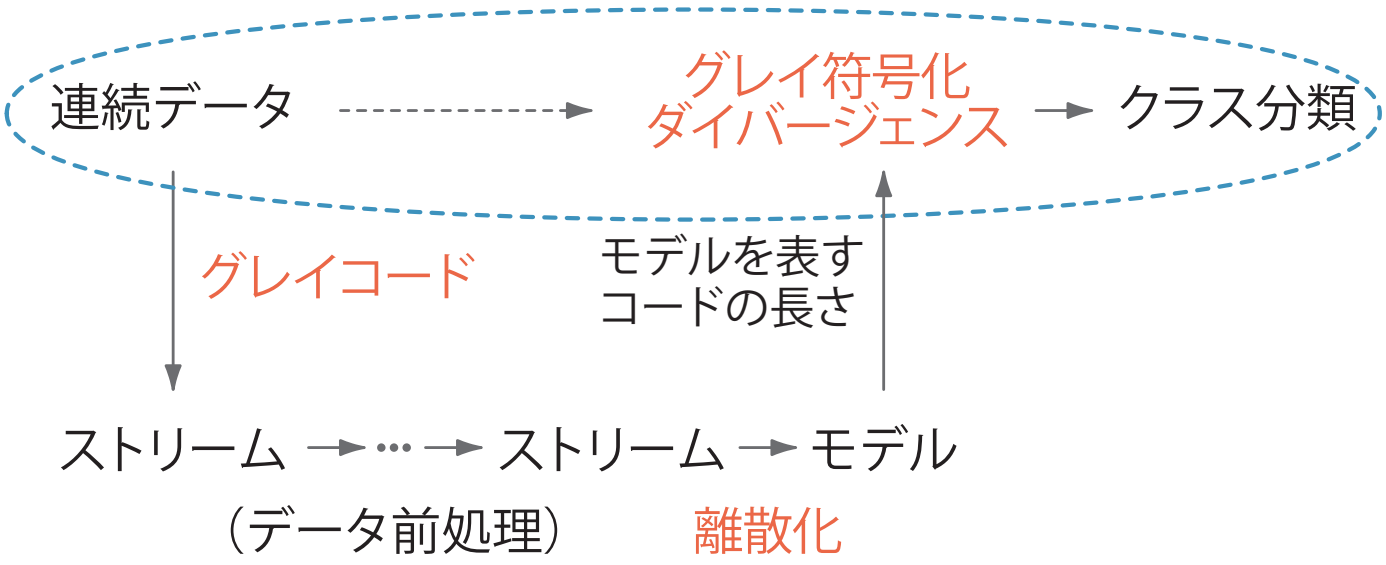
- グレイ符号化ダイバージェンスを用いた怠惰学習によってクラス分類をおこなう
- この分類器は，訓練データ X と Y（それぞれラベルは A と B とする）を受け取り，テストデータ Z を A か B へ分類する

$$Z \text{ は } \begin{cases} A \text{ に属する} & \text{if } C(X, Z) > C(Y, Z), \\ B \text{ に属する} & \text{otherwise.} \end{cases}$$



目次

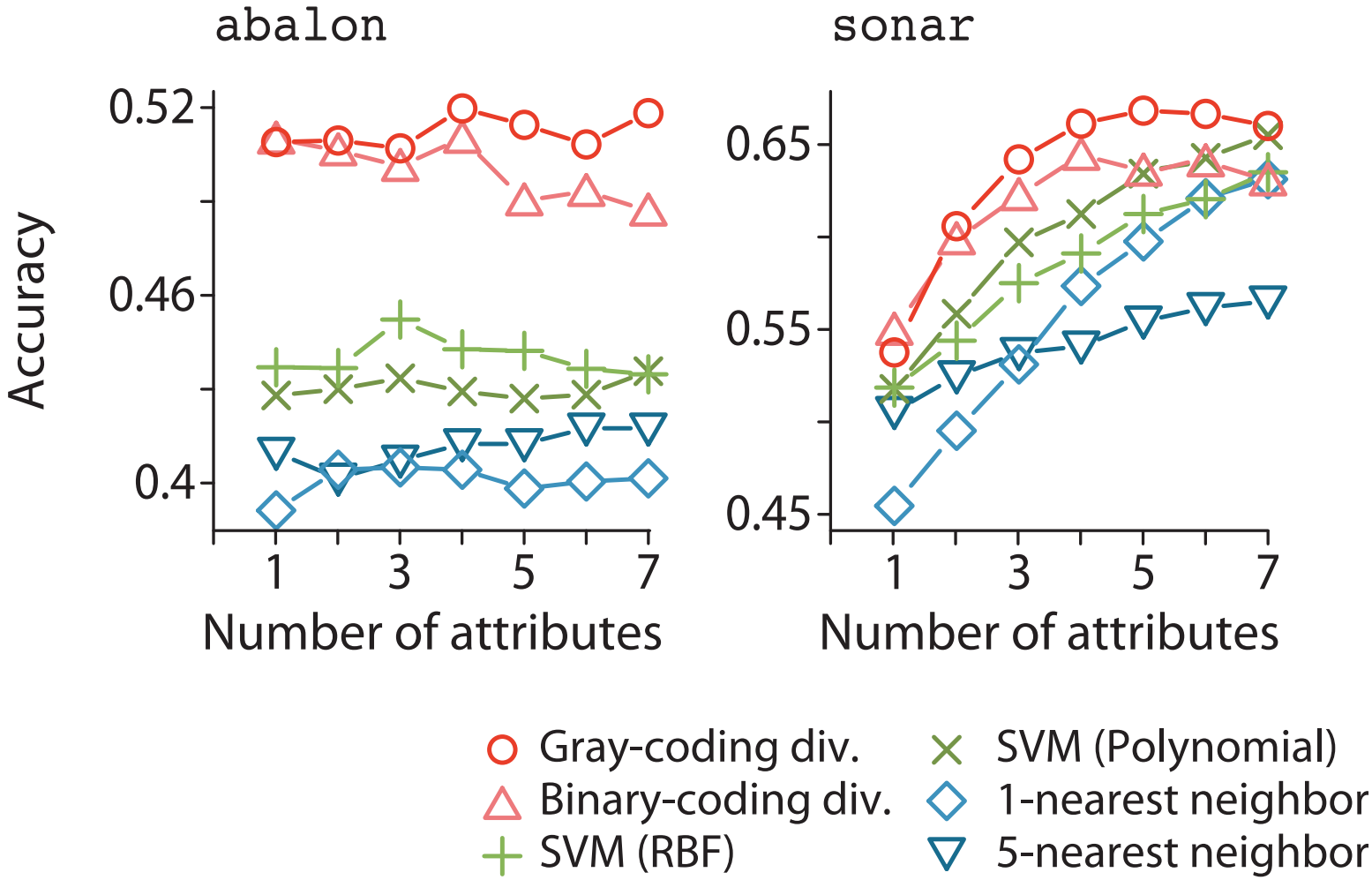
- 研究背景
- グレイコードを用いたストリーム計算
- グレイ符号化ダイバージェンス
- クラス分類
- 実験
- まとめ



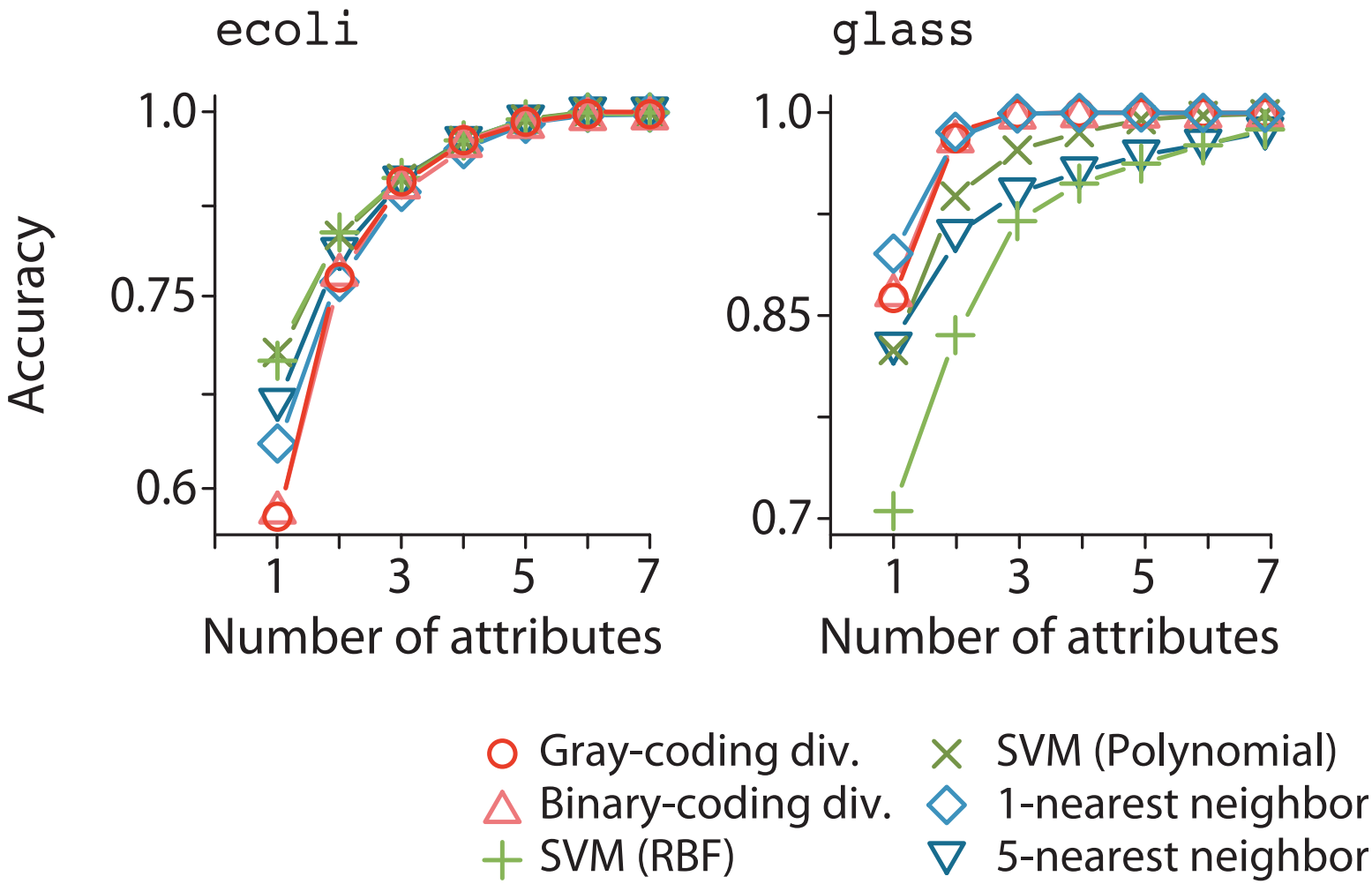
実験手法

- 分類器は R 2.10.1 で実装
- UCI のデータセット（`abalon`, `sonar`, ...）を用いる
- 以下を 10,000 回繰り返して，sensitivity と specificity から accuracy を求める
  - 識別に用いる属性をランダムに決める
  - 双方のラベルからそれぞれランダムに（非復元で） n 個サンプリングを 2 回（X, T_+ と Y, T_-）
    - X, Y は訓練データ， T_+, T_- はテストデータ
  - データを正規化（min-max normalization）
  - 本手法と他の手法で T_+ と T_- を分類
- 得られた真陽性の数を t_{pos}，真偽性の数を t_{neg} として，$(t_{\mathrm{pos}} + t_{\mathrm{neg}})/20000$ で accuracy を求める

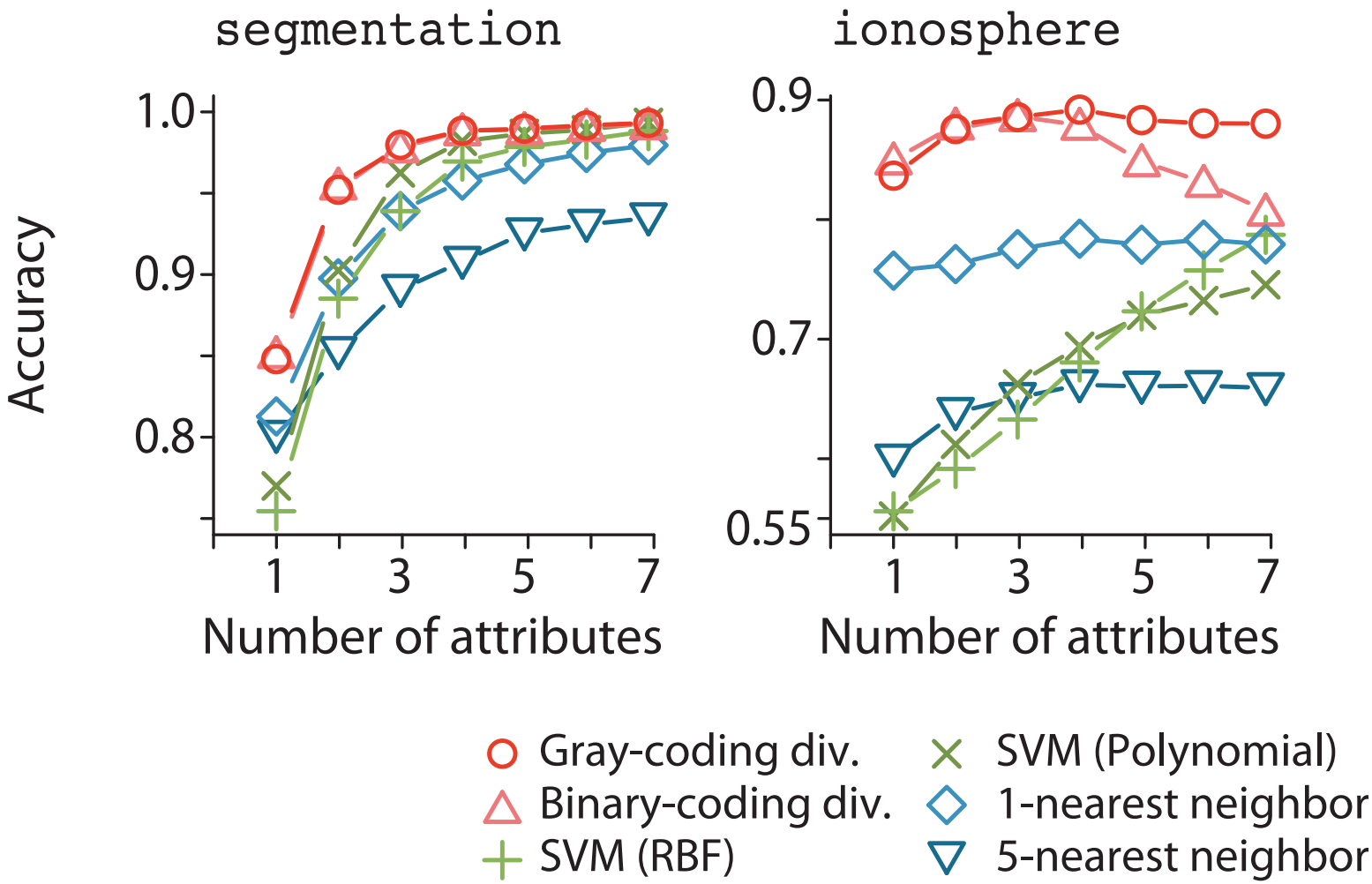
実験結果（各 $n = 10$）


abalon
sonar
Accuracy
0.52
0.46
0.4
0.65
0.55
0.45
1
3
5
7
Number of attributes
Number of attributes
Gray-coding div.
Binary-coding div.
SVM (RBF)
SVM (Polynomial)
1-nearest neighbor
5-nearest neighbor

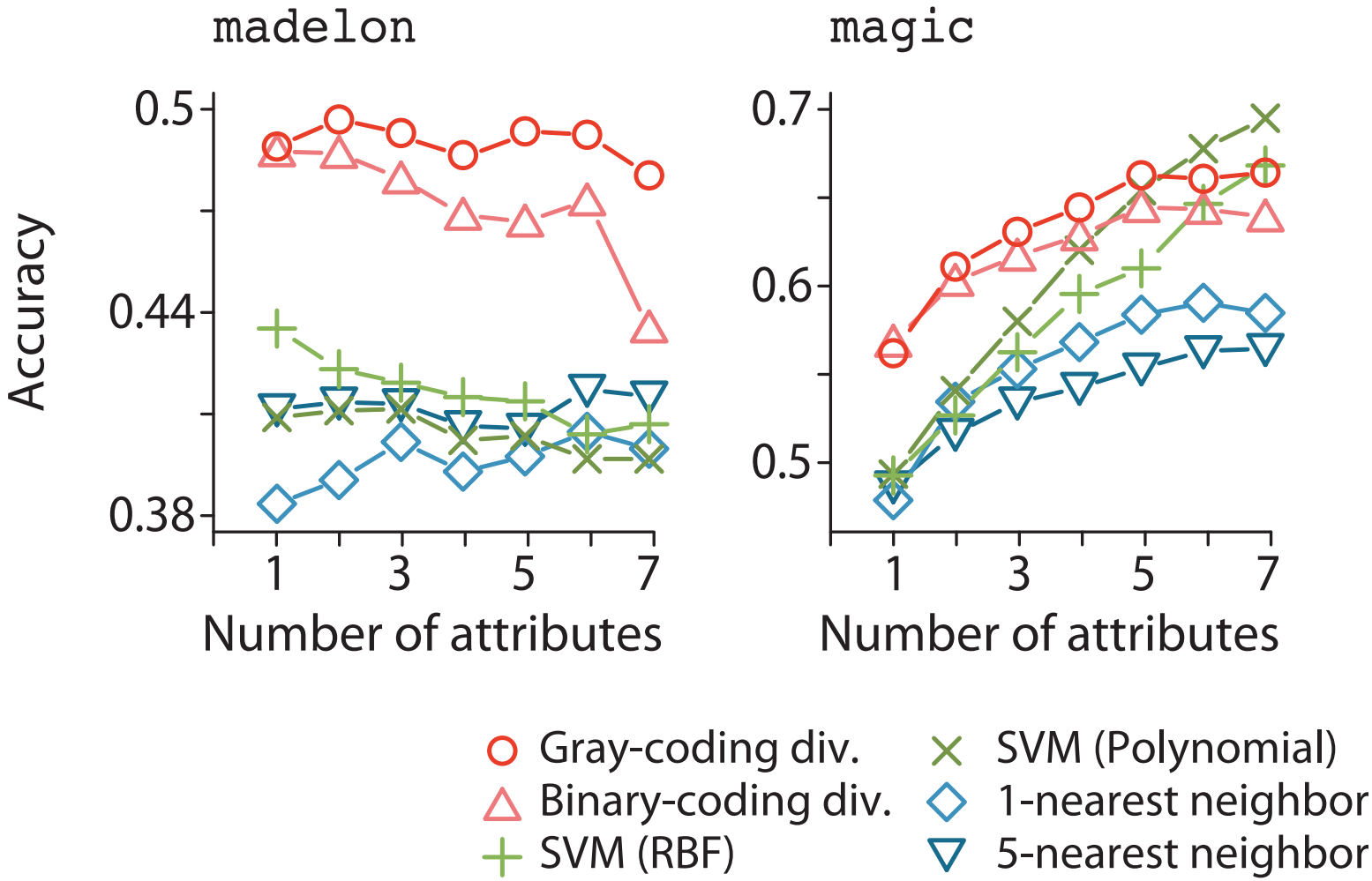
実験結果（各 $n = 10$）


ecoli
glass
Accuracy
1.0
0.75
0.6
0.85
0.7
1
3
5
7
Number of attributes
Gray-coding div.
Binary-coding div.
SVM (RBF)
SVM (Polynomial)
1-nearest neighbor
5-nearest neighbor

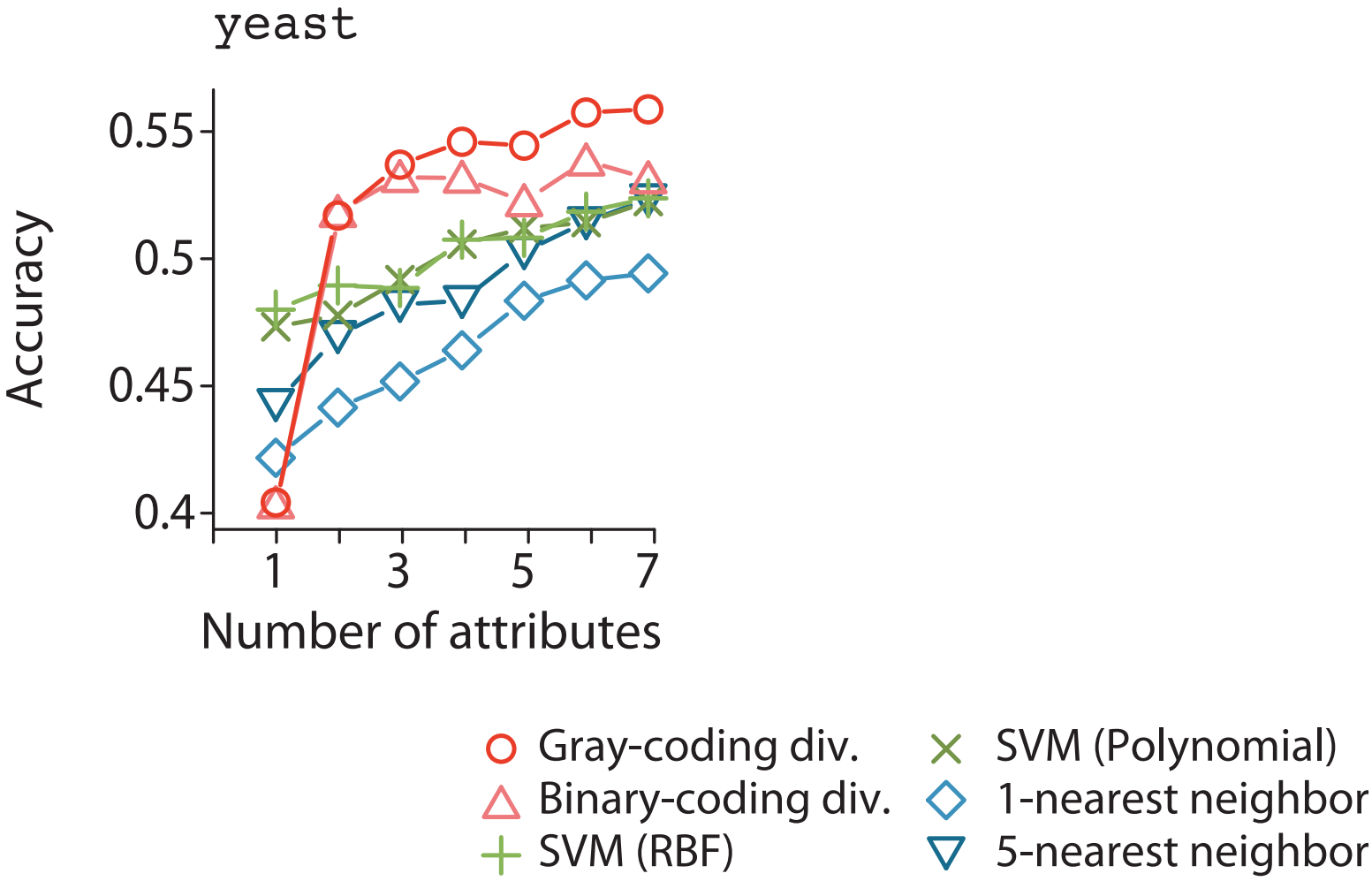
実験結果（各 $n = 10$）


segmentation
ionosphere
Accuracy
1.0
0.9
0.8
0.9
0.7
0.55
1
3
5
7
Number of attributes
Number of attributes
Gray-coding div.
Binary-coding div.
SVM (RBF)
SVM (Polynomial)
1-nearest neighbor
5-nearest neighbor

実験結果（各 $n = 10$）

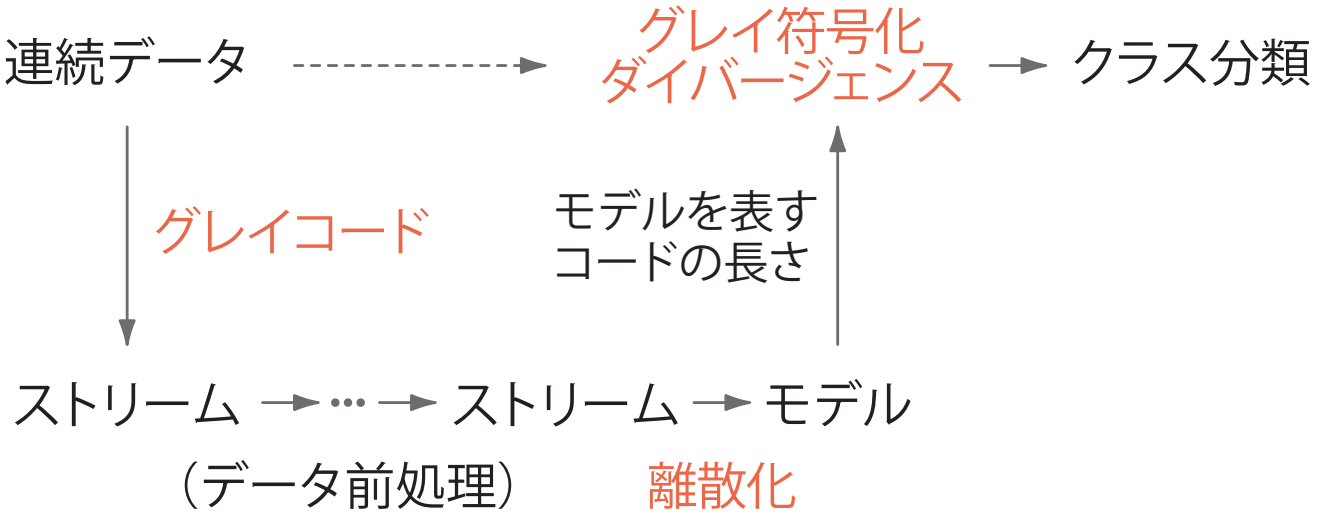

madelon
magic
Accuracy
0.5
0.44
0.38
0.7
0.6
0.5
1
3
5
7
Number of attributes
Number of attributes
Gray-coding div.
Binary-coding div.
SVM (RBF)
SVM (Polynomial)
1-nearest neighbor
5-nearest neighbor

実験結果（各 $n = 10$）


yeast
Accuracy
0.55
0.5
0.45
0.4
1
3
5
7
Number of attributes
Gray-coding div.
Binary-coding div.
SVM (RBF)
SVM (Polynomial)
1-nearest neighbor
5-nearest neighbor

目次

- 研究背景
- グレイコードを用いたストリーム計算
- グレイ符号化ダイバージェンス
- クラス分類
- 実験
- まとめ



まとめ

- 計算可能性解析学（Computable Analysis）と計算論的学習理論（Computational Learning Theory）を理論的背景とし，グレイ符号化ダイバージェンスを導入
  →数値誤差ゼロでの知識発見を可能にする計算論的な枠組
  - 学習手続きを構築
  - 怠惰学習をおこなう分類器を構築
- アイデア：被覆を許した分割によってデータの分離の困難さを測る
- 実データによる実験で，グレイ符号化ダイバージェンスを用いた分類の精度が頑健，かつ優れていることを示した
- 今後の展開：異常値検出など