広報

サイエンス社 ライブラリ データ科学

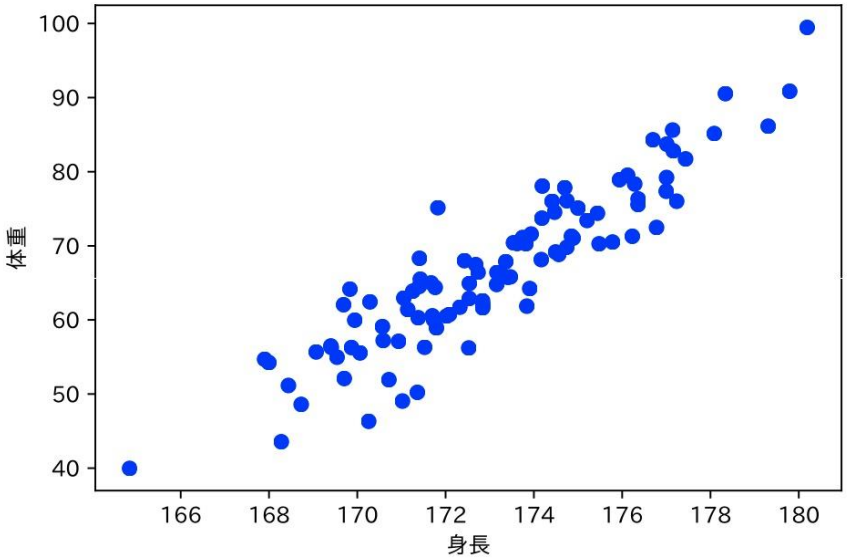
ここまでは，「データ科学入門I」の内容

ここからは，「データ科学入門II」の内容（近々発刊）

© Waseda Education team of Data Science (WEDS)

特徴記述としての回帰問題の例

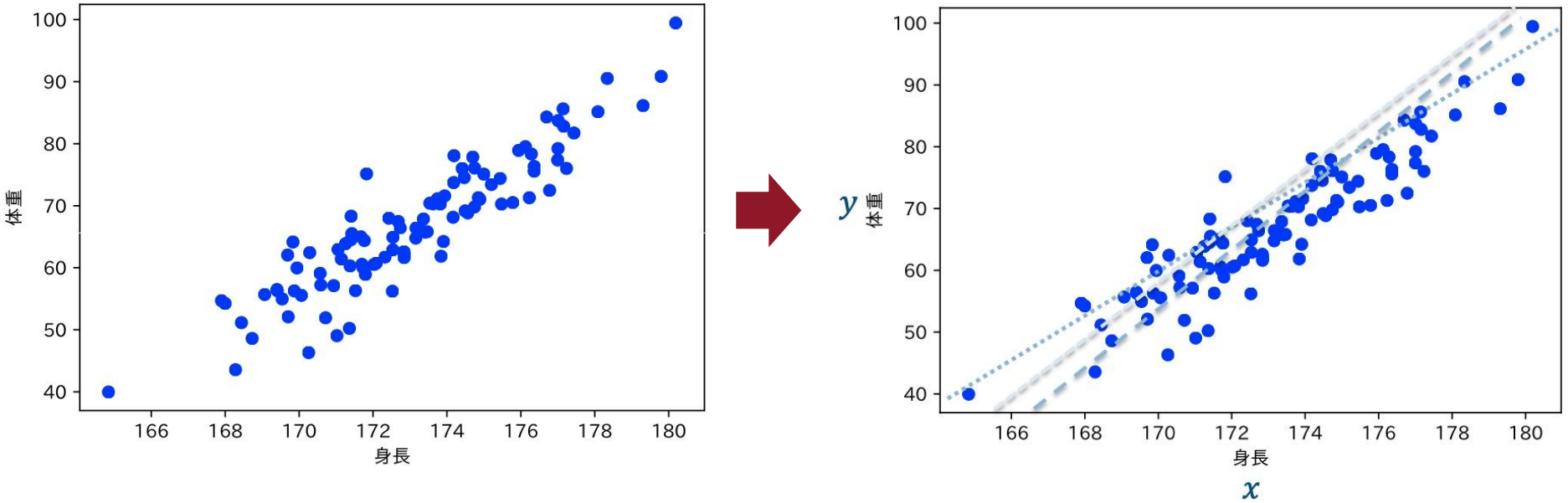
以下のような身長と体重のデータがある．このデータの特徴を記述したい



© Waseda Education team of Data Science (WEDS)

特徴記述としての回帰問題の例

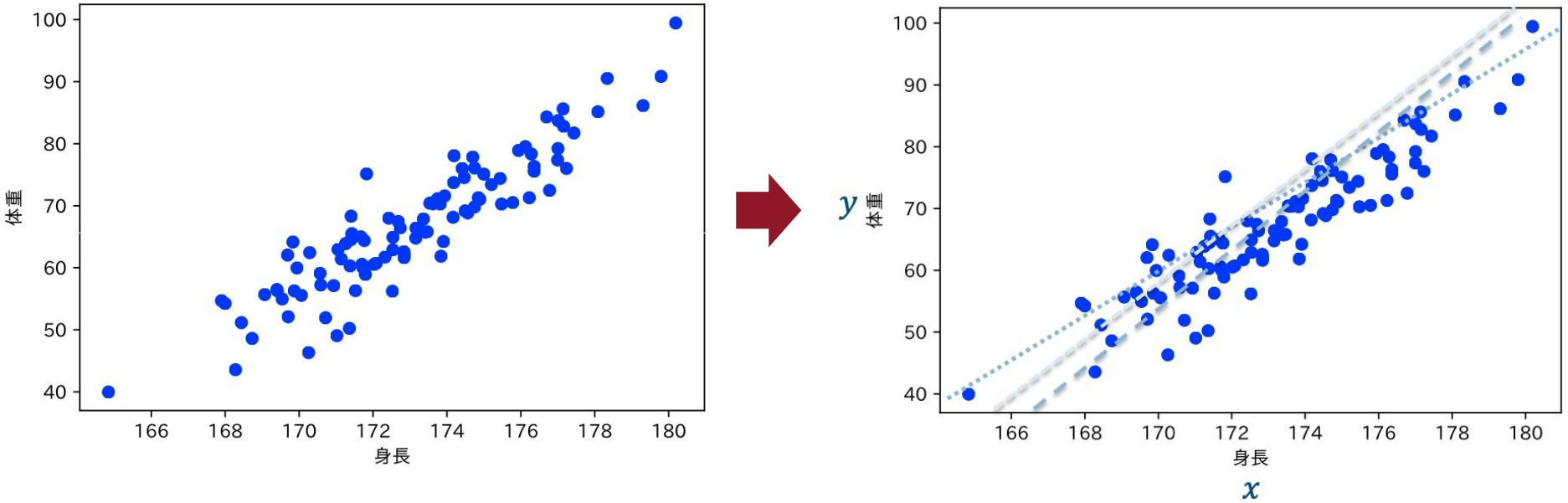
以下のような身長と体重のデータがある．このデータの特徴を記述したい



x と y の関係として，直線の関係で特徴を記述することにする．

特徴記述としての回帰問題の例

以下のような身長と体重のデータがある．このデータの特徴を記述したい



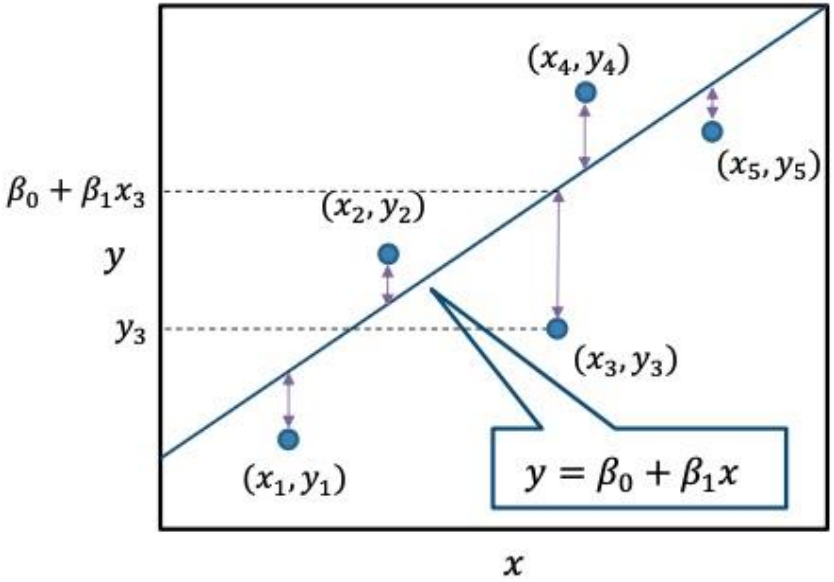
x と y の関係として，直線の関係で特徴を記述することにする．

直線はいくつも考えられるので，評価基準が必要

© Waseda Education team of Data Science (WEDS)

特徴記述としての回帰問題の例

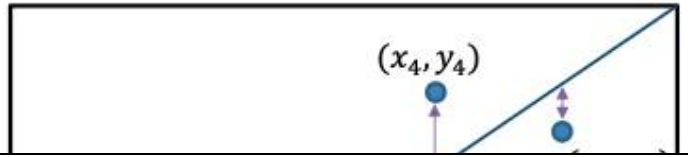
直線はいくつも考えられるので，評価基準が必要



y_i と $\beta_0 + \beta_1 x_i$ との間の「距離の合計」を「評価基準」とする

特徴記述としての回帰問題の例

直線はいくつも考えられるので，評価基準が必要



目的: 目的変数yをxの関数$f(x)$で特徴記述したい
設定: 関数は$f(x) = \beta_0 + \beta_1 x$という直線を表す関数
評価基準: y_iと$\beta_0 + \beta_1 x_i$の間の距離の合計値最小化

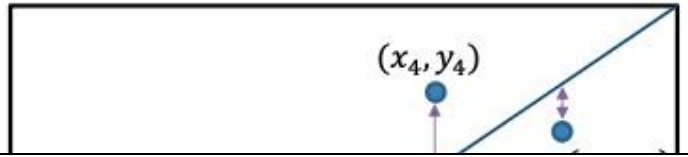
2変数データ
（量的データと量的データ）
$(x_1, y_1), \ldots, (x_n, y_n)$

→ **意思決定写像** →

回帰係数
$\hat{\beta}_0, \hat{\beta}_1$

特徴記述としての回帰問題の例

直線はいくつも考えられるので，評価基準が必要



目的: 目的変数yをxの関数$f(x)$で特徴記述したい
設定: 関数は$f(x) = \beta_0 + \beta_1 x$という直線を表す関数
評価基準: y_iと$\beta_0 + \beta_1 x_i$の間の距離の合計値最小化

2変数データ
（量的データと量的データ）
$(x_1, y_1), \ldots, (x_n, y_n)$

→ **意思決定写像** →

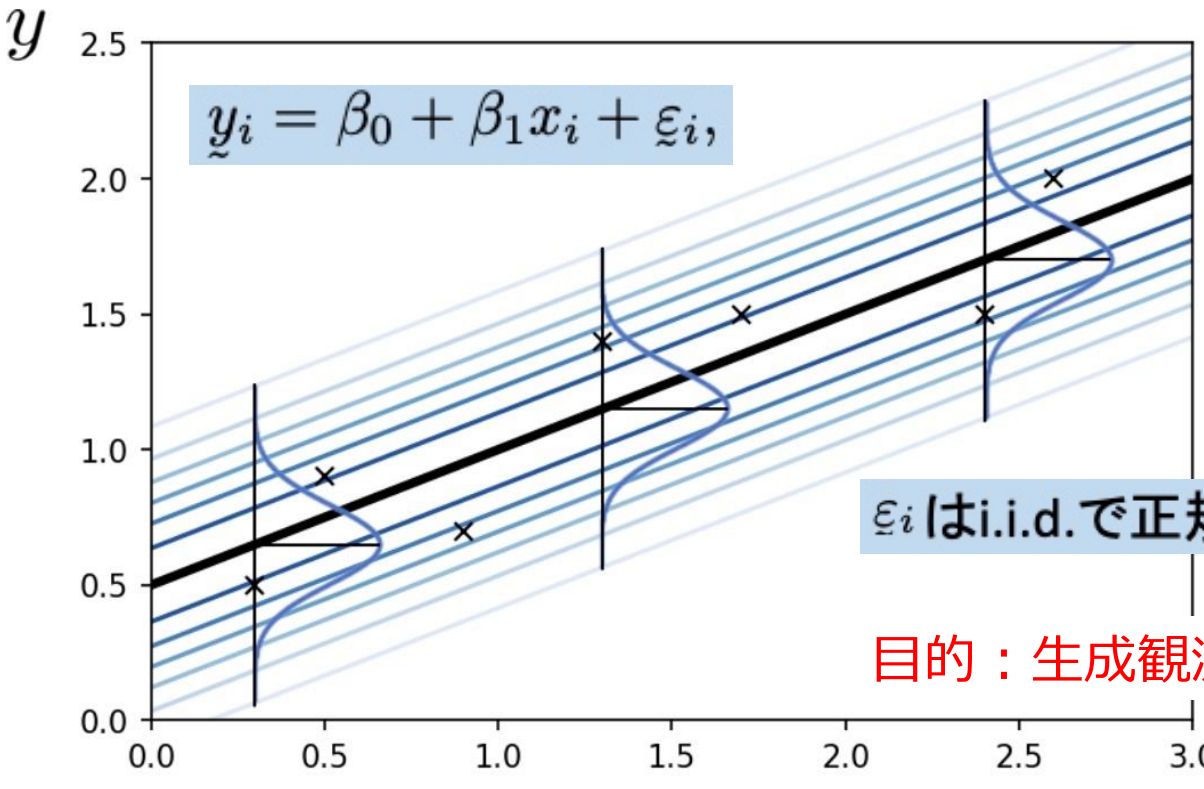
回帰係数
$\hat{\beta}_0, \hat{\beta}_1$

評価基準の「距離」を2乗距離にすると，この意思決定写像は「最小二乗法」

© Waseda Education team of Data Science (WEDS)

構造推定としての回帰問題の例

生成観測メカニズムとしての回帰の**設定**と構造推定（注：データは変わらない）



ε_i はi.i.d.で正規分布 $\mathcal{N}(0, \sigma_\varepsilon^2)$ に従う

目的：生成観測メカニズムの構造を知りたい

© Waseda Education team of Data Science (WEDS)

構造推定としての回帰問題の例

生成観測メカニズムとしての回帰の**設定**と構造推定の例１　（注：データは変わらない）

目的: 説明変数 $x_1, \ldots, x_n$ が与えられたもとでの $\underset{\sim}{y}_1, \ldots, \underset{\sim}{y}_n$ の確率的データ生成観測メカニズムを明らかにしたい

設定: $\underset{\sim}{y}_i = \beta_0 + \beta_1 x_i + \underset{\sim}{\varepsilon}_i, \quad i = 1, \ldots, n$

$\underset{\sim}{\varepsilon}_i$ はi.i.d.で正規分布 $\mathcal{N}(0, \sigma_\varepsilon^2)$ に従う

評価基準: 尤度最大化

2変数データ
（量的データと量的データ）
$(x_1, \underset{\sim}{y}_1), \ldots, (x_n, \underset{\sim}{y}_n)$

→ **意思決定写像** →

母回帰係数と誤差項の分散の推定量
$\hat{\underset{\sim}{\beta}}_0, \hat{\underset{\sim}{\beta}}_1, \widehat{\underset{\sim}{\sigma}_\varepsilon^2}$

© Waseda Education team of Data Science (WEDS)

構造推定としての回帰問題の例

生成観測メカニズムとしての回帰の**設定**と構造推定の例２ （注：データは変わらない）

目的: 説明変数$x_1, \ldots, x_n$が与えられたもとでの$\underset{\sim}{y}_1, \ldots, \underset{\sim}{y}_n$の確率的データ生成観測メカニズムを明らかにしたい

設定: $\underset{\sim}{y}_i = \beta_0 + \beta_1 x_i + \underset{\sim}{\varepsilon}_i, \quad i = 1, \ldots, n$

$\underset{\sim}{\varepsilon}_i$はi.i.d.で正規分布$\mathcal{N}(0, \sigma_\varepsilon^2)$に従う

評価基準: $[\underset{\sim}{l}_0, \underset{\sim}{u}_0]$と$[\underset{\sim}{l}_1, \underset{\sim}{u}_1]$がそれぞれ真の$\beta_0$と$\beta_1$を$1-\alpha$の確率で含むという条件のもとで区間幅が最小の区間

2変数データ
（量的データと量的データ）

$(x_1, \underset{\sim}{y}_1), \ldots, (x_n, \underset{\sim}{y}_n)$

→ **意思決定写像** →

母回帰係数の信頼区間

$[\underset{\sim}{l}_0, \underset{\sim}{u}_0]$ $[\underset{\sim}{l}_1, \underset{\sim}{u}_1]$

© Waseda Education team of Data Science (WEDS)

構造推定としての回帰問題の例

生成観測メカニズムとしての回帰の**設定**と構造推定の例３ （注：データは変わらない）
回帰係数も確率変数と**設定**した場合の意思決定写像の例

目的:	説明変数 $x_1, \ldots, x_n$ が与えられたもとでの $\underset{\sim}{y}_1, \ldots, \underset{\sim}{y}_n$ の確率的データ生成観測メカニズムを明らかにしたい
設定:	$\underset{\sim}{y}_i = \underset{\sim}{\beta}_0 + \underset{\sim}{\beta}_1 x_i + \underset{\sim}{\varepsilon}_i, \quad i = 1, \ldots, n$
	$\underset{\sim}{\varepsilon}_i$ はi.i.d.で正規分布 $\mathcal{N}(0, \sigma_\varepsilon^2)$ に従う
	$\underset{\sim}{\beta}_0, \underset{\sim}{\beta}_1$ は事前分布 $p(\beta_0, \beta_1)$ に従う
	損失関数: $\ell(\beta_0, \beta_1, d((x_1, y_1), \ldots, (x_n, y_n)))$
評価基準:	ベイズ危険関数最小化

2変数データ
（量的データと量的データ）
$(x_1, \underset{\sim}{y}_1), \ldots, (x_n, \underset{\sim}{y}_n)$

→ **意思決定写像** →

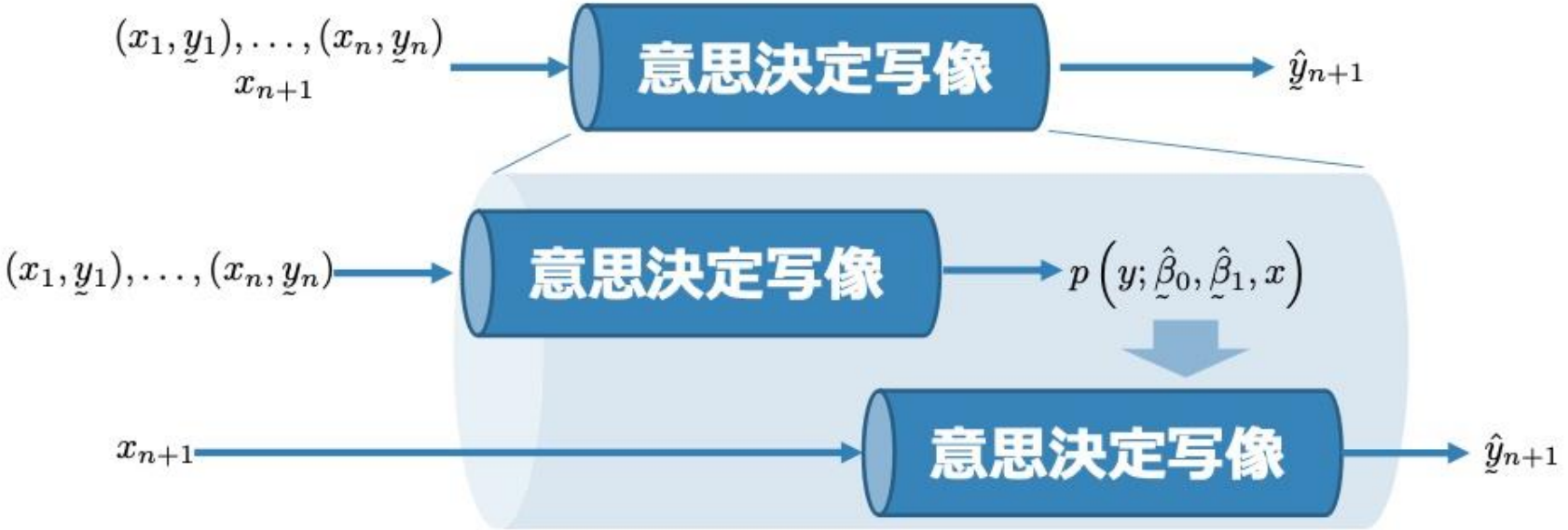
母回帰係数の推定量
$\underset{\sim}{\hat{\beta}}_0, \underset{\sim}{\hat{\beta}}_1$

© Waseda Education team of Data Science (WEDS)

予測としての回帰問題の例

生成観測メカニズムとして回帰を**設定**し，新しい説明変数に対する未知の目的変数の値を推定したい（注：データは変わらない）

[間接予測（新語？）] 一度構造推定を行い，そのパラメータを使った意思決定写像で予測



© Waseda Education team of Data Science (WEDS)

予測としての回帰問題の例

生成観測メカニズムとして回帰を**設定**し，新しい説明変数に対する未知の目的変数の値を推定したい（注：データは変わらない）

[直接予測（新語？）] パラメータを一つに決めることをせず，予測値そのものを推定する考え方

$\underset{\sim}{y}_{n+1}$ と $\hat{\underset{\sim}{y}}_{n+1}$ の間の損失を考える

目的: x_{n+1}に対応した $\underset{\sim}{y}_{n+1}$を予測したい

設定: $\underset{\sim}{y}_i = \underset{\sim}{\beta}_0 + \underset{\sim}{\beta}_1 x_i + \underset{\sim}{\varepsilon}_i, \quad i = 1, \ldots, n+1$

$\underset{\sim}{\varepsilon}_i$はi.i.d.で正規分布$\mathcal{N}(0, \sigma_\varepsilon^2)$に従う

$\underset{\sim}{\beta}_0, \underset{\sim}{\beta}_1$ は事前分布 $p(\beta_0, \beta_1)$ に従う

$\underset{\sim}{y}_{n+1}$と予測値の間の近さの基準

評価基準: ベイズ危険関数最小化

2変数データ
（量的データと量的データ）

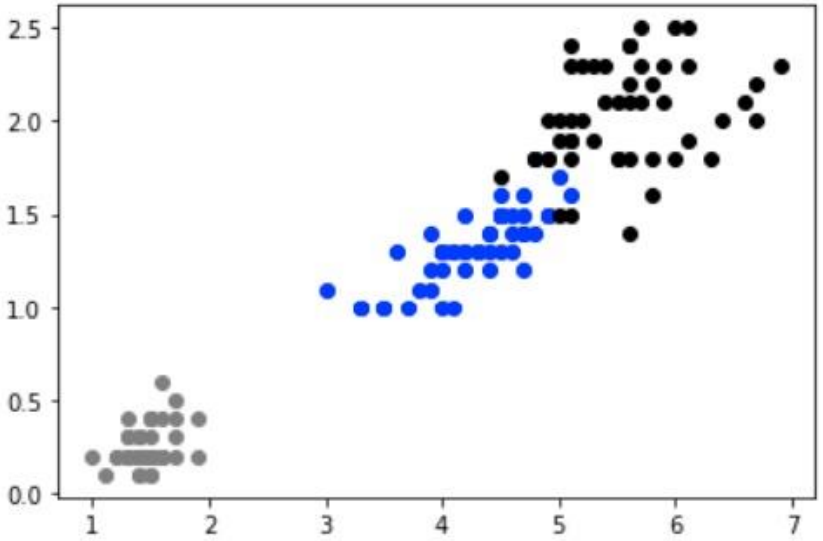
$(x_1, \underset{\sim}{y}_1), \ldots, (x_n, \underset{\sim}{y}_n)$
x_{n+1}

→ **意思決定写像** →

$\underset{\sim}{y}_{n+1}$**の予測値**
$\hat{\underset{\sim}{y}}_{n+1}$

特徴記述としての分類問題の例

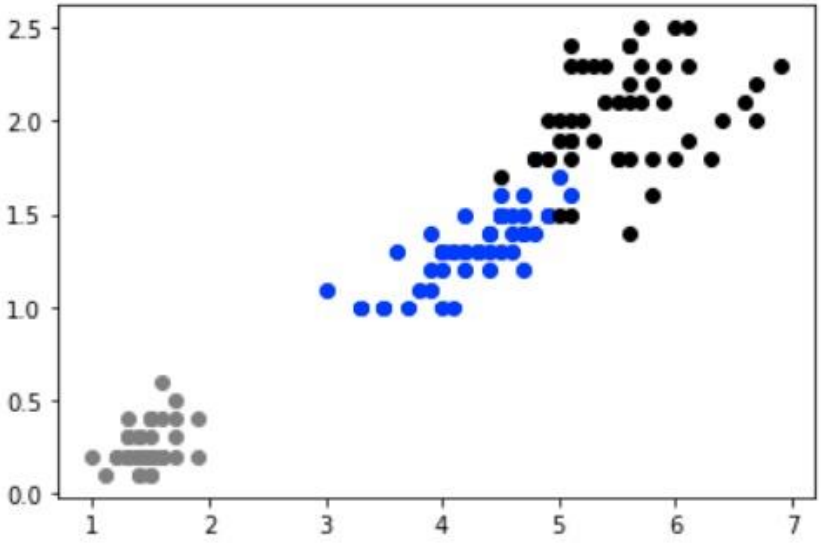
以下のようなデータがある．このデータの特徴を記述したい

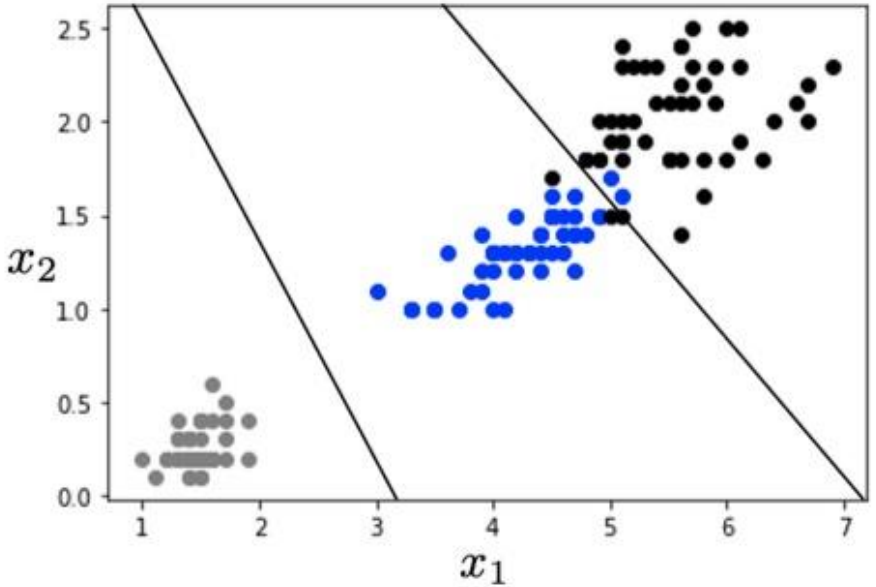


© Waseda Education team of Data Science (WEDS)

特徴記述としての分類問題の例

以下のようなデータがある．このデータの特徴を記述したい



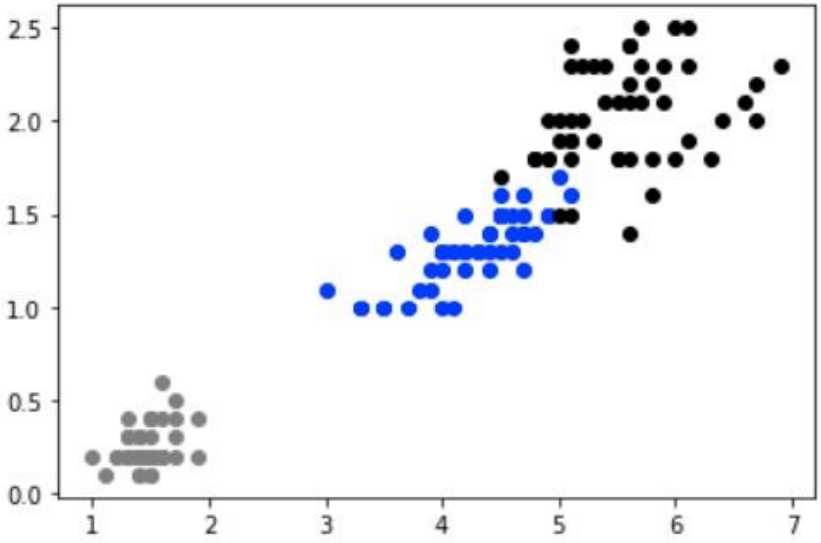


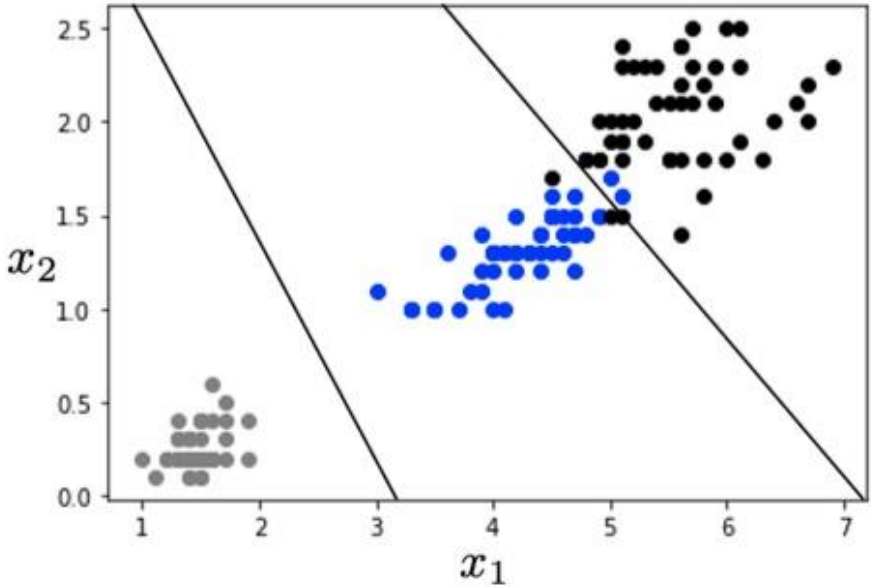
x_1 , x_2 と y（y は質的変数）の関係として，それぞれを別つ平面（ここでは直線）の関係で特徴を記述することにする．

© Waseda Education team of Data Science (WEDS)

特徴記述としての分類問題の例

以下のようなデータがある．このデータの特徴を記述したい





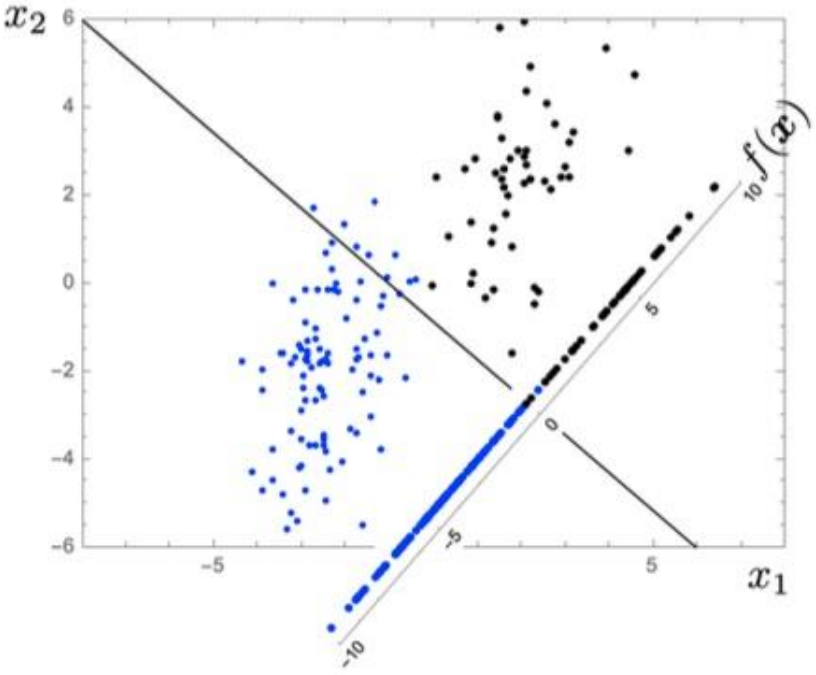
x_1 , x_2 と y（y は質的変数）の関係として，それぞれを別つ平面（ここでは直線）の関係で特徴を記述することにする．

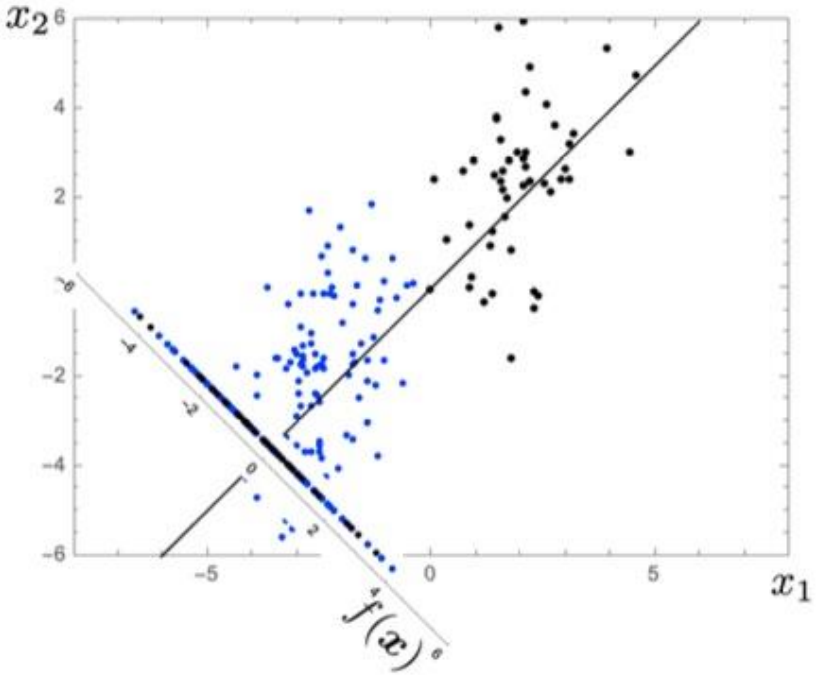
直線はいくつも考えられるので，評価基準が必要

© Waseda Education team of Data Science (WEDS)

特徴記述としての分類問題の例

評価基準を考えるために，別つ平面（ここでは直線）と垂直な直線への射影を考える



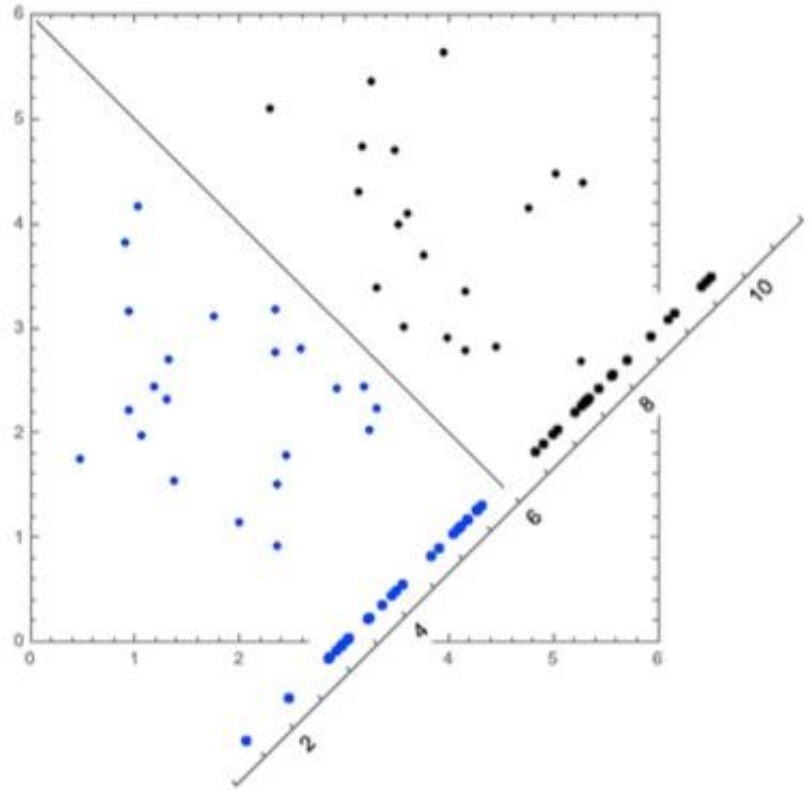


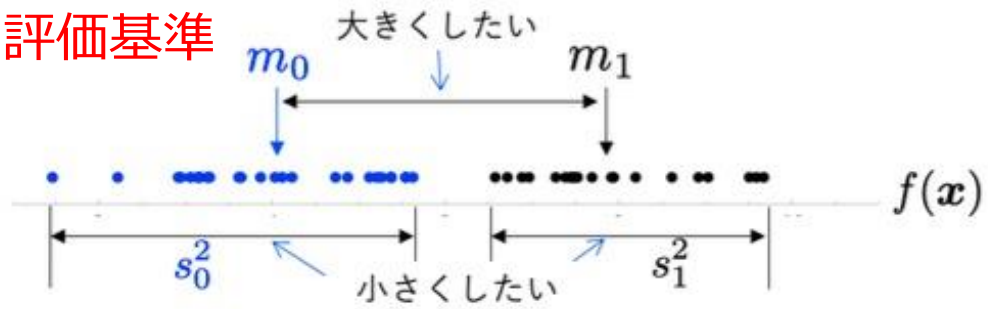
直線はいくつも考えられるので，評価基準が必要

© Waseda Education team of Data Science (WEDS)

特徴記述としての分類問題の例

評価基準を考えるために，別つ平面（ここでは直線）と垂直な直線への射影を考える

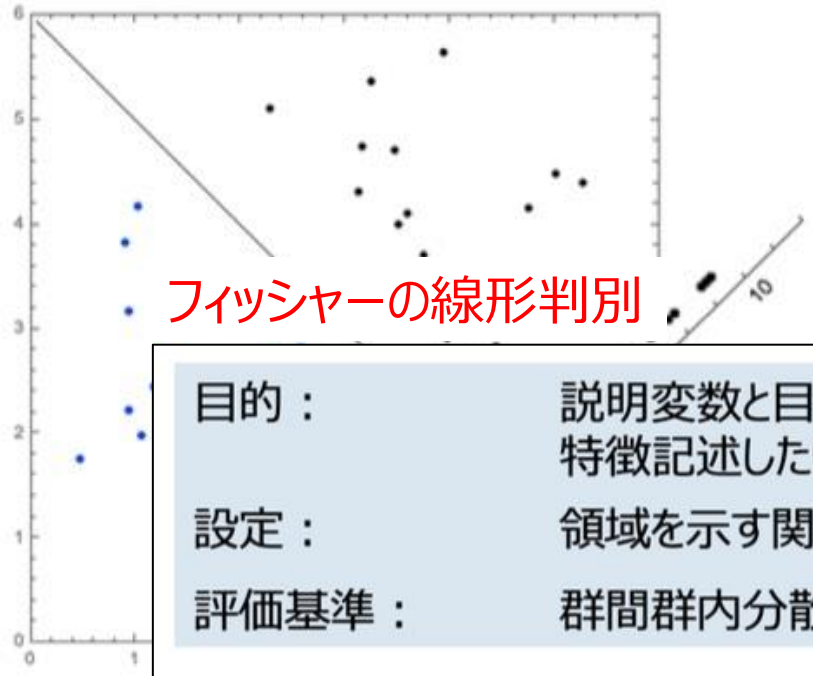




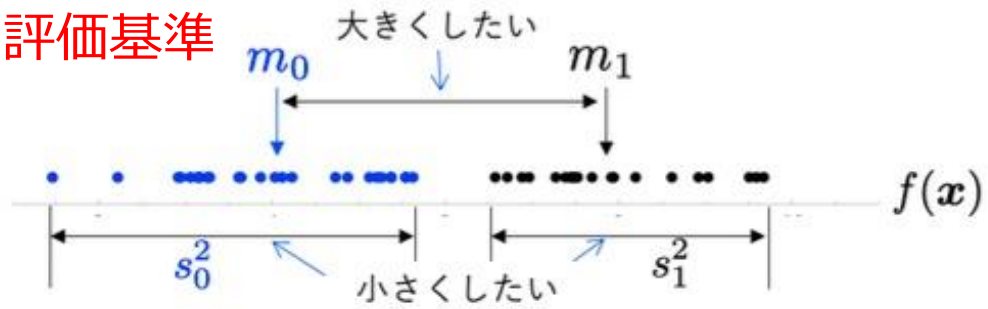
© Waseda Education team of Data Science (WEDS)

特徴記述としての分類問題の例

評価基準を考えるために，別つ平面（ここでは直線）と垂直な直線への射影を考える



フィッシャーの線形判別



目的：	説明変数と目的変数（質的）の関係を，領域を用いて特徴記述したい
設定：	領域を示す関数には線形関数$f(\boldsymbol{x}) = \boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{x}$ である
評価基準：	群間群内分散比最大化

$(\boldsymbol{x}_i = [1, x_{i1}, x_{i2}, \ldots, x_{ip}]^{\mathsf{T}}, y_i)$

$i = 1,2, \ldots, n+m$

→ 意思決定写像 →

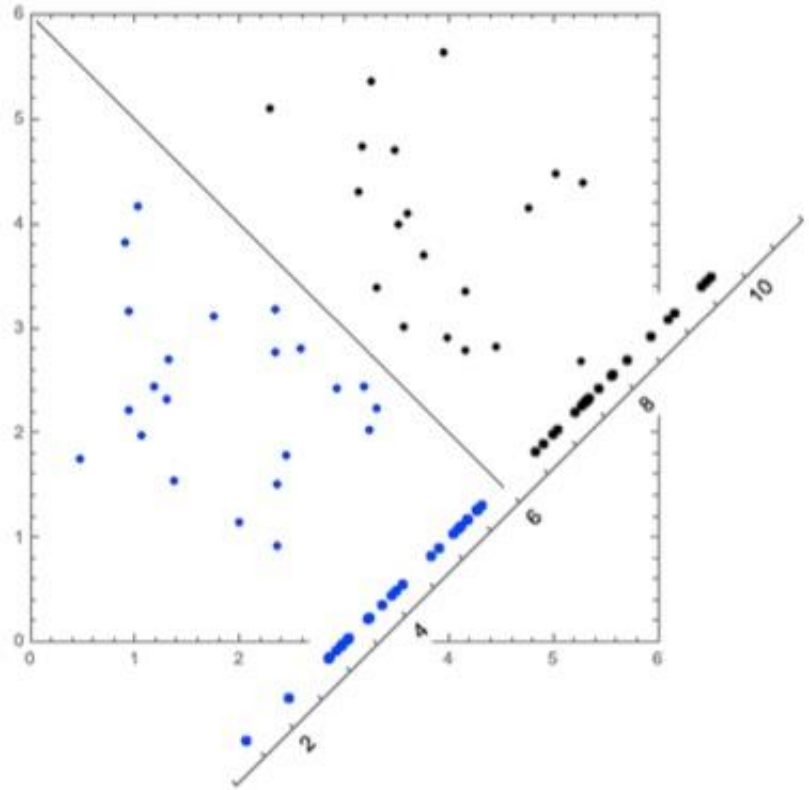
$f(\boldsymbol{x})$の係数

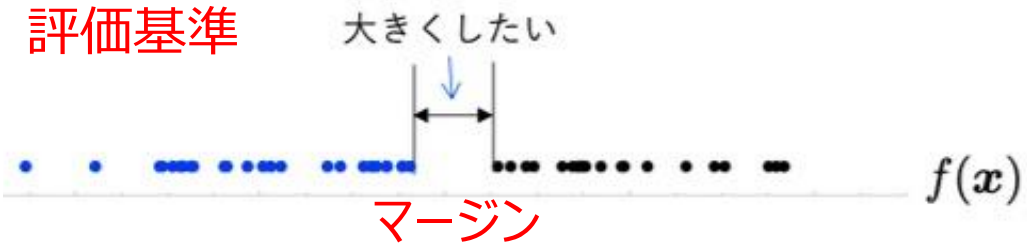
$[\beta_0, \beta_1, \ldots, \beta_p]^{\mathsf{T}}$

© Waseda

特徴記述としての分類問題の例

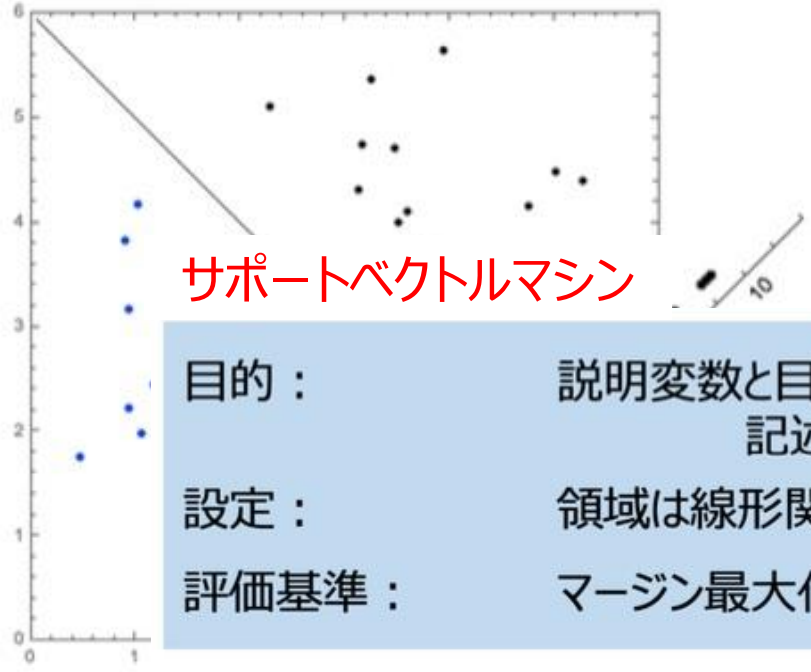
評価基準を考えるために，別つ平面（ここでは直線）と垂直な直線への射影を考える



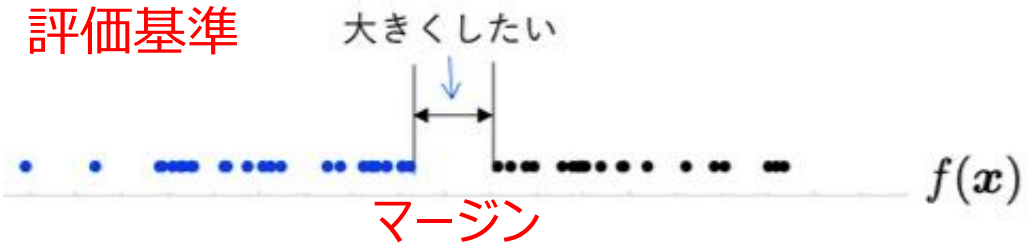


特徴記述としての分類問題の例

評価基準を考えるために，別つ平面（ここでは直線）と垂直な直線への射影を考える



サポートベクトルマシン



目的：	説明変数と目的変数（質的）の関係を，領域を用いて特徴記述したい
設定：	領域は線形関数$f(\boldsymbol{x}) = \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}$を用いて分離可能である
評価基準：	マージン最大化

$(\boldsymbol{x_i} = [1, x_{i1}, x_{i2}, \dots, x_{ip}]^{\mathrm{T}}, y_i)$

$i = 1,2, \dots, n+m$

→ **意思決定写像** →

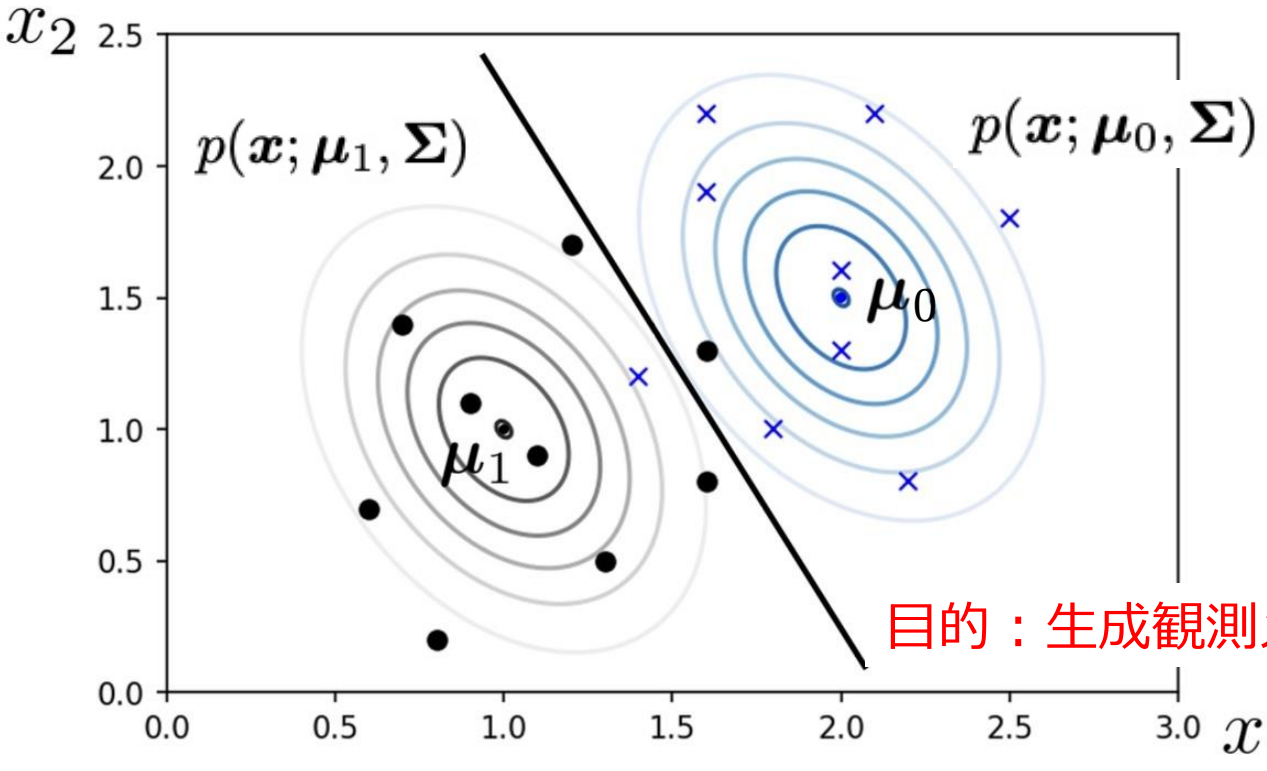
$f(\boldsymbol{x})$の係数

$[\beta_0, \beta_1, \dots, \beta_p]^{\mathrm{T}}$

© Wased

構造推定としての分類問題の例

生成観測メカニズムとしての分類の設定と構造推定（注：データは変わらない）



目的：生成観測メカニズムの構造を知りたい

構造推定としての分類問題の例

生成観測メカニズムとしての分類の設定と構造推定（注：データは変わらない）

生成観測メカニズムの構造推定としてのフィッシャーの線形判別の意思決定写像

目的：	確率的データ生成観測メカニズムを明らかにしたい（構造推定）
設定：	各群のもとでデータは多変量正規分布に従う
	$p(\boldsymbol{x}\|0): \mathcal{N}(\mu_0, \boldsymbol{\Sigma}),\ p(\boldsymbol{x}\|1): \mathcal{N}(\mu_1, \boldsymbol{\Sigma})$
評価基準：	尤度最大

$(\underset{\sim}{x}_i = [x_{i1}, \ldots, x_{ip}]^\top, \underset{\sim}{y}_i)$

$i = 1,2, \ldots, n+m$

→ 意思決定写像 →

パラメータ $\mu_0, \mu_1, \boldsymbol{\Sigma}$ の最尤推定量

© Waseda Education team of Data Science (WEDS)

構造推定としての分類問題の例

生成観測メカニズムとしての分類の設定と構造推定（注：データは変わらない）

ロジスティック回帰の意思決定写像

目的：	確率的データ生成観測メカニズムを明らかにしたい
設定：	群y_iの発生確率はロジスティック回帰モデルに従う
評価基準：	尤度最大

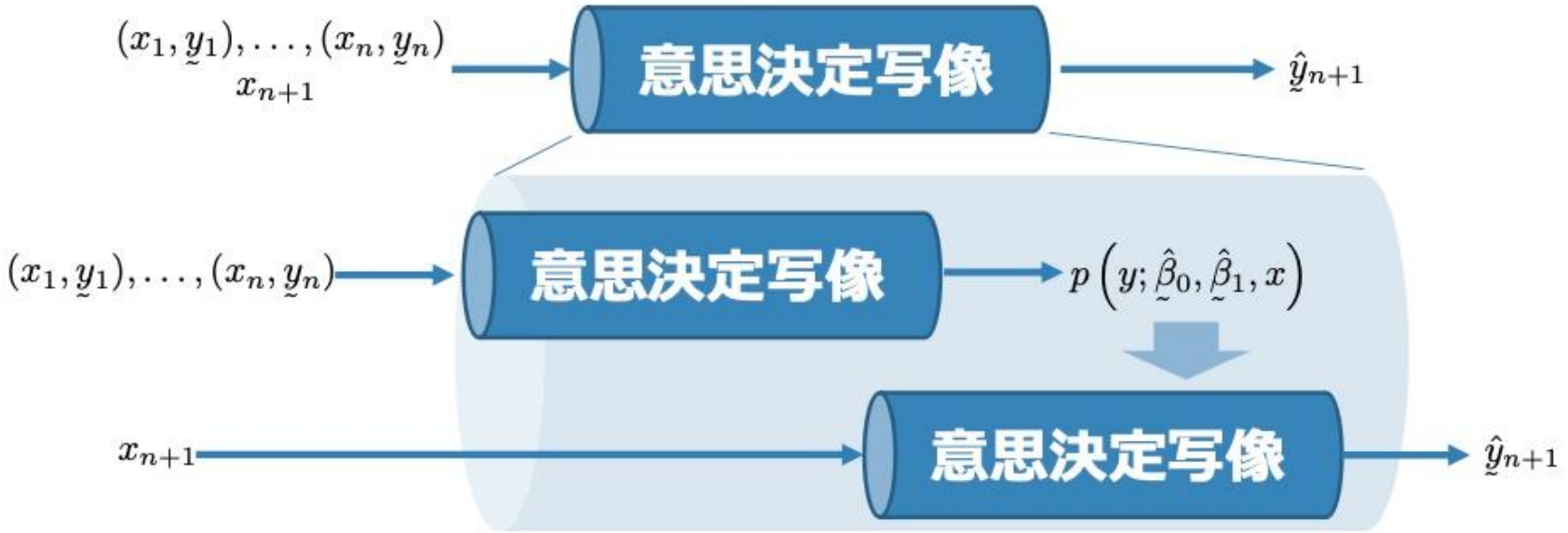
$(\underset{\sim}{x}_i, \underset{\sim}{y}_i)$
$i = 1,2,\ldots,n+m$

© Waseda Education team of Data Science (WEDS)

予測としての分類問題の例

生成観測メカニズムとしての分類の設定と予測（注：データは変わらない）

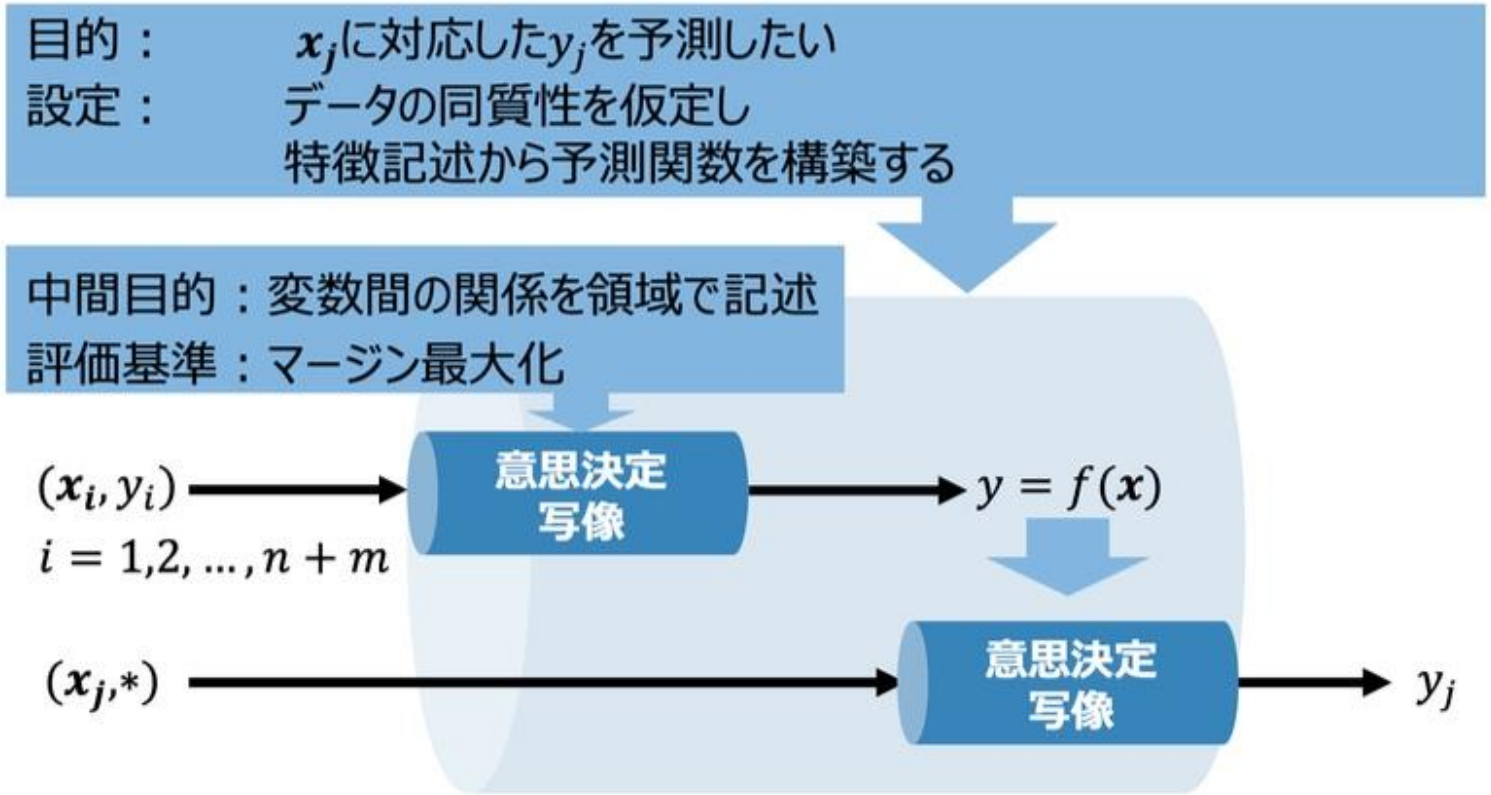
[間接予測（新語？）] 一度構造推定を行い，そのパラメータを使った意思決定写像で予測



© Waseda Education team of Data Science (WEDS)

予測としての分類問題の例

[データの同質性（新語）] 学習データと予測の対象となる新規データは同じメカニズムで発生する



意思決定写像の標準化図

目的：
設定：
評価基準：

- 分析者が分析の目的ややりたいことに合わせて，これらのパーツを選択できるようになることが到達目標

© Waseda Education team of Data Science (WEDS)

データ科学の統一的体系化の取り組みの背景

- 2017年12月：早稲田大学データ科学センター発足
- 2018年4月　：実質的始動（カリキュラム検討開始）

早稲田大学全学でのデータ科学関連科目のシラバス調査
（どんなことを教えているか？必要としているか？どこまで教えているか？…）

➡ 統計学や機械学習，データマイニング，パターン認識，AIなど様々な文脈で山ほどの科目があるが，統一的な科目群のパッケージは見当たらない
➡ データ科学センターで作る必要があると判断

「データ科学センター教員全員」と「GECデータ科学教育部門の一部教員」が「金曜日丸一日使って」体系化の検討
（この検討のための事前準備は別途実施…）

今後の「サイエンス社 ライブラリ データ科学」シリーズ発刊予定

2. データ科学入門II
3. データ科学入門III
4. データ科学入門IV
5. データ科学実践
6. 回帰と分類のデータ科学
7. 時系列構造のデータ科学
8. 潜在構造のデータ科学
9. 空間構造のデータ科学
10. 因果構造のデータ科学
11. データ科学のためのモデリング

© Waseda Education team of Data Science (WEDS)