知能システム開発特論:
第2回

データマイニングの基礎2:
データセット

ダビド (david@iwate-pu.ac.jp)
IPU
15 December 2016

データ(Data)

- すべてのデータセットはデータ行列ではなく、順列、テキスト、時系列、画像、音声、ビデオ、などもある。
- 特徴抽出による上記のデータはデータ行列に変換できる。
- データ分析はエンティティごとに独立ということを想定している。実世界にはそれぞれのエンティティはさまざまな関係もある。データグラフのモデルも考えられる。

属性の種類:量的変数 vs 質的変数

- 量的変数(quantitative or numeric):

例:重さ、温度、物の数

- 質的変数(qualitative or categorical):演算不可能

例:人間の目の色、郵便番号、IPアドレス、

数字の桁{0,1,..,9}、

Iris花のクラス{Setosa,Versicolor,Virginica}

目標変数(target variable):二つのカテゴリの表現が多い。

アイリスデータセットのサンプル

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
12	4.8	3.4	1.6	0.2	setosa
25	4.8	3.4	1.9	0.2	setosa
7	4.6	3.4	1.4	0.3	setosa
125	6.7	3.3	5.7	2.1	virginica
145	6.7	3.3	5.7	2.5	virginica
101	6.3	3.3	6.0	2.5	virginica
57	6.3	3.3	4.7	1.6	versicolor
24	5.1	3.3	1.7	0.5	setosa
50	5.0	3.3	1.4	0.2	setosa
126	7.2	3.2	6.0	1.8	virginica
51	7.0	3.2	4.7	1.4	versicolor
121	6.9	3.2	5.7	2.3	virginica

データの代数的な視点

- データ行列(Data Matrix)

行：レコード、サンプル、エンティティ、オブジェクト、特徴ベクター、

5-タプル

$$x_1 = (5.9, 3.0, 4.2, 1.5, versicolor)$$

列：属性、プロパティ、特徴、変数、フィールド

データ行列(DATA MATRIX)

$n \times d$ データ行列 (data matrix), n 行(rows) and d 列(columns)

行 (rows) = データセットのエンティティ(entities in the dataset)

列 (columns) = 属性、プロパーティ(attibutes or properties)

$$D = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix}$$

x_i : i-th row d-tuple : $x_i = (x_{i1}, x_{i2}, \cdots, x_{id})$

X_j : j-th column n-tuple : $X_j = (x_{1j}, x_{2j}, \ldots, x_{nj})$

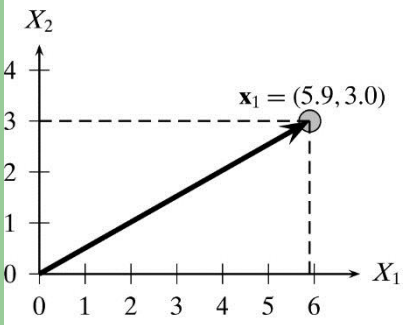
row=={instance, example, record, transaction, object, point, feature-vector, tuple}

column=={attribute, property, feature, dimension, variable, field}

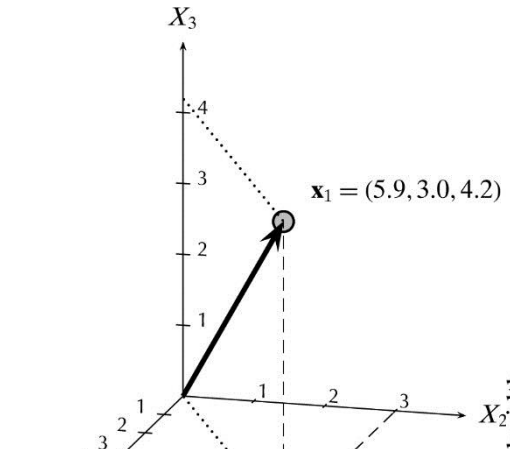
n : データのサイズ(size of the data)

d : データの次元(dimensionality of the data)

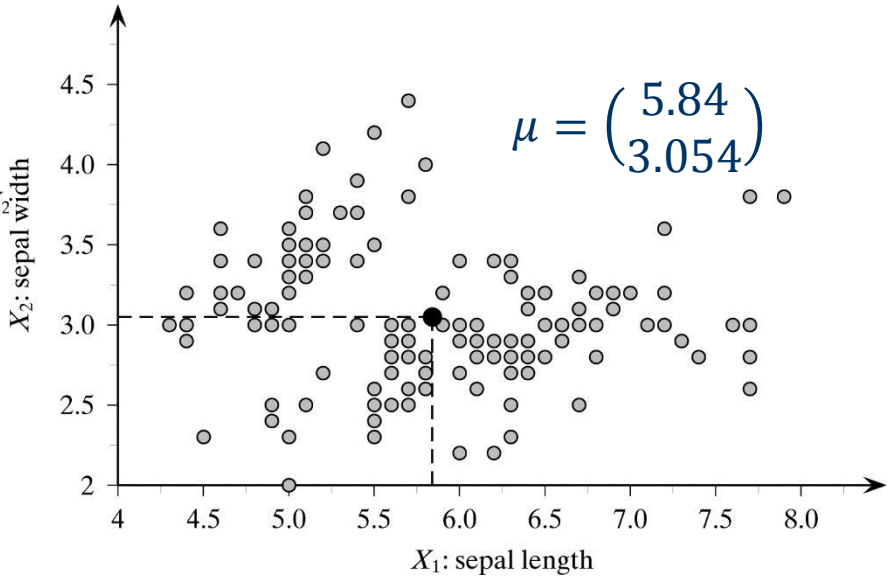
幾何学の観点：3次元、2次元、1次元の投影



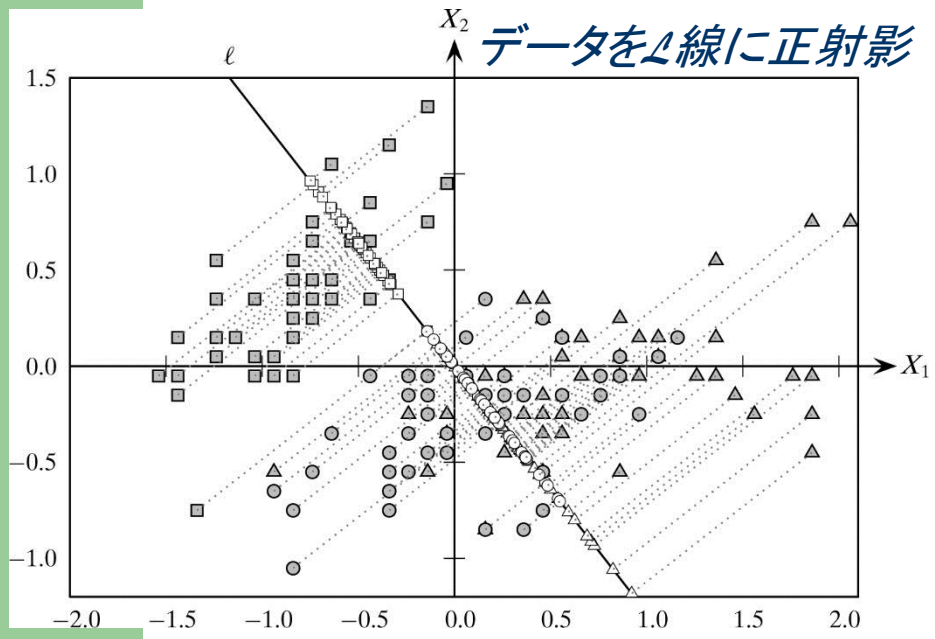
(a)



(b)



データをℓ線に正射影



離散属性　vs 連続属性

離散属性（Numeric attribute）

属性の領域：domain(Age) = $\mathbb{N}$, domain(petal length)= $\mathbb{R}^+$

種類：離散、連続、バイナリ

カテゴリ型属性（Categorical attribute）

属性の領域：domain(血液型) ={A,B,AB,O}, domain(性別)={M,F}

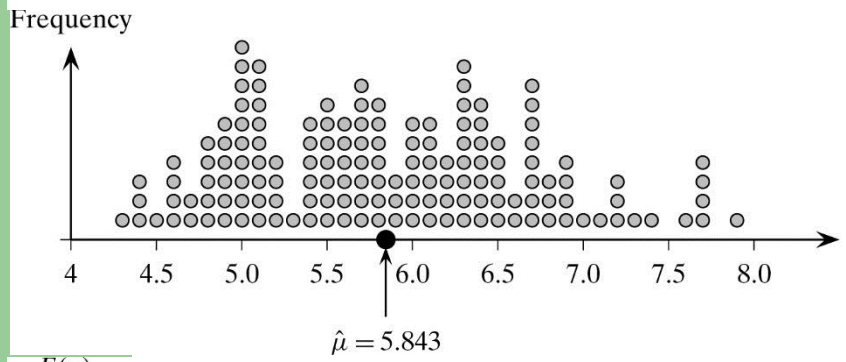
アイリスデータセットの統計結果

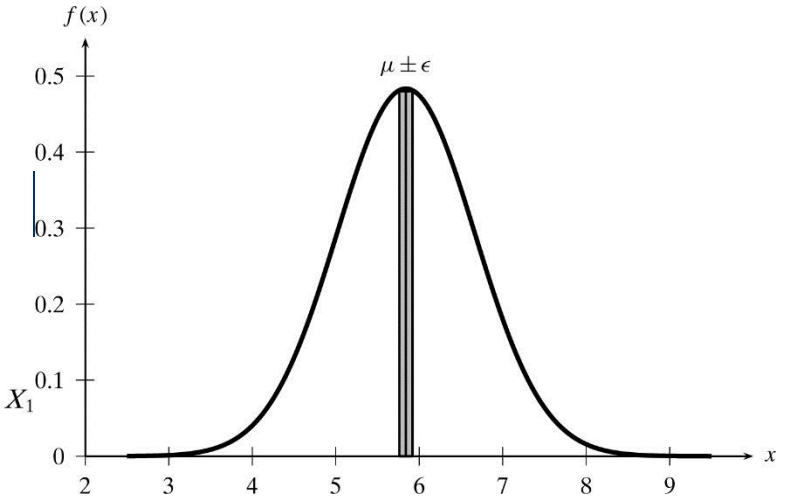
```
> summary(iris)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
 Median :5.800   Median :3.000   Median :4.350   Median :1.300
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
       Species
 setosa    :50
 versicolor:50
 virginica :50
```

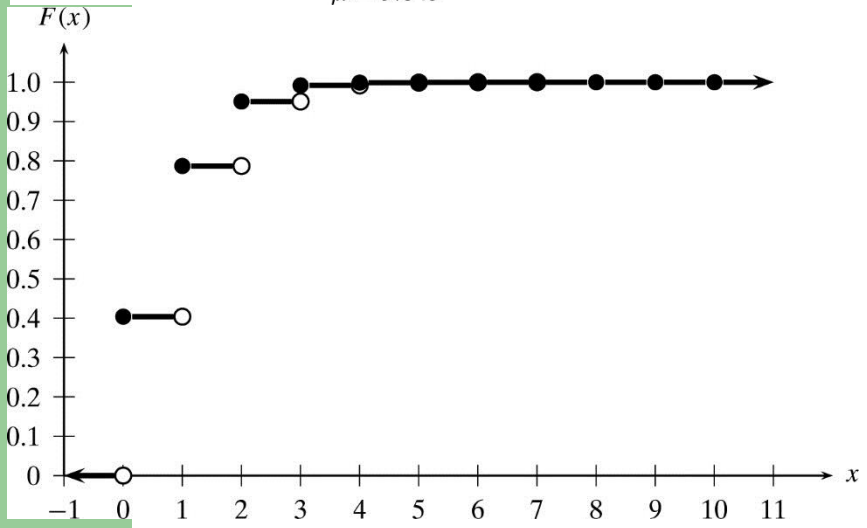
サンプリング Univariate

f, F, μ, σ, r: 母集団の確率分布,累積分布関数,平均,標準偏差,レンジ

$\hat{f}, \hat{F}, \hat{\mu}, \hat{\sigma}, \hat{r}$:サンプルの内容



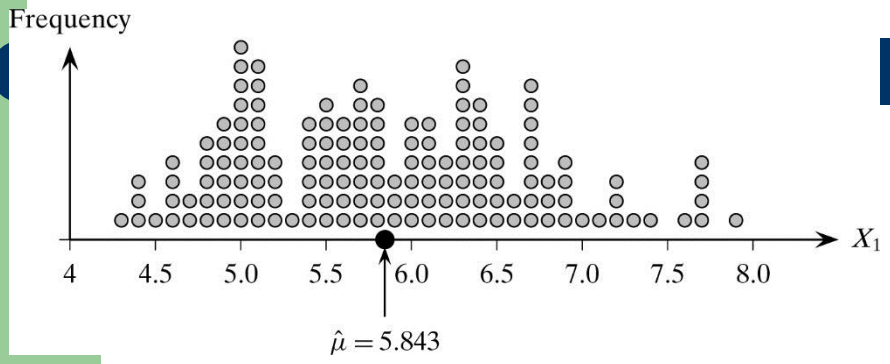


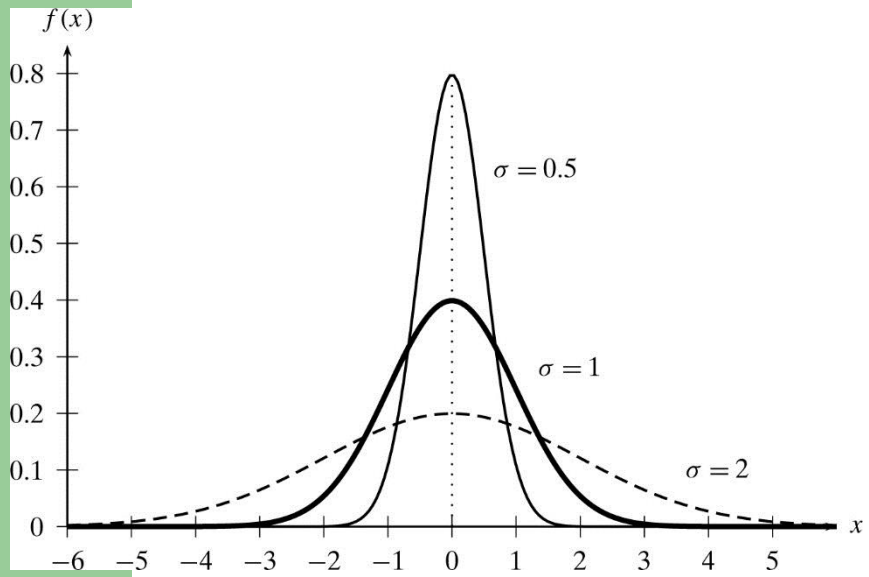


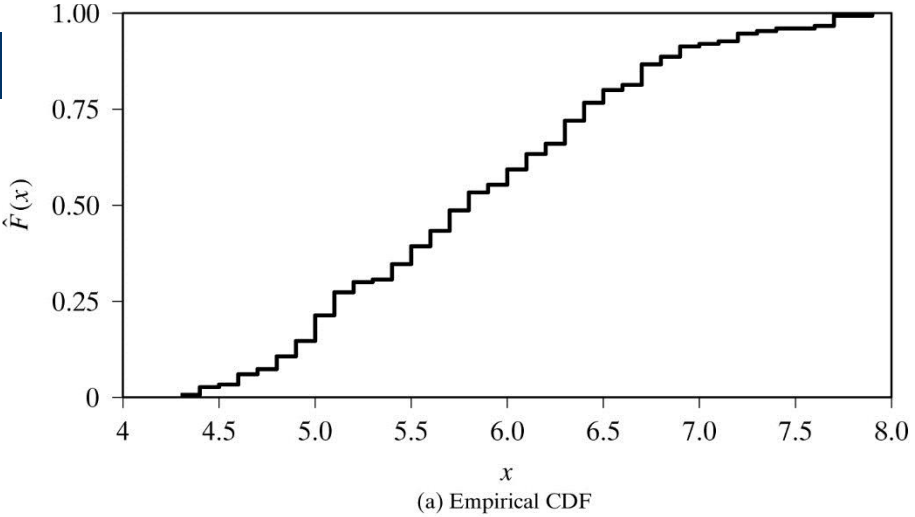
サンプリング Univariate

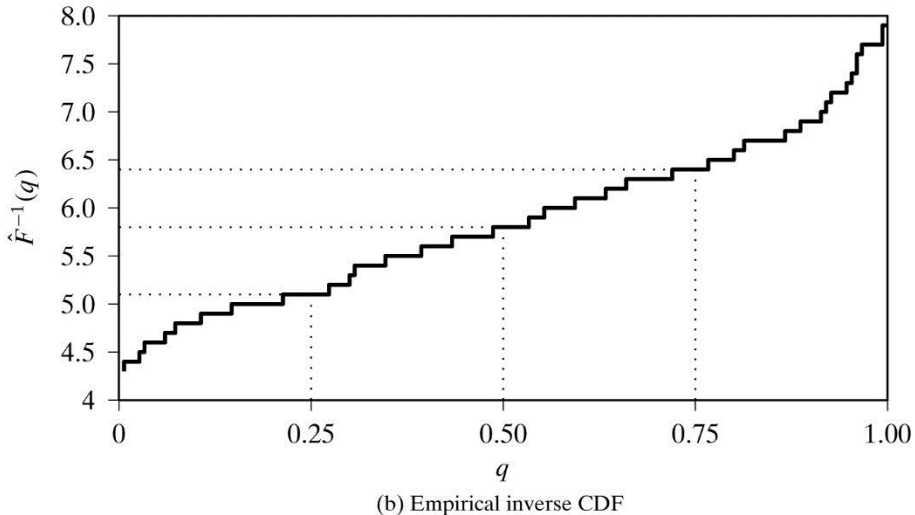
f, F, μ, σ, r: 母集団の確率分布,累積分布関数,平均,標準偏差,レンジ

$\hat{f}, \hat{F}, \hat{\mu}, \hat{\sigma}, \hat{r}$:サンプルの内容

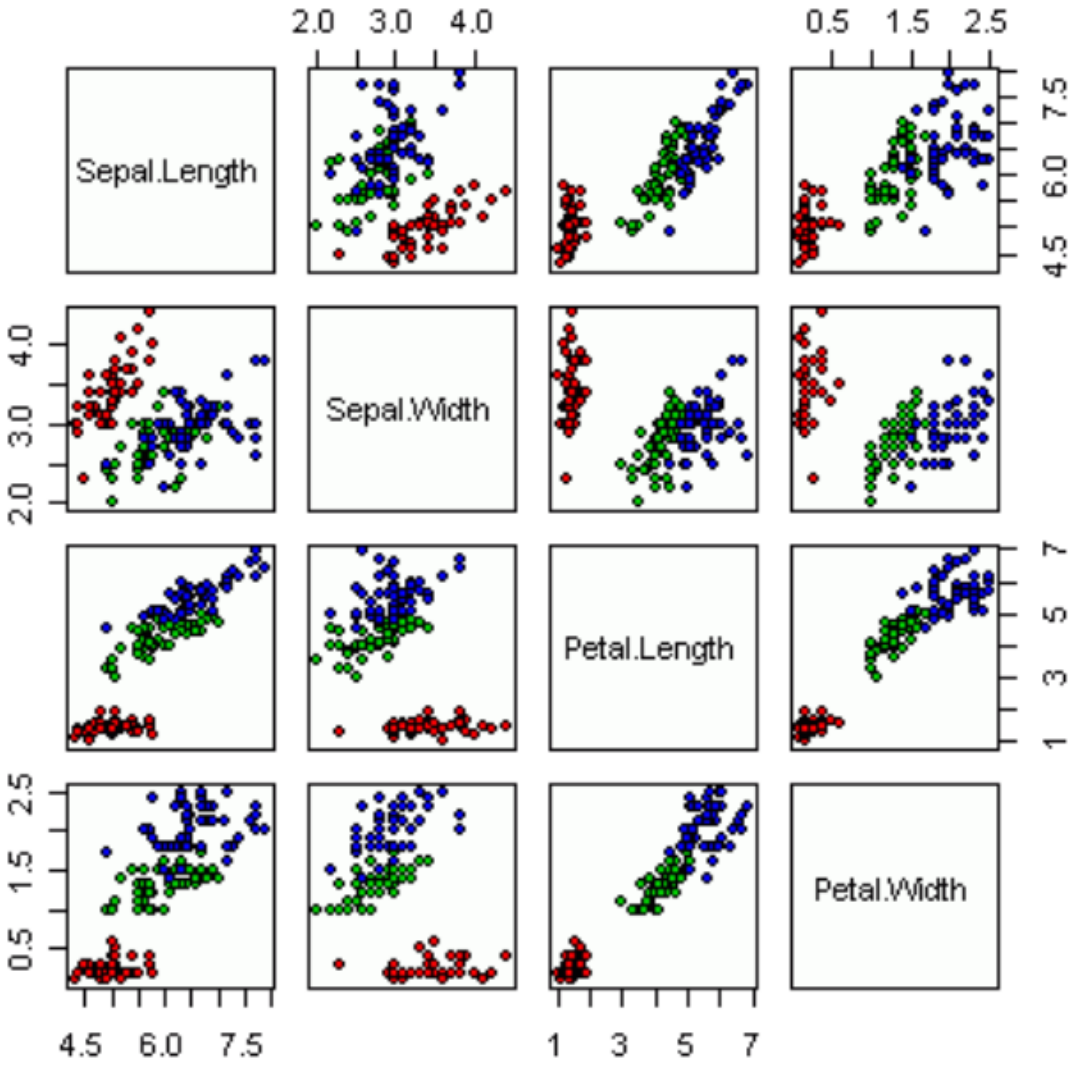




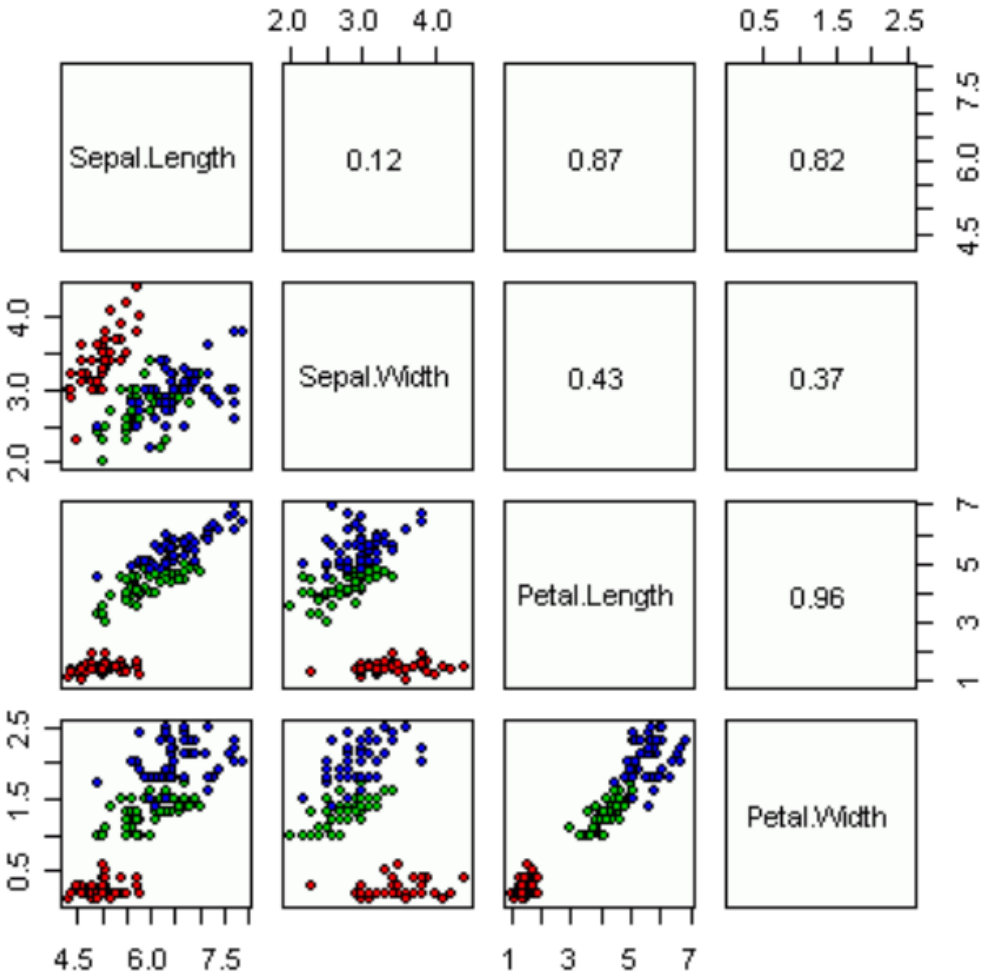




アイリスデータセットの属性の可視化
4次元の属性を2次元のプロット
(Draftsman's display)

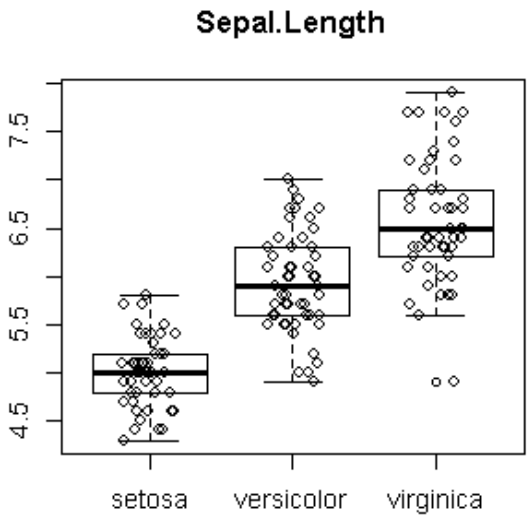


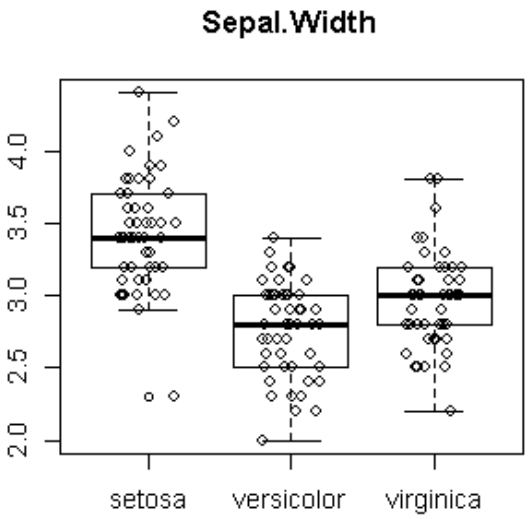
アイリスデータセットの属性の可視化
4次元の属性とピアソン相関性(相関係数)

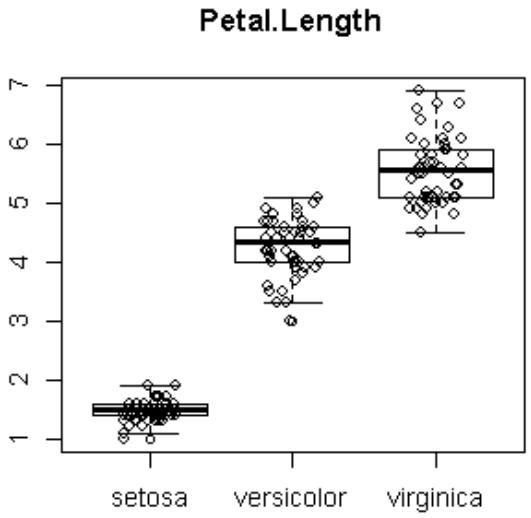


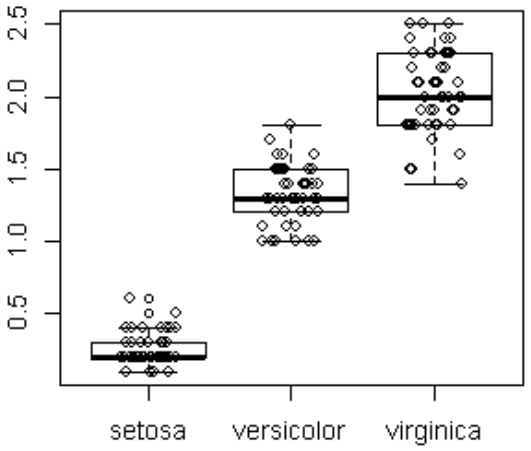
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\left(\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right)\right)^{1/2}}$$

アイリスデータセットの属性の可視化
Boxplot









確率変数の変換（Probabilistic Data view）

- 量的変数を離散変数に変換する
  - ビンの対応と作成
- 確率的な機械学習（ベイズ学習モデル）のデータを利用するために、量的変数を確率変数に変換必要ある

サンプリング

- 質的属性の扱い、一つの変数(univariate)
- ベルヌーイ変数

$$X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = (x_1 \quad x_2 \quad \cdots \quad x_n)^T$$

$$x_i \in dom(X) = \{a_1, a_2, \ldots, a_m\}$$

$$m = 2$$

ベルヌーイ変数: $X(v) = \begin{cases} 1 & if \quad v = a_1 \\ 0 & if \quad v = a_2 \end{cases}$

確率質量関数 (Probability mass function PMF): $P(X = x) = f(x) = \begin{cases} p_1 & if \quad x = 1 \\ p_0 & if \quad x = 0 \end{cases}$

サンプリング Univariate

f, F, μ, σ, r: 母集団の確率分布,累積分布関数,平均,標準偏差,レンジ

$\hat{f}, \hat{F}, \hat{\mu}, \hat{\sigma}, \hat{r}$:サンプルの内容

サンプリング

- 質的属性の扱い、一つの変数(univariate)
- ベルヌーイ変数

$p_1 + p_0 = 1$

$P(X = x) = f(x) = p^x(1-p)^{1-x}$, where $p_0 = p$ and $p_1 = 1 - p$

$\hat{\mu} = \hat{p}$

$\hat{\sigma} = var(X) = \hat{p}(1-\hat{p})$

Iris dataset: $dom(sepal\_length) = \{Long, Short\}$

$Long = [7, +\infty), Short = (-\infty, 7)$

平均: $\hat{\mu} = \hat{p} = 13/150 = 0.087$

分散: $\hat{\sigma} = var(X) = \hat{p}(1-\hat{p}) = 0.087(1-0.087) = 0.079$

出現数N: $E[N] = n\hat{p} = 150 \times 0.087 = 13$

出現数Nの分散: $var(N) = n\hat{p}(1-\hat{p}) = 150 \times 0.079 = 11.9$

サンプリング

- 質的属性の扱い、一つの変数(univariate)
- 複数ベルヌーイ変数

複数ベルヌーイ変数:

$$e_i = \begin{pmatrix} 0 & \cdots & 0 & 1(\text{成分}i) & 0 & \cdots & 0 \end{pmatrix}^T$$

$$X(v) = e_i \text{ if } v = a_i$$

確率質量関数 (Probability mass function PMF): $P(X = e_i) = f(e_i) = p_i,\ \Sigma_{i=1}^{m} p_i = 1$

- 共分散行列

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1m} & \sigma_{2m} & \cdots & \sigma_m^2 \end{pmatrix} = \begin{pmatrix} p_1(1-p_1) & -p_1 p_2 & \cdots & -p_1 p_m \\ -p_1 p_2 & p_2(1-p_2) & \cdots & -p_2 p_m \\ \vdots & \vdots & \ddots & \vdots \\ -p_1 p_m & -p_2 p_m & \cdots & p_m(1-p_m) \end{pmatrix}$$

$$\hat{\Sigma} = \hat{\mathbf{P}} - \hat{\mathbf{p}} \cdot \hat{\mathbf{p}}^T \quad \text{where } \hat{\mathbf{P}} = \text{diag}(\hat{\mathbf{p}}), \text{ and } \hat{\mathbf{p}} = \hat{\mu} = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_m)^T$$

サンプリング

- 質的属性の扱い、一つの変数(univariate)
- 複数ベルヌーイ変数

Table 3.1. Discretized `sepal length` attribute

Bins	Domain	Counts
[4.3, 5.2]	Very Short (a_1)	$n_1 = 45$
(5.2, 6.1]	Short (a_2)	$n_2 = 50$
(6.1, 7.0]	Long (a_3)	$n_3 = 43$
(7.0, 7.9]	Very Long (a_4)	$n_4 = 12$

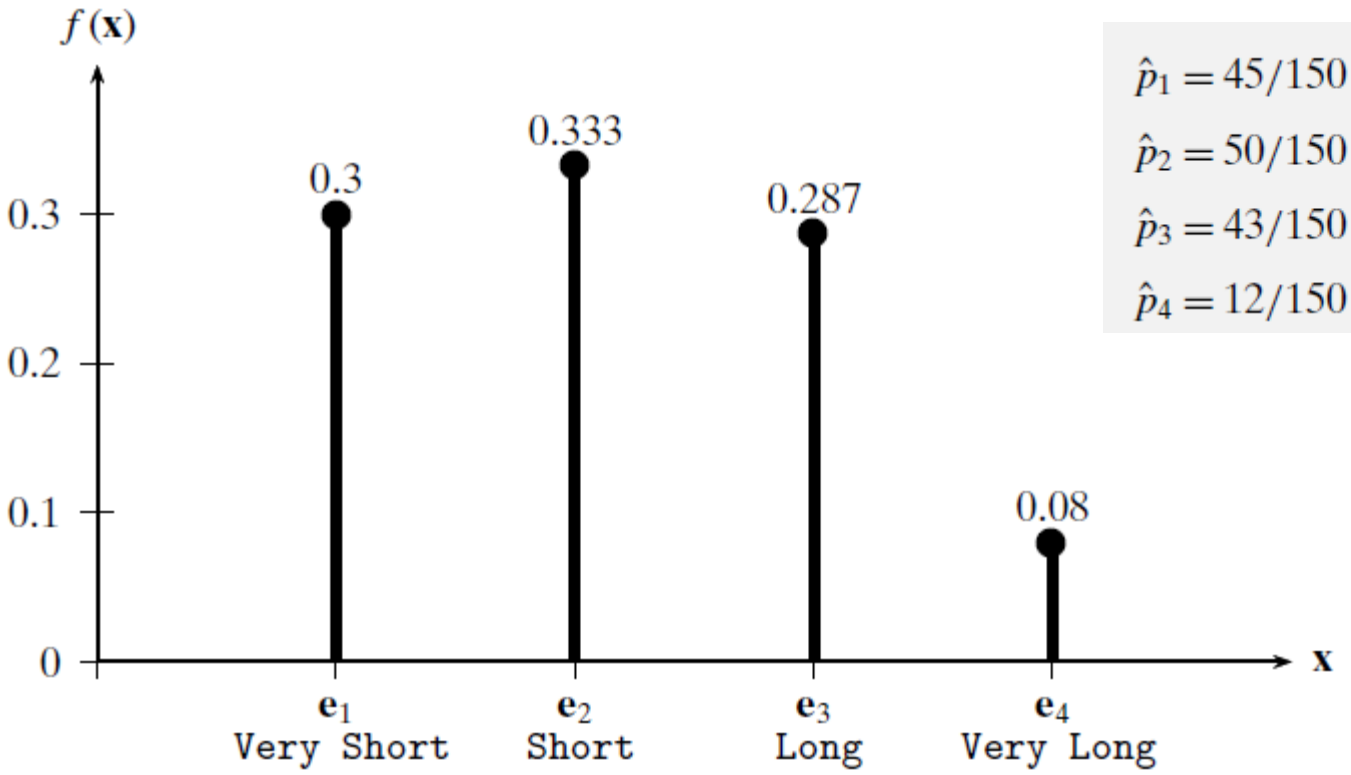
$$\{a_1 = \texttt{VeryShort}, a_2 = \texttt{Short}, a_3 = \texttt{Long}, a_4 = \texttt{VeryLong}\}$$

$$\mathbf{X}(v) = \begin{cases} \mathbf{e}_1 = (1,0,0,0) & \text{if } v = a_1 \\ \mathbf{e}_2 = (0,1,0,0) & \text{if } v = a_2 \\ \mathbf{e}_3 = (0,0,1,0) & \text{if } v = a_3 \\ \mathbf{e}_4 = (0,0,0,1) & \text{if } v = a_4 \end{cases}$$

サンプリング

- 質的属性の扱い、一つの変数(univariate)
- 複数ベルヌーイ変数

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i = \sum_{i=1}^{m}\frac{n_i}{n}\mathbf{e}_i = \begin{pmatrix} n_1/n \\ n_2/n \\ \vdots \\ n_m/n \end{pmatrix} = \begin{pmatrix} \hat{p}_1 \\ \hat{p}_2 \\ \vdots \\ \hat{p}_m \end{pmatrix} = \hat{\mathbf{p}}$$



$\hat{p}_1 = 45/150 = 0.3$

$\hat{p}_2 = 50/150 = 0.333$

$\hat{p}_3 = 43/150 = 0.287$

$\hat{p}_4 = 12/150 = 0.08$

Figure 3.1. Probability mass function: `sepal length`.

サンプリング

- 質的属性の扱い、一つの変数(univariate)
- 複数ベルヌーイ変数

$$\hat{\boldsymbol{\Sigma}} = \hat{\mathbf{P}} - \hat{\mathbf{p}} \cdot \hat{\mathbf{p}}^T$$

$$= \begin{pmatrix} 0.3 & 0 & 0 & 0 \\ 0 & 0.333 & 0 & 0 \\ 0 & 0 & 0.287 & 0 \\ 0 & 0 & 0 & 0.08 \end{pmatrix} - \begin{pmatrix} 0.3 \\ 0.333 \\ 0.287 \\ 0.08 \end{pmatrix} \begin{pmatrix} 0.3 & 0.333 & 0.287 & 0.08 \end{pmatrix}$$

$$= \begin{pmatrix} 0.3 & 0 & 0 & 0 \\ 0 & 0.333 & 0 & 0 \\ 0 & 0 & 0.287 & 0 \\ 0 & 0 & 0 & 0.08 \end{pmatrix} - \begin{pmatrix} 0.09 & 0.1 & 0.086 & 0.024 \\ 0.1 & 0.111 & 0.096 & 0.027 \\ 0.086 & 0.096 & 0.082 & 0.023 \\ 0.024 & 0.027 & 0.023 & 0.006 \end{pmatrix}$$

$$= \begin{pmatrix} 0.21 & -0.1 & -0.086 & -0.024 \\ -0.1 & 0.222 & -0.096 & -0.027 \\ -0.086 & -0.096 & 0.204 & -0.023 \\ -0.024 & -0.027 & -0.023 & 0.074 \end{pmatrix}$$

サンプリング

- 質的属性の扱い、二つの変数(bivariate)

$$\mathbf{D} = \begin{pmatrix} X_1 & X_2 \\ \hline x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{pmatrix}$$

$$dom(X_1) = \{a_{11}, a_{12}, \ldots, a_{1m_1}\}$$

$$dom(X_2) = \{a_{21}, a_{22}, \ldots, a_{2m_2}\}$$

$$P(\mathbf{X}_1 = \mathbf{e}_{1i}) = f_1(\mathbf{e}_{1i}) = p_i^1 = \prod_{k=1}^{m_1} (p_i^1)^{e_{ik}^1}$$

$$P(\mathbf{X}_2 = \mathbf{e}_{2j}) = f_2(\mathbf{e}_{2j}) = p_j^2 = \prod_{k=1}^{m_2} (p_j^2)^{e_{jk}^2}$$

$$\mathbf{X}\left((v_1, v_2)^T\right) = \begin{pmatrix} \mathbf{X}_1(v_1) \\ \mathbf{X}_2(v_2) \end{pmatrix} = \begin{pmatrix} \mathbf{e}_{1i} \\ \mathbf{e}_{2j} \end{pmatrix}$$

共分布、条件:

$$P\left(\mathbf{X} = (\mathbf{e}_{1i}, \mathbf{e}_{2j})^T\right) = f(\mathbf{e}_{1i}, \mathbf{e}_{2j}) = p_{ij} = \prod_{r=1}^{m_1} \prod_{s=1}^{m_2} p_{ij}^{e_{ir}^1 \cdot e_{js}^2}$$

$$\sum_{i=1}^{m_1} \sum_{j=1}^{m_2} p_{ij} = 1$$

サンプリング

- 質的属性の扱い、二つの変数(bivariate)
- 平均、共分散行列

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i = \frac{1}{n}\begin{pmatrix}\sum_{i=1}^{m_1} n_i^1 \mathbf{e}_{1i} \\ \sum_{j=1}^{m_2} n_j^2 \mathbf{e}_{2j}\end{pmatrix} = \frac{1}{n}\begin{pmatrix} n_1^1 \\ \vdots \\ n_{m_1}^1 \\ n_1^2 \\ \vdots \\ n_{m_2}^2 \end{pmatrix} = \begin{pmatrix} \hat{p}_1^1 \\ \vdots \\ \hat{p}_{m_1}^1 \\ \hat{p}_1^2 \\ \vdots \\ \hat{p}_{m_2}^2 \end{pmatrix} = \begin{pmatrix}\hat{\mathbf{p}}_1 \\ \hat{\mathbf{p}}_2\end{pmatrix} = \begin{pmatrix}\hat{\boldsymbol{\mu}}_1 \\ \hat{\boldsymbol{\mu}}_2\end{pmatrix}$$

$$\hat{\boldsymbol{\Sigma}} = \begin{pmatrix}\hat{\boldsymbol{\Sigma}}_{11} & \hat{\boldsymbol{\Sigma}}_{12} \\ \hat{\boldsymbol{\Sigma}}_{12}^T & \hat{\boldsymbol{\Sigma}}_{22}\end{pmatrix}$$

$$I_{ij}(\mathbf{x}_k) = \begin{cases} 1 & \text{if } \mathbf{x}_{k1} = \mathbf{e}_{1i} \text{ and } \mathbf{x}_{k2} = \mathbf{e}_{2j} \\ 0 & \text{otherwise} \end{cases}$$

$$\hat{\mathbf{P}}_1 = \text{diag}(\hat{\mathbf{p}}_1) \text{ and } \hat{\mathbf{P}}_2 = \text{diag}(\hat{\mathbf{p}}_2)$$

$$\hat{\boldsymbol{\Sigma}}_{11} = \hat{\mathbf{P}}_1 - \hat{\mathbf{p}}_1\hat{\mathbf{p}}_1^T$$

$$\hat{\boldsymbol{\Sigma}}_{22} = \hat{\mathbf{P}}_2 - \hat{\mathbf{p}}_2\hat{\mathbf{p}}_2^T \quad \hat{\mathbf{P}}_{12}(i,j) = \hat{f}(\mathbf{e}_{1i}, \mathbf{e}_{2j}) = \frac{1}{n}\sum_{k=1}^{n} I_{ij}(\mathbf{x}_k) = \frac{n_{ij}}{n} = \hat{p}_{ij}$$

$$\hat{\boldsymbol{\Sigma}}_{12} = \hat{\mathbf{P}}_{12} - \hat{\mathbf{p}}_1\hat{\mathbf{p}}_2^T$$

サンプリング

- 質的属性の扱い、二つの変数(bivariate)

$$dom(X_1) = \{a_{11} = \texttt{VeryShort}, a_{12} = \texttt{Short}, a_{13} = \texttt{Long}, a_{14} = \texttt{VeryLong}\}$$
$$dom(X_2) = \{a_{21} = \texttt{Short}, a_{22} = \texttt{Medium}, a_{23} = \texttt{Long}\}$$

Table 3.3. Discretized `sepal width` attribute

Bins	Domain	Counts
[2.0, 2.8]	Short (a_1)	47
(2.8, 3.6]	Medium (a_2)	88
(3.6, 4.4]	Long (a_3)	15

$$\hat{\boldsymbol{\mu}}_1 = \hat{\mathbf{p}}_1 = \begin{pmatrix} 0.3 \\ 0.333 \\ 0.287 \\ 0.08 \end{pmatrix} \qquad \hat{\boldsymbol{\mu}}_2 = \hat{\mathbf{p}}_2 = \frac{1}{150}\begin{pmatrix} 47 \\ 88 \\ 15 \end{pmatrix} = \begin{pmatrix} 0.313 \\ 0.587 \\ 0.1 \end{pmatrix}$$

$$\hat{\boldsymbol{\mu}} = \begin{pmatrix} \hat{\boldsymbol{\mu}}_1 \\ \hat{\boldsymbol{\mu}}_2 \end{pmatrix} = \begin{pmatrix} \hat{\mathbf{p}}_1 \\ \hat{\mathbf{p}}_2 \end{pmatrix} = (0.3, 0.333, 0.287, 0.08 \mid 0.313, 0.587, 0.1)^T$$

サンプリング

- 質的属性の扱い、二つの変数(bivariate)

Table 3.4. Observed Counts (n_{ij}): `sepal length` and `sepal width`

		X_2		
		Short ($\mathbf{e}_{21}$)	Medium ($\mathbf{e}_{22}$)	Long ($\mathbf{e}_{23}$)
X_1	Very Short ($\mathbf{e}_{11}$)	7	33	5
	Short ($\mathbf{e}_{22}$)	24	18	8
	Long ($\mathbf{e}_{13}$)	13	30	0
	Very Long ($\mathbf{e}_{14}$)	3	7	2

$$E[\mathbf{X}_1]E[\mathbf{X}_2]^T = \hat{\boldsymbol{\mu}}_1\hat{\boldsymbol{\mu}}_2^T = \hat{\mathbf{p}}_1\hat{\mathbf{p}}_2^T$$

$$= \begin{pmatrix} 0.3 \\ 0.333 \\ 0.287 \\ 0.08 \end{pmatrix} \begin{pmatrix} 0.313 & 0.587 & 0.1 \end{pmatrix}$$

$$= \begin{pmatrix} 0.094 & 0.176 & 0.03 \\ 0.104 & 0.196 & 0.033 \\ 0.09 & 0.168 & 0.029 \\ 0.025 & 0.047 & 0.008 \end{pmatrix}$$

サンプリング

- 質的属性の扱い、二つの変数(bivariate)

$$\hat{\Sigma}_{11} = \begin{pmatrix} 0.21 & -0.1 & -0.086 & -0.024 \\ -0.1 & 0.222 & -0.096 & -0.027 \\ -0.086 & -0.096 & 0.204 & -0.023 \\ -0.024 & -0.027 & -0.023 & 0.074 \end{pmatrix} \quad \hat{\Sigma}_{22} = \begin{pmatrix} 0.215 & -0.184 & -0.031 \\ -0.184 & 0.242 & -0.059 \\ -0.031 & -0.059 & 0.09 \end{pmatrix}$$

$$E[\mathbf{X}_1\mathbf{X}_2^T] = \hat{\mathbf{P}}_{12} = \frac{1}{150}\begin{pmatrix} 7 & 33 & 5 \\ 24 & 18 & 8 \\ 13 & 30 & 0 \\ 3 & 7 & 2 \end{pmatrix} = \begin{pmatrix} 0.047 & 0.22 & 0.033 \\ 0.16 & 0.12 & 0.053 \\ 0.087 & 0.2 & 0 \\ 0.02 & 0.047 & 0.013 \end{pmatrix}$$

$$\hat{\Sigma}_{12} = \hat{\mathbf{P}}_{12} - \hat{\mathbf{p}}_1\hat{\mathbf{p}}_2^T$$

$$= \begin{pmatrix} -0.047 & 0.044 & 0.003 \\ 0.056 & -0.076 & 0.02 \\ -0.003 & 0.032 & -0.029 \\ -0.005 & 0 & 0.005 \end{pmatrix}$$

サンプリング

- 質的属性の扱い、二つの変数(bivariate)

$$\hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_{11} & \hat{\Sigma}_{12} \\ \hat{\Sigma}_{12}^T & \hat{\Sigma}_{22} \end{pmatrix}$$

$$= \left(\begin{array}{cccc|ccc} 0.21 & -0.1 & -0.086 & -0.024 & -0.047 & 0.044 & 0.003 \\ -0.1 & 0.222 & -0.096 & -0.027 & 0.056 & -0.076 & 0.02 \\ -0.086 & -0.096 & 0.204 & -0.023 & -0.003 & 0.032 & -0.029 \\ -0.024 & -0.027 & -0.023 & 0.074 & -0.005 & 0 & 0.005 \\ \hline -0.047 & 0.056 & -0.003 & -0.005 & 0.215 & -0.184 & -0.031 \\ 0.044 & -0.076 & 0.032 & 0 & -0.184 & 0.242 & -0.059 \\ 0.003 & 0.02 & -0.029 & 0.005 & -0.031 & -0.059 & 0.09 \end{array}\right)$$

サンプリング

- 質的属性の扱い、二つの変数(bivariate)

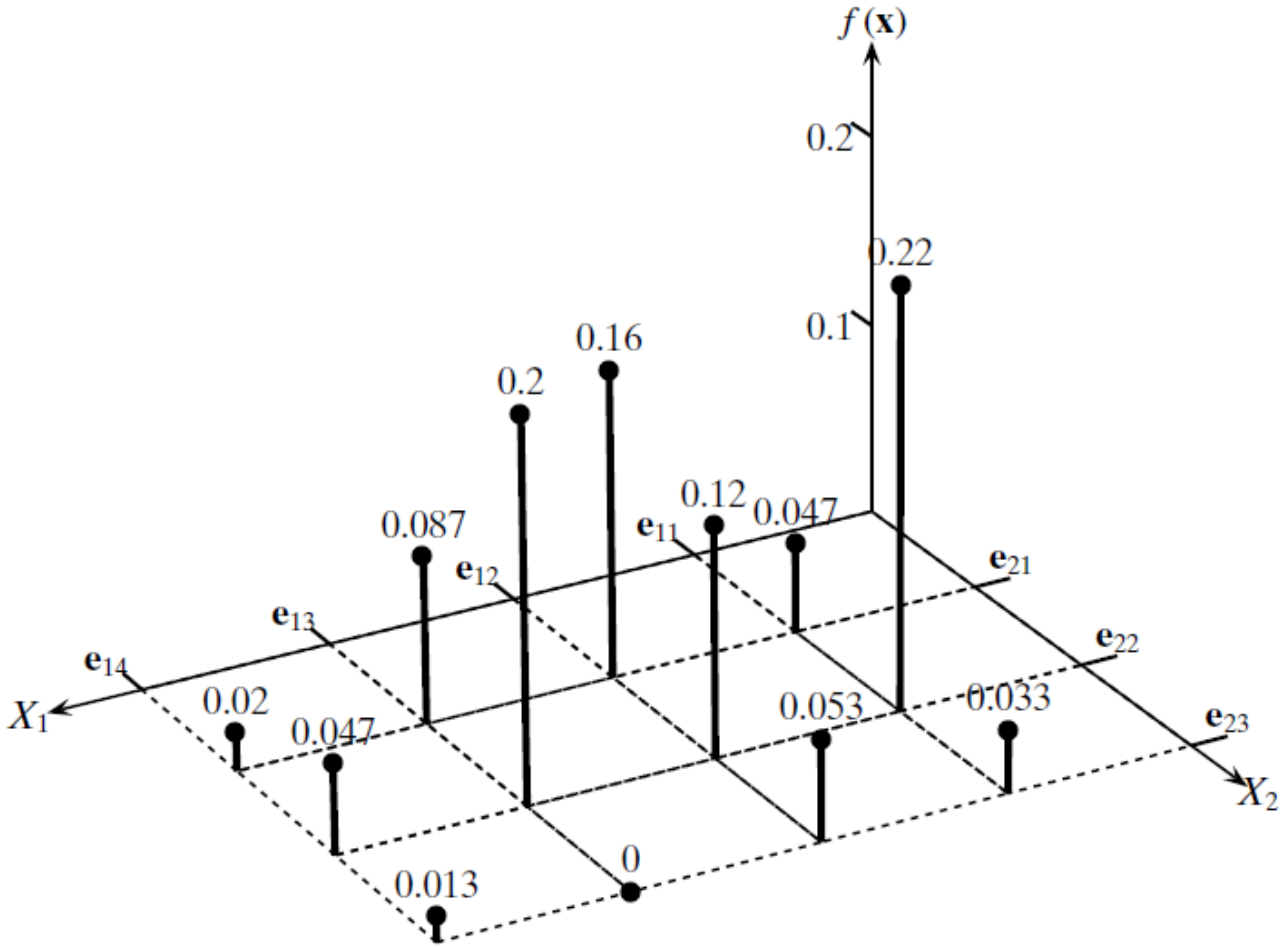


Figure 3.2. Empirical joint probability mass function: sepal length and sepal width.

サンプリング

- 質的属性の扱い、二つの変数(bivariate)
- 属性の独立性:2つの変数間の「付随性」

Table 3.5. Contingency table: sepal length vs. sepal width

Sepal length (X_1)	Sepal width (X_2)			
	Short a_{21}	Medium a_{22}	Long a_{23}	Row Counts
Very Short (a_{11})	7	33	5	$n_1^1 = 45$
Short (a_{12})	24	18	8	$n_2^1 = 50$
Long (a_{13})	13	30	0	$n_3^1 = 43$
Very Long (a_{14})	3	7	2	$n_4^1 = 12$
Column Counts	$n_1^2 = 47$	$n_2^2 = 88$	$n_3^2 = 15$	$n = 150$

$$\mathbf{N}_{12} = n \cdot \widehat{\mathbf{P}}_{12} = \begin{pmatrix} n_{11} & n_{12} & \cdots & n_{1m_2} \\ n_{21} & n_{22} & \cdots & n_{2m_2} \\ \vdots & \vdots & \ddots & \vdots \\ n_{m_1 1} & n_{m_1 2} & \cdots & n_{m_1 m_2} \end{pmatrix}$$

サンプリング

- 質的属性の扱い、二つの変数(bivariate)
- 仮説検定、カイ二統計、q自由度,

$$\chi^2 = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\frac{(n_{ij}-e_{ij})^2}{e_{ij}} \qquad e_{ij} = n\cdot\hat{p}_{ij} = n\cdot\hat{p}_i^1\cdot\hat{p}_j^2 = n\cdot\frac{n_i^1}{n}\cdot\frac{n_j^2}{n} = \frac{n_i^1 n_j^2}{n}$$

$$f(x|q) = \frac{1}{2^{q/2}\Gamma(q/2)}x^{\frac{q}{2}-1}\mathrm{e}^{-\frac{x}{2}} \qquad \Gamma(k>0) = \int_0^\infty x^{k-1}\mathrm{e}^{-x}dx$$

$$e_{11} = \frac{n_1^1 n_1^2}{n} = \frac{45\cdot 47}{150} = \frac{2115}{150} = 14.1$$

$$q = (m_1-1)\cdot(m_2-1) = 3\cdot 2 = 6$$

$$\chi^2 = 21.8.$$

サンプリング

- 質的属性の扱い、二つの変数(bivariate)
- 仮説検定、カイ二統計、q自由度,

Table 3.6. Expected counts

		X_2		
		Short (a_{21})	Medium (a_{22})	Short (a_{23})
X_1	Very Short (a_{11})	14.1	26.4	4.5
	Short (a_{12})	15.67	29.33	5.0
	Long (a_{13})	13.47	25.23	4.3
	Very Long (a_{14})	3.76	7.04	1.2

F検定を行う

サンプリング

- 質的属性の扱い、二つの変数(bivariate)
- 仮説検定、カイ二統計、q自由度,

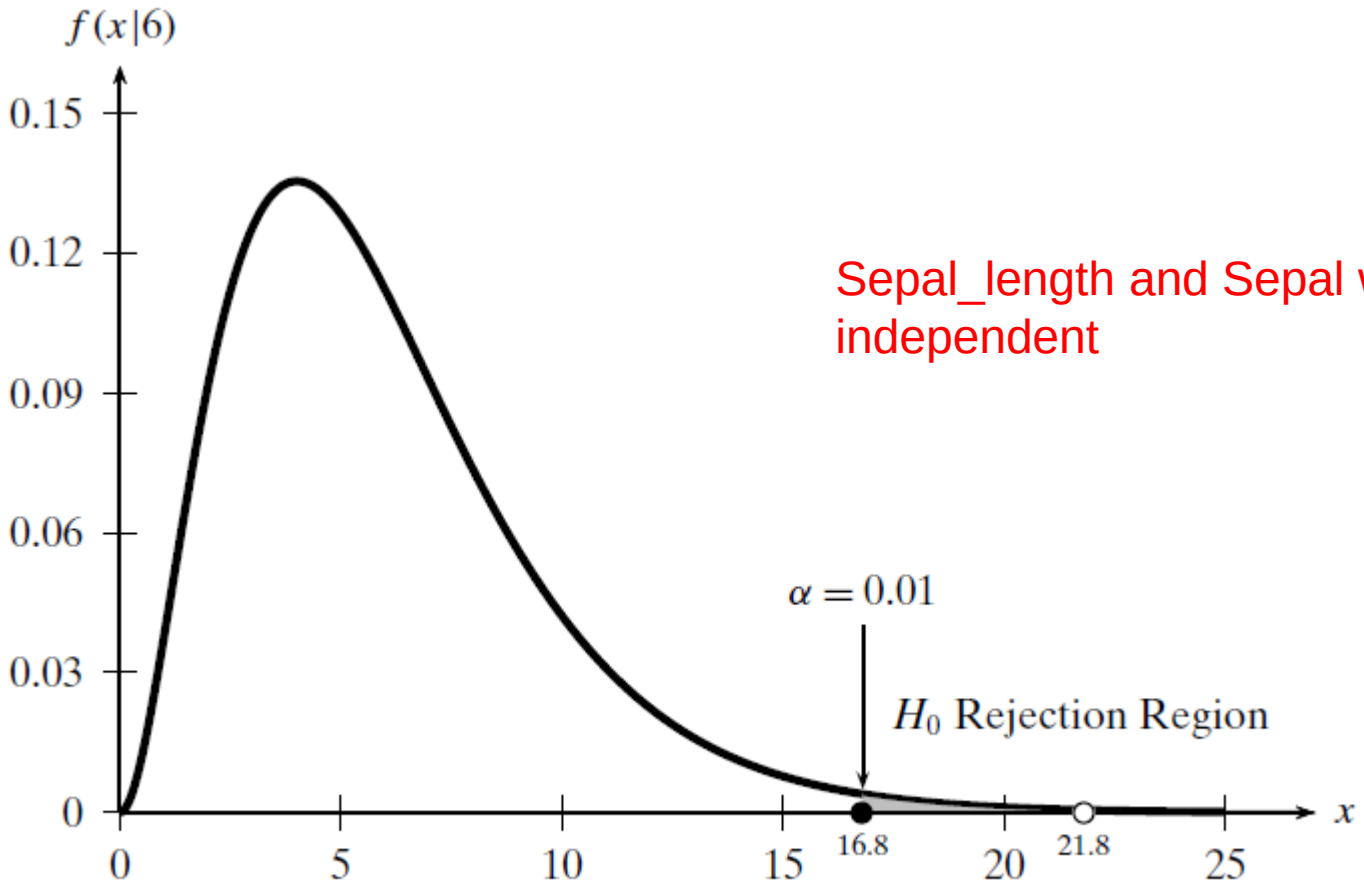


Figure 3.3. Chi-squared distribution ($q = 6$).

Sepal_length and Sepal width are independent

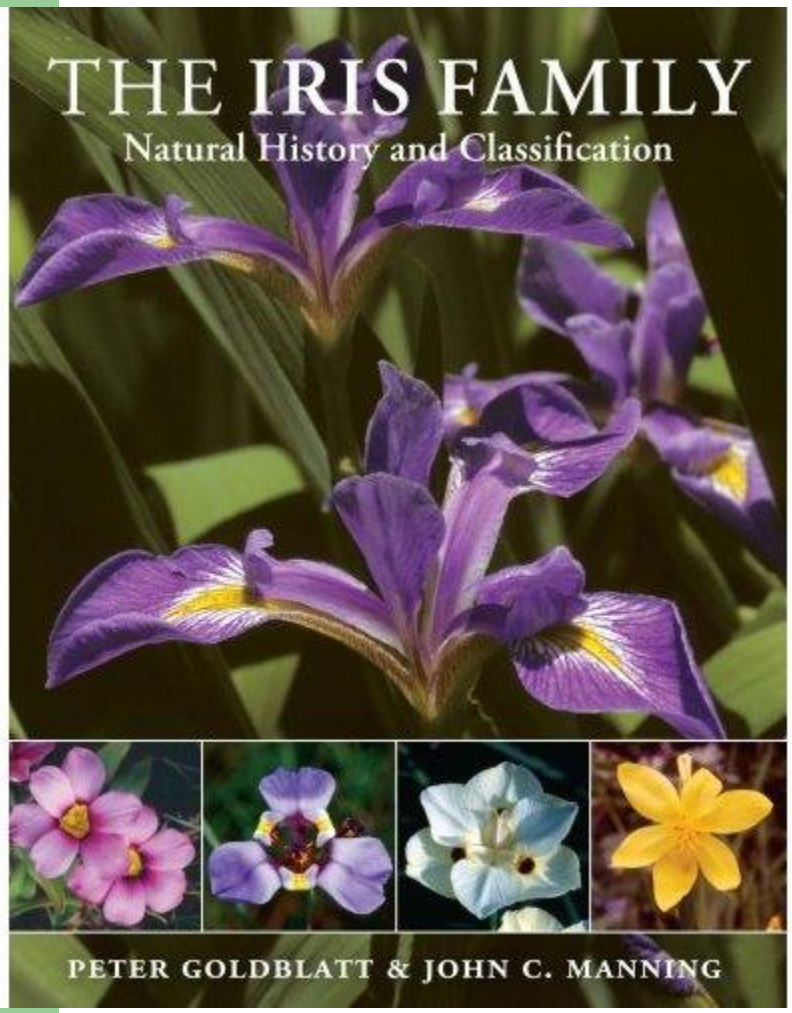
サンプリング

- 質的属性の扱い、二つ以上の変数 (multivariate)

サンプリング Multivariate

$\hat{f}, \hat{F}, \hat{\mu}, \hat{\sigma}, \hat{\Sigma}$:ベクトル平均、共分散、行列共分散

問題定式化



- 入力:
- C: クラス（classes）
- F: 特徴(Features)
- 出力:
- 特徴データとクラスの分類

データセット例(Iris dataset)

Setosa

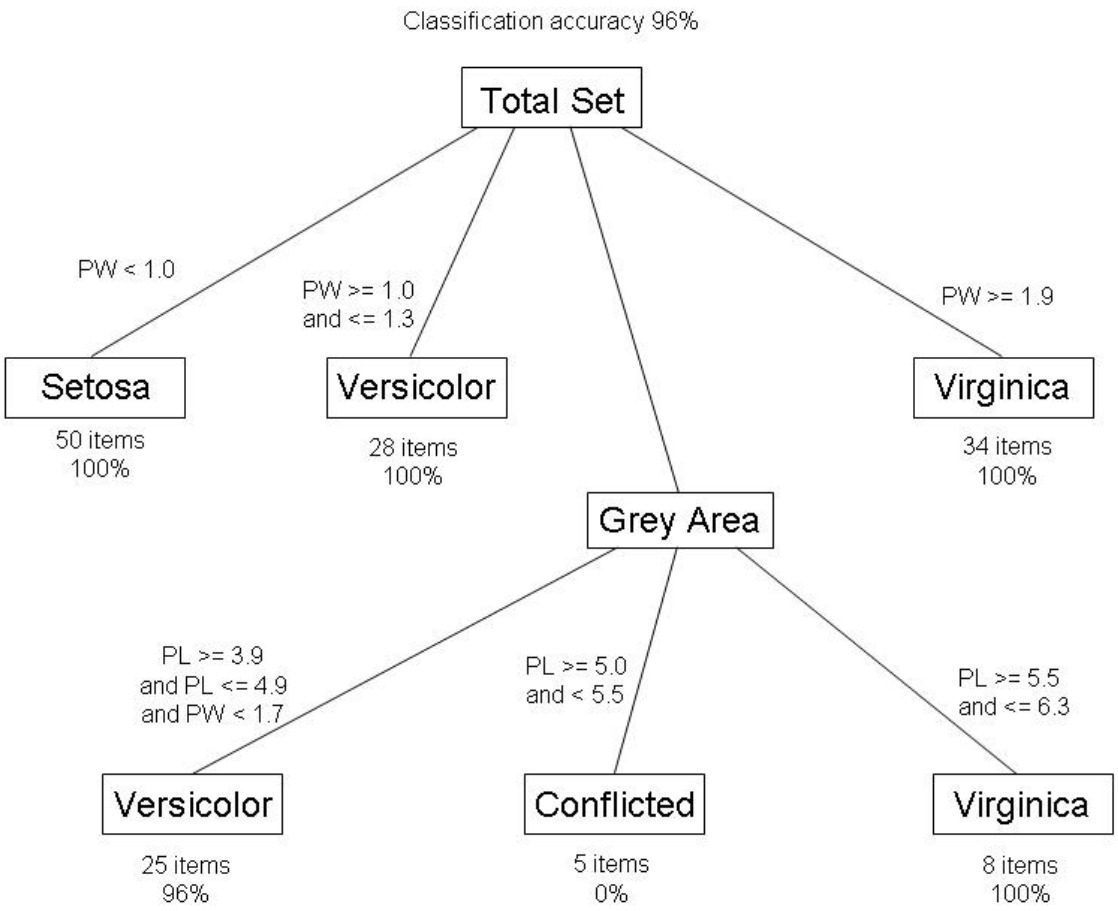
D1:{PL=1.4,PW=0.2,SL=5.1,SW=3.5}
D2:{PL=1.4,PW=0.2,SL=4.9,SW=3.0}
D3:{PL=1.3,PW=0.2,SL=4.7,SW=3.2}
…

D51:{PL=4.7,PW=1.4,SL=7.0,SW=3.2}
D52:{PL=4.9,PW=1.5,SL=6.4,SW=3.2}
D53:{PL=4.0,PW=1.5,SL=6.9,SW=3.1}
…

D101:{PL=6.0,PW=2.5,SL=6.3,SW=3.3}
D102:{PL=5.1,PW=1.9,SL=5.8,SW

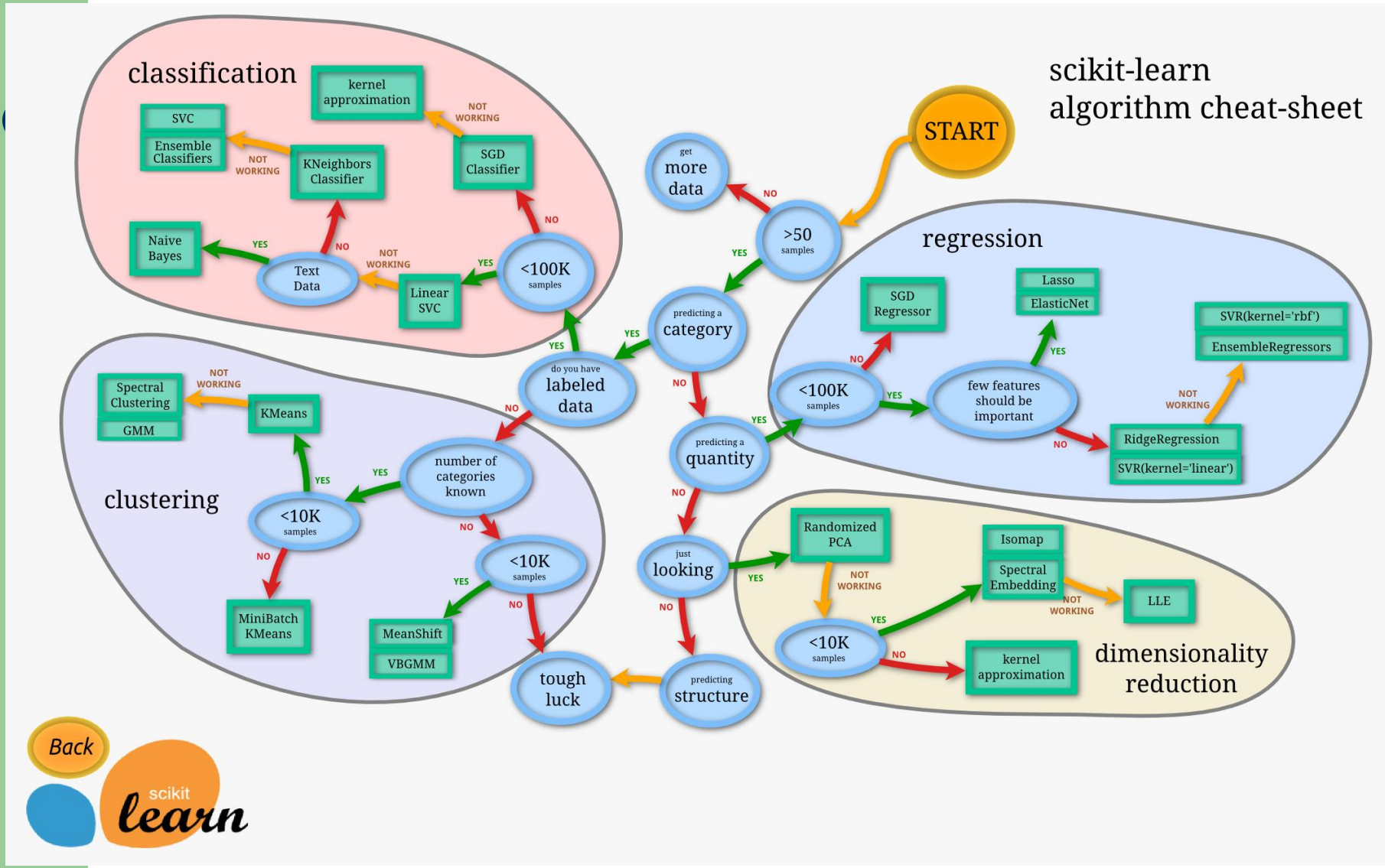
- 3つのクラスに50サンプル毎に収集
- 特徴
  - **PW: Petal Width**
  - **PL: Petal Length**
  - **SW: SepalWidth**
  - **SL: Sepal Length**

データセットの分類(Iris dataset)



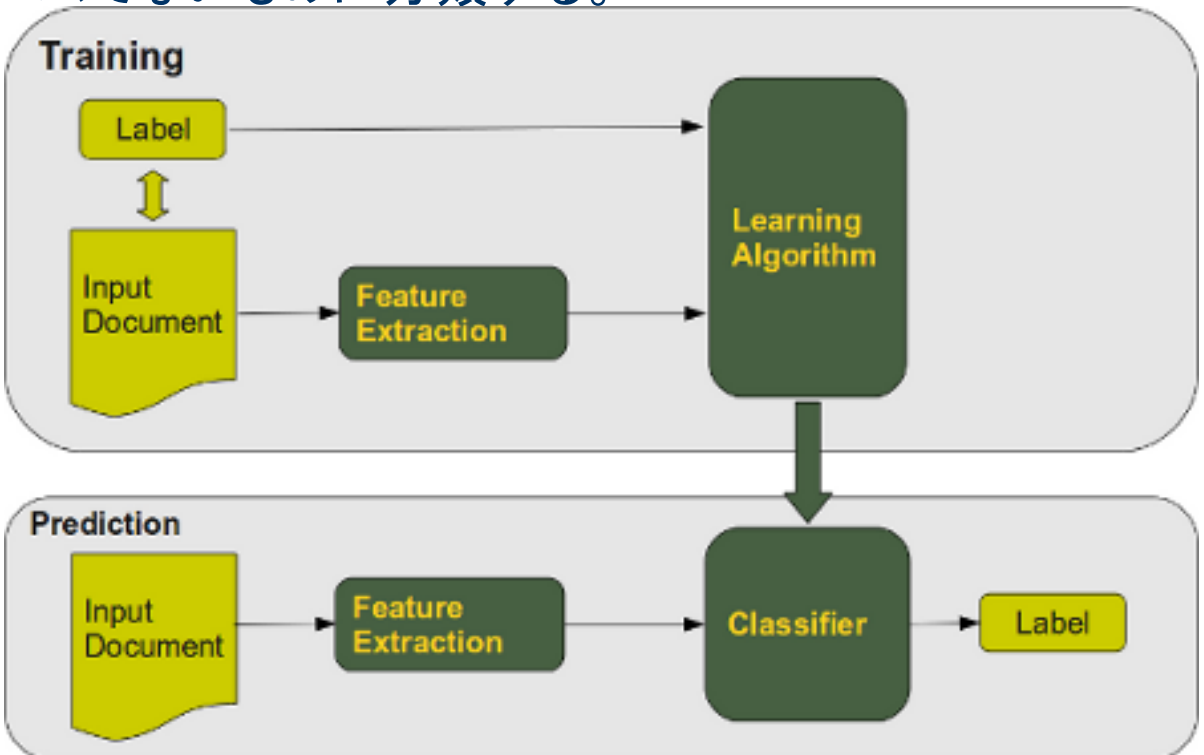
- 特徴
  - PW: Petal Width
  - PL: Petal Length

分類器/クラスタ分析器/回帰など

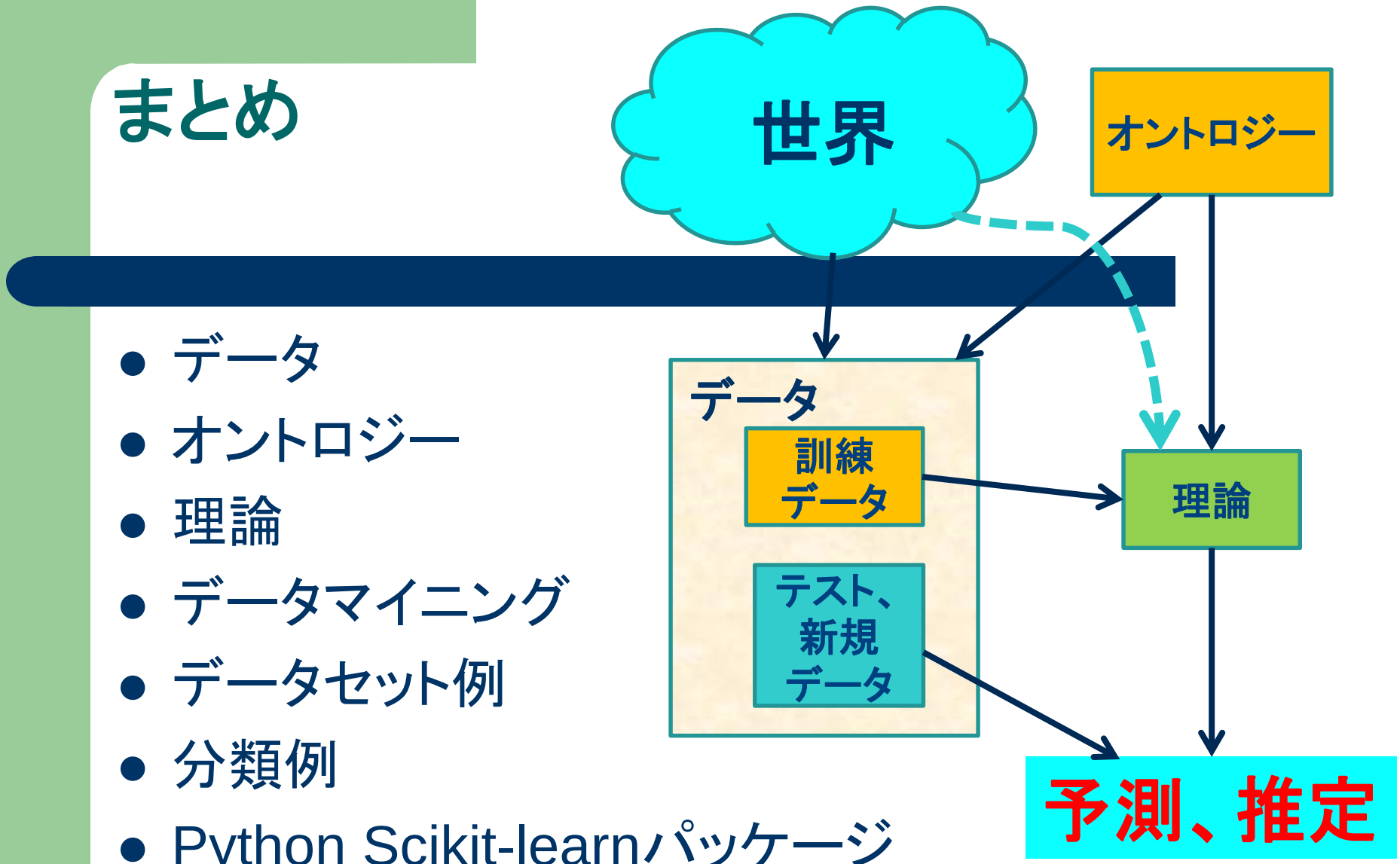


文書の自動分類

- 単純ベイズ分類器を文書分類問題に適用した例を示す。文書群をその内容によって分類する問題であり、例えば、電子メールをスパムとスパムでないものに分類する。



まとめ



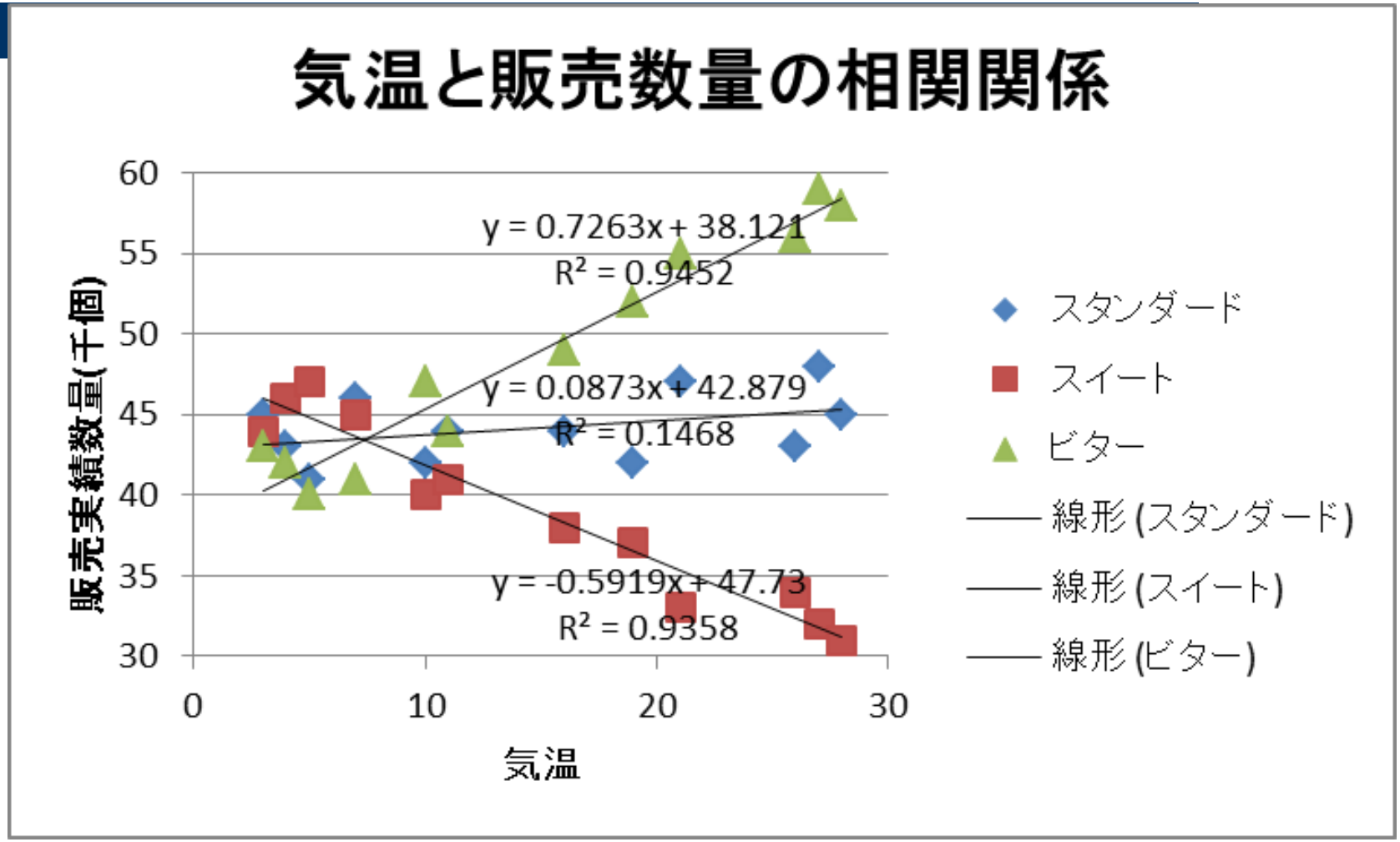
- データ
- オントロジー
- 理論
- データマイニング
- データセット例
- 分類例
- Python Scikit-learnパッケージ
- データテキストの分類器

EXAMPLE

データ例:コーヒのテストを分析する

月	昨年の実測気温	スタンダード	スイート	ビター
1	4	43	46	42
2	3	45	44	43
3	5	41	47	40
4	10	42	40	47
5	16	44	38	49
6	21	47	33	55
7	27	48	32	59
8	28	45	31	58
9	26	43	34	56
10	19	42	37	52
11	11	44	41	44
12	7	46	45	41

回帰分析の例


気温と販売数量の相関関係
販売実績数量(千個)
気温
60
55
50
45
40
35
30
0
10
20
30
y = 0.7263x + 38.121
R² = 0.9452
y = 0.0873x + 42.879
R² = 0.1468
y = -0.5919x + 47.73
R² = 0.9358
スタンダード
スイート
ビター
線形 (スタンダード)
線形 (スイート)
線形 (ビター)

気温予測に基づく販売数量		
本年の予測気温	スイート	ビター
8	42.99529	43.93098
10	41.8115	45.38353
9	42.4034	44.65725
14	39.44392	48.28863
16	38.26013	49.74118
23	34.11686	54.8251
28	31.15739	58.45647
29	30.56549	59.18275
26	32.34118	57.00392
20	35.89255	52.64627
9	42.4034	44.65725
6	44.17908	42.47843

	前年比	
月	スイート	ビター
1	−7%	5%
2	−5%	6%
3	−10%	12%
4	−1%	3%
5	1%	2%
6	3%	0%
7	−3%	−1%
8	−1%	2%
9	−5%	2%
10	−3%	1%
11	3%	1%
12	−2%	4%

1年間の販売数量の予測