知能システム開発特論: 第5回

データマイニング:
ベイズ、単純ベイズ分類器、k近傍法

ダビド (david@iwate-pu.ac.jp)
IPU

18 January 2021

データマイニングは機械学習か？

- 分類器
- クラスター器
- 回帰器
- ニュラルネット
- PCA
- 決定木器
- …

知識を獲得したい！！！

学習の手法

- 情報から従来使えそうな知識をどんな方法で見つけ出す
- 例題（事例）からアルゴリズムに基づいて概念記述（structural descriptions)を取得
- 概念記述には明示的にパターンを定義される

データマイニングの機能

- 分類と予測
  - –予測を目的として、クラスや概念を区別するための記述を求める。
  - –例国を気候により分類する。自動車を燃費で分類する。
  - –表現: 決定木、分類規則、ニューラルネットワーク
  - –予測:未知の、あるいは欠落した数値(missing values)を予測する。
- クラスター分析
  - –分類カテゴリが未知: 新しいクラスを作るためにグループ化する。たとえば、分布パターンを見つけるために顧客の住居をクラスター化する。
  - –クラスタリングの原理: クラス内の類似性を最大化し、同時に、クラス間の類似性を最小にする。

ベイズ分類器

- データセットDが与える時、n サンプルのx_iと $y_i \epsilon \{c_1, c_2, \ldots, c_k\}$
- ベイズ分類器は新たなサンプルx　の$\hat{\mathrm{y}}$クラスを予測する。
- すべてのクラスc_iについて、条件付確率$P(c_i|\mathrm{x})$を見積もり、事後確率最大の値を選ぶ。

$$\hat{y} = \arg\max_{c_i}\{P(c_i|\mathbf{x})\}$$

1. ベイズ推定

- 「結果から原因を求める推定方法」
- 「ある結果のときにある原因である確率が最も大きいような原因を選ぶ」

「発熱の原因」を確率的に推定する場合：　事前に分かっている確率

P(風邪|発熱) = ?

P(エボラ|発熱) = ?

P(赤痢|発熱) = ?

P(発熱|風邪) = 0.8

P(発熱|エボラ) = 0.7

P(発熱|赤痢) = 0.9

発熱したときに{風邪／エボラ／赤痢}である確率。
この確率が大きいものが原因と推定出来る。

2. ベイズの定理

ベイズの定理

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A)} = \frac{P(A|B_i)P(B_i)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}$$

ベイズの定理（簡略版）

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

「Aが起きた時にBが起きる確率」＝ $\dfrac{\text{「Bが起きた時にAが起きる確率」×「Bが起きる確率」}}{\text{「事象Aが起きる確率」}}$

3. ベイズの定理の導出(1)

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$\Downarrow$$

$$P(A \cap B) = P(B|A)P(A) = P(A|B)P(B)$$

$$\Downarrow$$

ベイズの定理(簡略版)

$$\boxed{P(B|A) = \frac{P(A|B)P(B)}{P(A)}}$$

3. ベイズの定理の導出（2）

$$P(A)=\sum_{i=1}^{n} P(A|B_i)P(B_i)$$

$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2)$

※B_1, B_2 … B_n = Aを引き起こす全ての原因の集合
nが無限個の場合も定理が成り立つ。
（発熱の全ての原因は？）

ベイズの定理

$$P(B_i|A)=\frac{P(A|B_i)P(B_i)}{P(A)}=\frac{P(A|B_i)P(B_i)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}$$

4. ベイズ推定～例題～

A = 発熱する
B_1 = 風邪になる
B_2 = エボラになる
B_3 = 赤痢になる

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A)}$$

分子の大小により比較
（「発熱」の確率=P(A)が分からなくても推定可）

$P(B_1) = 0.1$
$P(B_2) = 0.01$
$P(B_3) = 0.01$
$P(A|B_1) = 0.8$
$P(A|B_2) = 0.7$
$P(A|B_3) = 0.9$

$P(B_1|A) = P(A|B_1)\ P(B_1)/\ P(A)= 0.08/\ P(A)$
$P(B_2|A) = P(A|B_2)\ P(B_2)\ /\ P(A)= 0.007/\ P(A)$
$P(B_3|A) = P(A|B_3)\ P(B_3)\ /\ P(A)= 0.009/\ P(A)$

「BiのときにAが起きる確率」×「Biが起きる確率」

ベイズ定理

- ベイズ定理による、事後確率を尤度確率$P(\mathrm{x}|c_i)$と事前確率$P(c_i)$で変換する。

$$P(c_i|\mathbf{x}) = \frac{P(\mathbf{x}|c_i) \cdot P(c_i)}{P(\mathbf{x})}$$

$$P(\mathbf{x}) = \sum_{j=1}^{k} P(\mathbf{x}|c_j) \cdot P(c_j)$$

ベイズ分類器

- 事後最大確率は以下のの様に書き換える。

$$\hat{y} = \arg\max_{c_i}\{P(c_i|\mathbf{x})\}$$

$$= \arg\max_{c_i}\left\{\frac{P(\mathbf{x}|c_i)P(c_i)}{P(\mathbf{x})}\right\} = \arg\max_{c_i}\{P(\mathbf{x}|c_i)P(c_i)\}$$

ベイズ分類器

- 尤度確率$P(\mathrm{x}|c_i)$の計算
  - 訓練データから尤度確率と事前確率を求める
  - D_iはクラスc_iのみのサンプル

$$\mathbf{D}_i = \{\mathbf{x}_j \in \mathbf{D} \mid \mathbf{x}_j \text{ has class } y_j = c_i\}$$

  - $n = |D|, n_i = |D_i|$の時、事前確率は

$$\hat{P}(c_i) = \frac{n_i}{n}$$

  - 尤度確率：　$P(\mathbf{x}|c_i) = 2\epsilon \cdot f_i(\mathbf{x})$

$$f_i(\mathbf{x}) = f(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{(\sqrt{2\pi})^d \sqrt{|\boldsymbol{\Sigma}_i|}} \exp\left\{-\frac{(\mathbf{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\mathbf{x}-\boldsymbol{\mu}_i)}{2}\right\}$$

ベイズ分類器

- 事後確率$P(c_i|\mathrm{x})$の計算
  - 量的属性の場合：

$$P(c_i|\mathbf{x}) = \frac{P(\mathbf{x}|c_i) \cdot P(c_i)}{P(\mathbf{x})}$$

$$P(c_i|\mathbf{x}) = \frac{2\epsilon \cdot f_i(\mathbf{x}) P(c_i)}{\sum_{j=1}^{k} 2\epsilon \cdot f_j(\mathbf{x}) P(c_j)} = \frac{f_i(\mathbf{x}) P(c_i)}{\sum_{j=1}^{k} f_j(\mathbf{x}) P(c_j)}$$

- 事後最大確率の値： $\hat{y} = \arg\max_{c_i} \left\{ f_i(\mathbf{x}) P(c_i) \right\}$

ベイズ分類器：アルゴリズム

- 量的属性の場合：サンプルc_iの平均値：$\hat{\mu}_i = \frac{1}{n_i} \sum_{\mathbf{x}_j \in \mathbf{D}_i} \mathbf{x}_j$

サンプルc_iの共分散：$\widehat{\boldsymbol{\Sigma}}_i = \frac{1}{n_i} \mathbf{Z}_i^T \mathbf{Z}_i$ $\quad \mathbf{Z}_i = \mathbf{D}_i - \mathbf{1} \cdot \hat{\mu}_i^T$.

BayesClassifier ($\mathbf{D} = \{(\mathbf{x}_j, y_j)\}_{j=1}^n$):

for $i = 1, \ldots, k$ **do**
- $\mathbf{D}_i \leftarrow \{\mathbf{x}_j \mid y_j = c_i, j = 1, \ldots, n\}$ `// class-specific subsets`
- $n_i \leftarrow |\mathbf{D}_i|$ `// cardinality`
- $\hat{P}(c_i) \leftarrow n_i / n$ `// prior probability`
- $\hat{\mu}_i \leftarrow \frac{1}{n_i} \sum_{\mathbf{x}_j \in \mathbf{D}_i} \mathbf{x}_j$ `// mean`
- $\mathbf{Z}_i \leftarrow \mathbf{D}_i - \mathbf{1}_{n_i} \hat{\mu}_i^T$ `// centered data`
- $\widehat{\boldsymbol{\Sigma}}_i \leftarrow \frac{1}{n_i} \mathbf{Z}_i^T \mathbf{Z}_i$ `// covariance matrix`

return $\hat{P}(c_i), \hat{\mu}_i, \widehat{\boldsymbol{\Sigma}}_i$ *for all* $i = 1, \ldots, k$

Testing ($\mathbf{x}$ **and** $\hat{P}(c_i)$, $\hat{\mu}_i$, $\widehat{\boldsymbol{\Sigma}}_i$, **for all** $i \in [1, k]$):

$\hat{y} \leftarrow \arg\max_{c_i} \{ f(\mathbf{x} | \hat{\mu}_i, \widehat{\boldsymbol{\Sigma}}_i) \cdot P(c_i) \}$

return $\hat{y}$

Irisデータセットの例

- 2次元のデータセット、属性Iris-sepal-lengthとIris-sepal-width、クラスc_1はIris-Setosa、

  クラスc_2はその他。
- $n_1 = 50, n_2 = 100$

$$\hat{P}(c_1) = \frac{n_1}{n} = \frac{50}{150} = 0.33 \qquad \hat{P}(c_2) = \frac{n_2}{n} = \frac{100}{150} = 0.67$$

- $$\hat{\mu}_1 = \begin{pmatrix} 5.01 \\ 3.42 \end{pmatrix} \qquad \hat{\mu}_2 = \begin{pmatrix} 6.26 \\ 2.87 \end{pmatrix}$$

$$\widehat{\boldsymbol{\Sigma}}_1 = \begin{pmatrix} 0.122 & 0.098 \\ 0.098 & 0.142 \end{pmatrix} \qquad \widehat{\boldsymbol{\Sigma}}_2 = \begin{pmatrix} 0.435 & 0.121 \\ 0.121 & 0.110 \end{pmatrix}$$

- $\mathbf{x} = (6.75, 4.25)^T$ を与える時、クラスを予測する

Irisデータセットの例

- 2次元のデータセット、属性Iris-sepal-lengthとIris-sepal-width

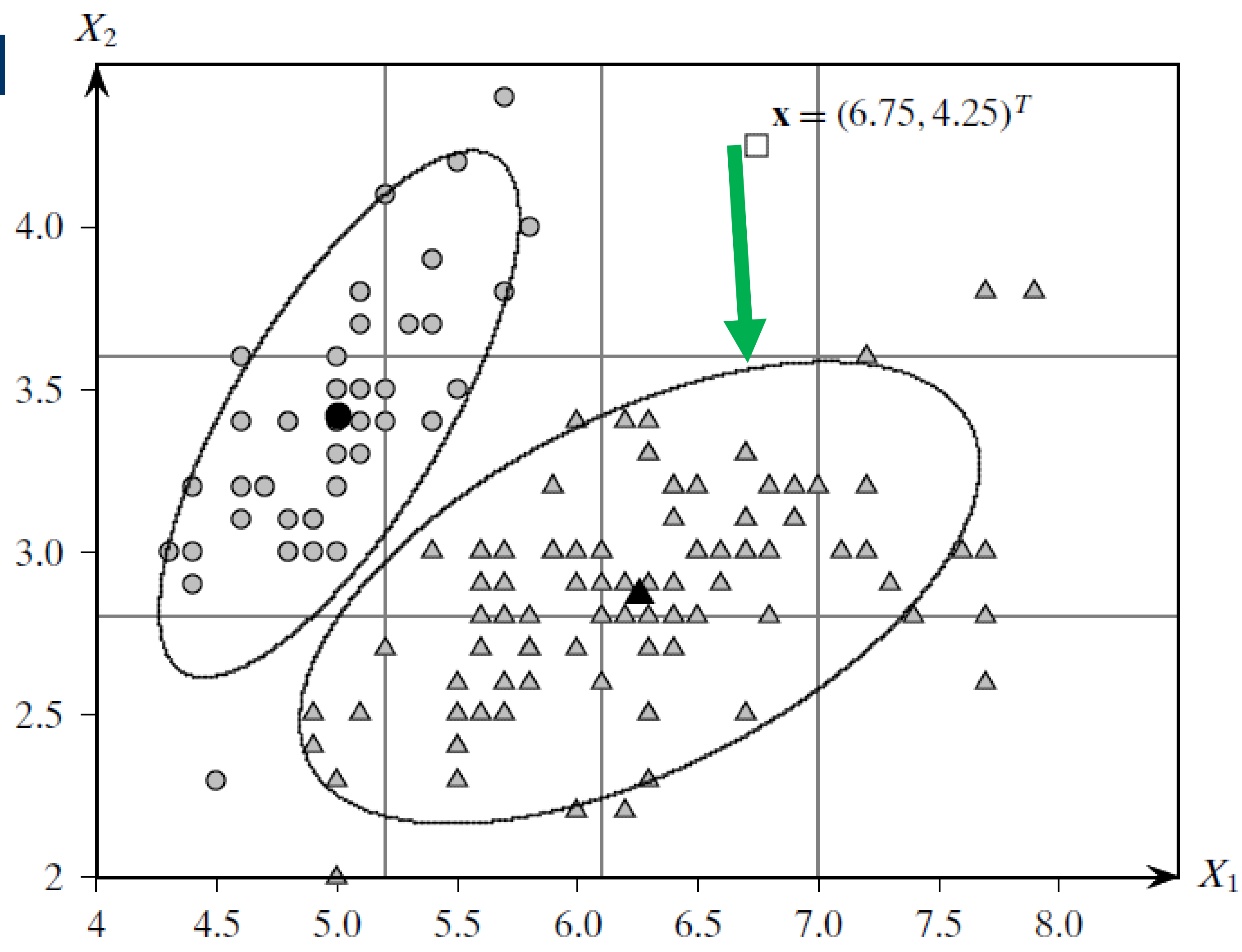
$$\hat{P}(c_1|\mathbf{x}) \propto \hat{f}(\mathbf{x}|\hat{\mu}_1, \widehat{\Sigma}_1)\hat{P}(c_1) = (4.914 \times 10^{-7}) \times 0.33 = 1.622 \times 10^{-7}$$

$$\hat{P}(c_2|\mathbf{x}) \propto \hat{f}(\mathbf{x}|\hat{\mu}_2, \widehat{\Sigma}_2)\hat{P}(c_2) = (2.589 \times 10^{-5}) \times 0.67 = 1.735 \times 10^{-5}$$

- 結論：$\hat{y} = c_2$

Irisデータセットの例

- 2次元のデータセット、X1=Iris-sepal-lengthとX2=Iris-sepal-width



単純ベイズ分類器

- 単純ベイズ分類器の元となる確率モデルは強い（単純な）独立性仮定と共にベイズの定理を適用することに基づいており、より正確に言えば「独立特徴モデル; independent feature model」と呼ぶべきものである。

$$P(\mathbf{x}|c_i) = P(x_1, x_2, \ldots, x_d|c_i) = \prod_{j=1}^{d} P(x_j|c_i)$$

単純ベイズ分類器

- 単純ベイズ分類器

$$P(x_j|c_i) \propto f(x_j|\mu_{ij},\sigma_{ij}^2) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} \exp\left\{-\frac{(x_j-\mu_{ij})^2}{2\sigma_{ij}^2}\right\}$$

$$(\mathbf{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) = \sum_{j=1}^{d}\frac{(x_j-\mu_{ij})^2}{\sigma_{ij}^2}$$

$$\boldsymbol{\Sigma}_i^{-1} = \begin{pmatrix} \frac{1}{\sigma_{i1}^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_{i2}^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \cdots & \frac{1}{\sigma_{id}^2} \end{pmatrix}$$

単純ベイズ分類器:アルゴリズム

NaiveBayes $(\mathbf{D} = \{(\mathbf{x}_j, y_j)\}_{j=1}^{n})$:

for $i = 1, \ldots, k$ **do**

  $\mathbf{D}_i \leftarrow \{\mathbf{x}_j \mid y_j = c_i, j = 1, \ldots, n\}$ // class-specific subsets

  $n_i \leftarrow |\mathbf{D}_i|$ // cardinality

  $\hat{P}(c_i) \leftarrow n_i / n$ // prior probability

  $\hat{\mu}_i \leftarrow \frac{1}{n_i} \sum_{\mathbf{x}_j \in \mathbf{D}_i} \mathbf{x}_j$ // mean

  $\mathbf{Z}_i = \mathbf{D}_i - \mathbf{1} \cdot \hat{\mu}_i^T$ // centered data for class c_i

  for $j = 1, .., d$ **do** // class-specific variance for X_j

    $\hat{\sigma}_{ij}^2 \leftarrow \frac{1}{n_i} Z_{ij}^T Z_{ij}$ // variance

  $\hat{\sigma}_i = \left(\hat{\sigma}_{i1}^2, \ldots, \hat{\sigma}_{id}^2\right)^T$ // class-specific attribute variances

return $\hat{P}(c_i), \hat{\mu}_i, \hat{\sigma}_i$ for all $i = 1, \ldots, k$

Testing ($\mathbf{x}$ **and** $\hat{P}(c_i)$, $\hat{\mu}_i$, $\hat{\sigma}_i$, **for all** $i \in [1, k]$):

$$\hat{y} \leftarrow \arg\max_{c_i} \left\{ \hat{P}(c_i) \prod_{j=1}^{d} f(x_j | \hat{\mu}_{ij}, \hat{\sigma}_{ij}^2) \right\}$$

return $\hat{y}$

単純ベイズ分類器：Irisデータセットの例

- 2次元のデータセット、属性Iris-sepal-lengthとIris-sepal-width

$$\hat{\mathbf{\Sigma}}_1 = \begin{pmatrix} 0.122 & 0 \\ 0 & 0.142 \end{pmatrix} \qquad \hat{\mathbf{\Sigma}}_2 = \begin{pmatrix} 0.435 & 0 \\ 0 & 0.110 \end{pmatrix}$$

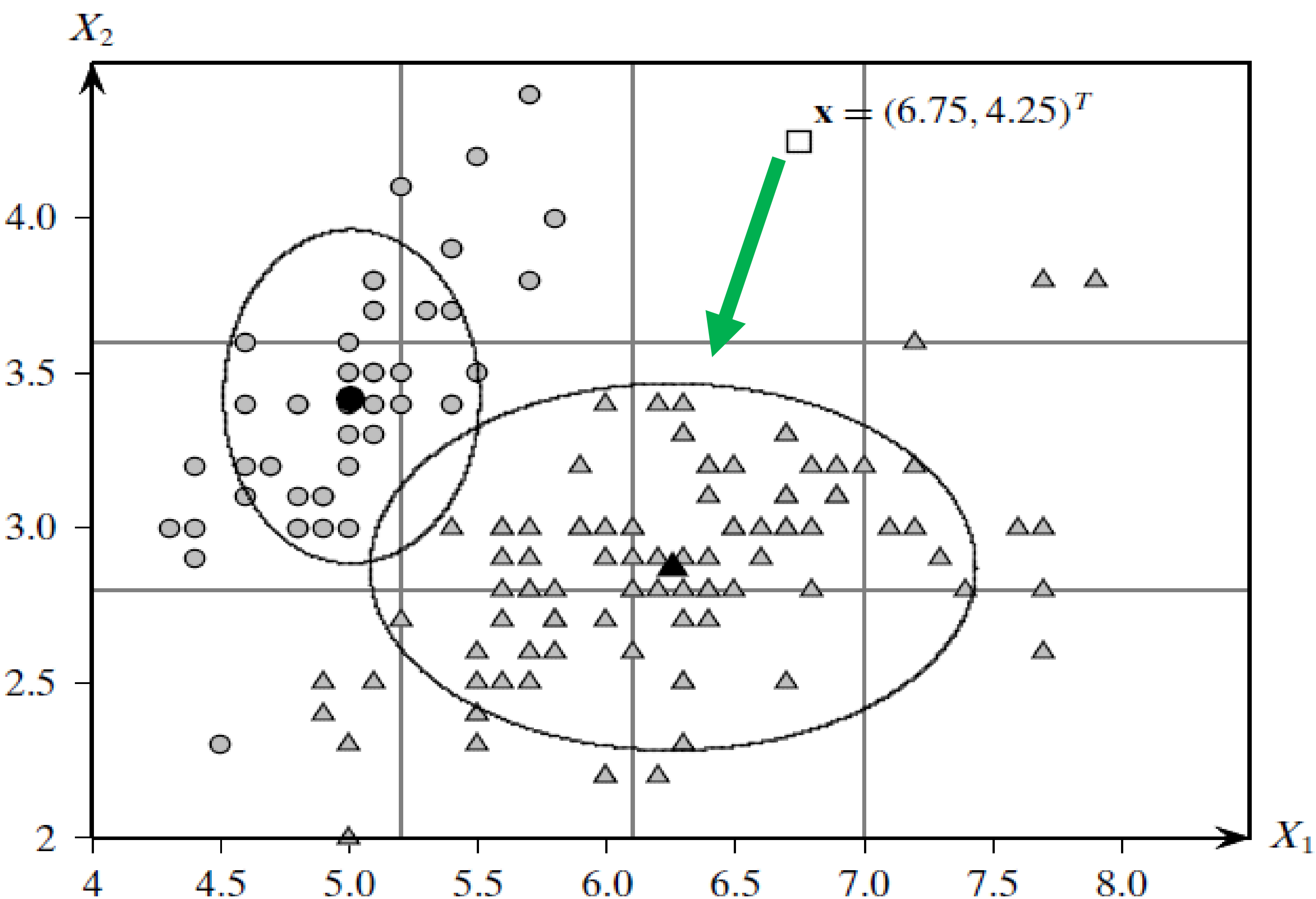
$$\hat{P}(c_1|\mathbf{x}) \propto \hat{f}(\mathbf{x}|\hat{\mu}_1, \hat{\mathbf{\Sigma}}_1)\hat{P}(c_1) = (3.99 \times 10^{-7}) \times 0.33 = 1.32 \times 10^{-7}$$

$$\hat{P}(c_2|\mathbf{x}) \propto \hat{f}(\mathbf{x}|\hat{\mu}_2, \hat{\mathbf{\Sigma}}_2)\hat{P}(c_2) = (9.597 \times 10^{-5}) \times 0.67 = 6.43 \times 10^{-5}$$

- 結論：$\hat{y} = c_2$

単純ベイズ分類器：Irisデータセットの例

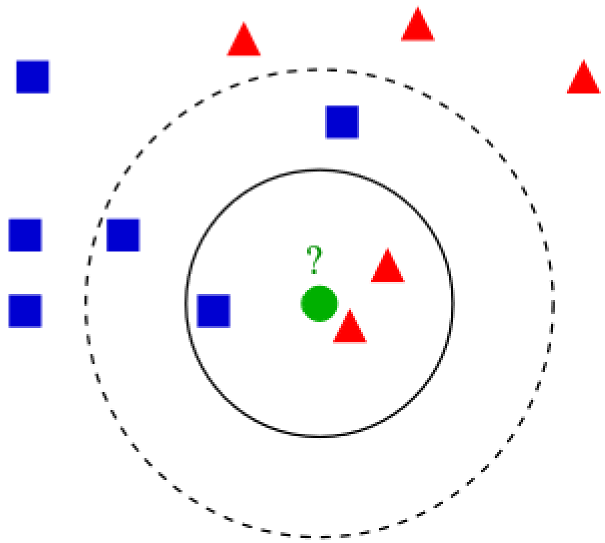
- 2次元のデータセット、X1=Iris-sepal-lengthとX2=Iris-sepal-width



k近傍法 （kNN：k-Nearest Neighbors）

- k近傍法は、インスタンスに基づく学習の一種であり、怠惰(たいだ)学習 (lazy learning) の一種である。

- 近傍のオブジェクト群の投票によって決定される



k近傍法の例。標本（緑の丸）は、
第一のクラス（青の四角）と第二のクラス（赤の三角）のいずれかに分類される。
k = 3 なら、内側の円内にあるオブジェクトが近傍となるので、第二のクラスに分類される（赤の三角の方が多い）。しかし、k = 5 なら、それが逆転する

ソース：ウイキペディア

k近傍法

- 定式化:
  - 訓練データから尤度確率と事前確率を求める
  - D_iはクラスc_iのみのサンプル

$$\mathbf{D}_i = \{\mathbf{x}_j \in \mathbf{D} \mid \mathbf{x}_j \text{ has class } y_j = c_i\}$$

  - $n = |D|, n_i = |D_i|$
  - K:近傍の数、r:距離
  - 内側の円 $B_d(\mathbf{x}, r) = \left\{\mathbf{x}_i \in \mathbf{D} \mid \delta(\mathbf{x}, \mathbf{x}_i) \le r\right\}$
  - 内側の円内にあるオブジェクト

$$K_i = \left\{\mathbf{x}_j \in B_d(\mathbf{x}, r) \mid y_j = c_i\right\}$$

k近傍法

- 確率$P(\mathrm{x}|c_i)$の計算

$$\hat{f}(\mathbf{x}|c_i) = \frac{K_i/n_i}{V} = \frac{K_i}{n_i V}$$

- 事後確率$P(c_i|\mathrm{x})$の計算

$$P(c_i|\mathbf{x}) = \frac{\hat{f}(\mathbf{x}|c_i)\hat{P}(c_i)}{\sum_{j=1}^{k} \hat{f}(\mathbf{x}|c_j)\hat{P}(c_j)}$$

$$\hat{P}(c_i) = \frac{n_i}{n}$$

$$P(c_i|\mathbf{x}) = \frac{\frac{K_i}{nV}}{\sum_{j=1}^{k} \frac{K_j}{nV}} = \frac{K_i}{K}$$

k近傍法

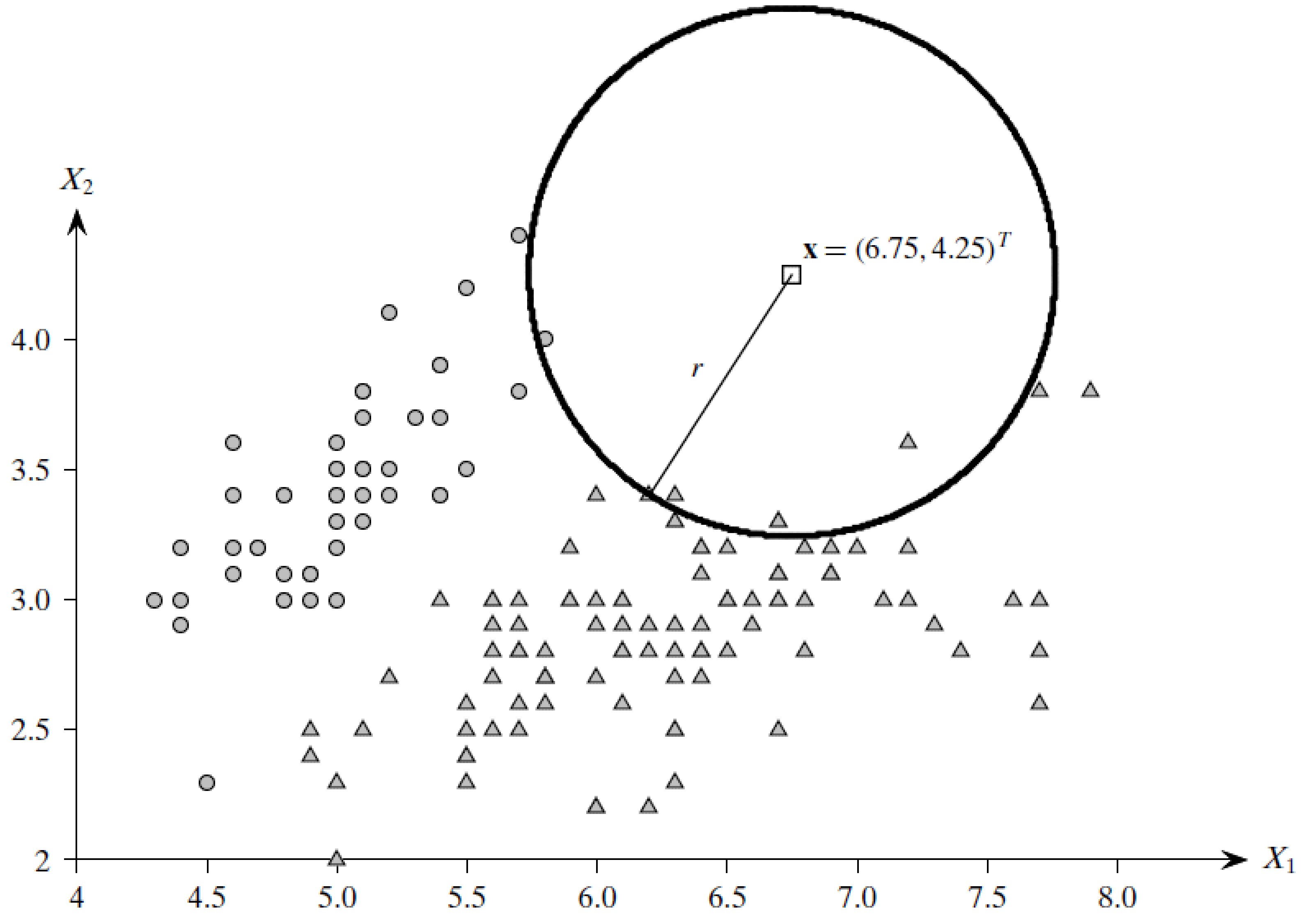
- 事後最大確率の計算

$$\hat{y} = \arg\max_{c_i} \{P(c_i | \mathbf{x})\} = \arg\max_{c_i} \left\{ \frac{K_i}{K} \right\} = \arg\max_{c_i} \{K_i\}$$

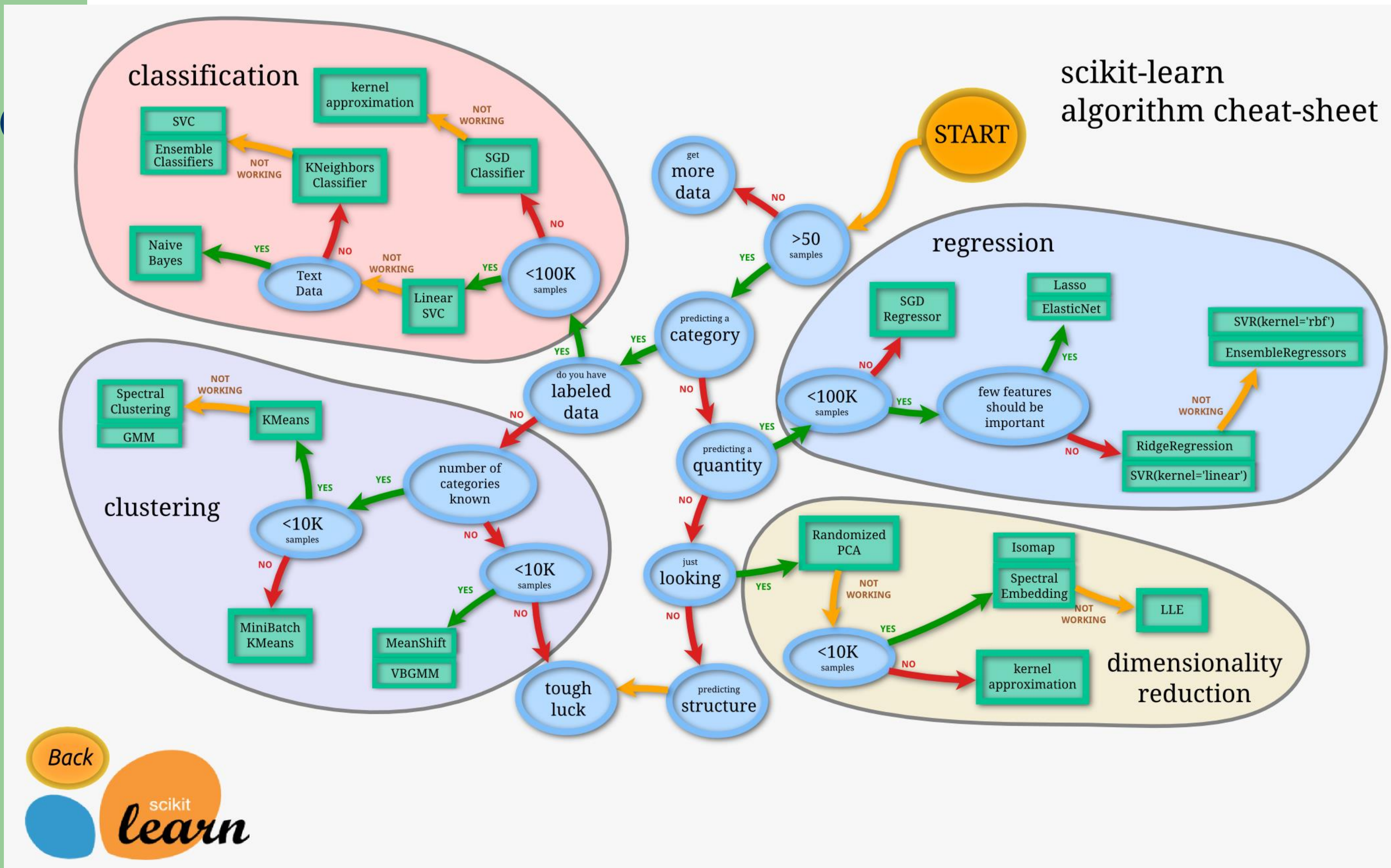
Irisデータセットの例

- 2次元のデータセット、属性Iris-sepal-lengthとIris-sepal-width、クラスc_1はIris-Setosa、
  クラスc_2はその他。
- $n_1 = 50, n_2 = 100$
- $\mathbf{x} = (6.75, 4.25)^T$ を与える時、クラスを予測する
- K=5
- $r = \sqrt{1.025} = 1.012$
- 結論：$\hat{y} = c_2$

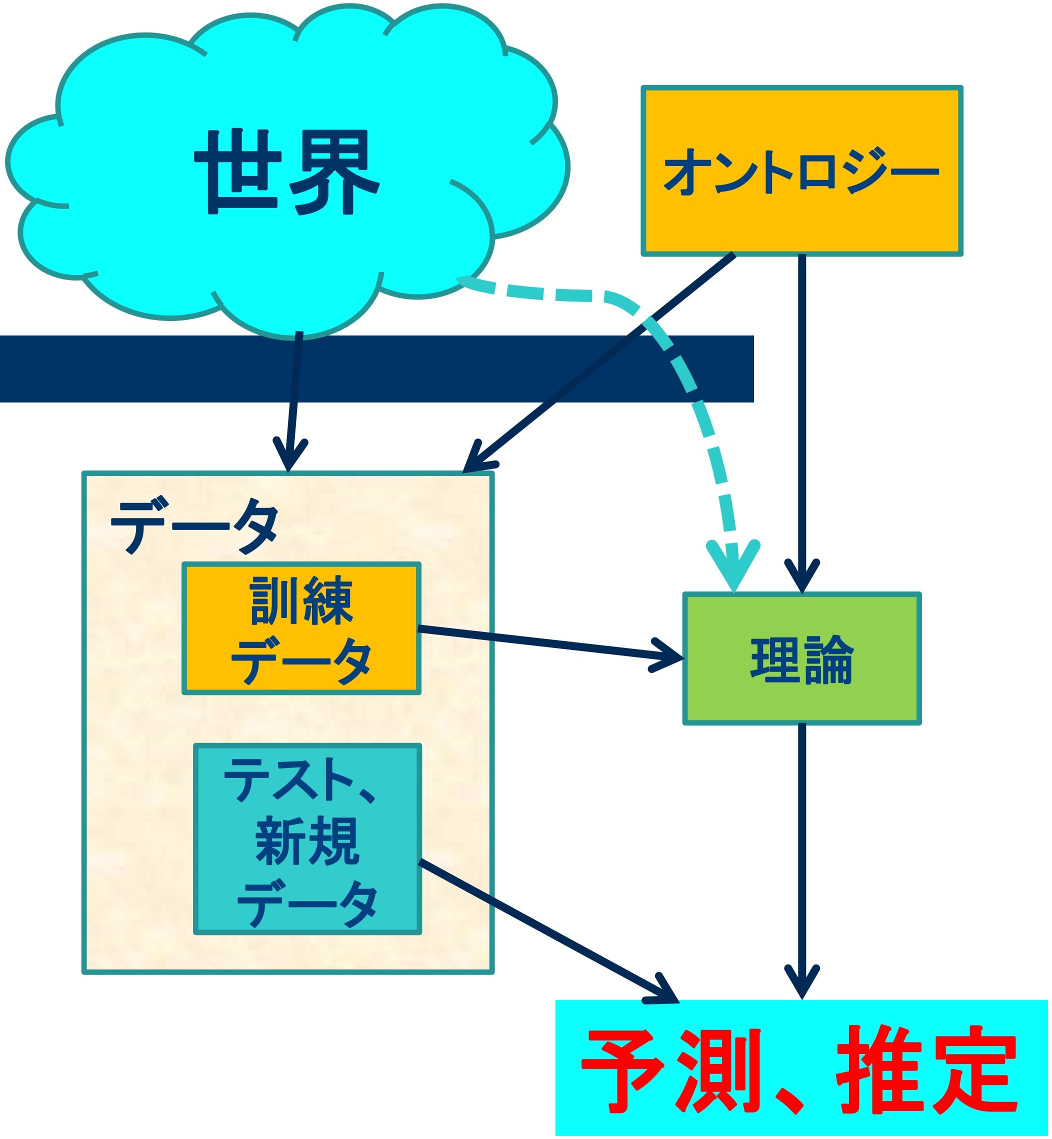
k近傍法: Irisデータセット、K=5



分類器/クラスタ分析器/回帰など



まとめ



- ベイズ分類器
- 単純ベイズ分類器
- K近傍法
- 例題

参考文献

- Data Mining (Mohamed Zaki and Wagner Meira)

HOW NAIVE BAYES CLASSIFIER WORKS?

Whether	Temperature	Play
Sunny	Hot	No
Sunny	Hot	No
Overcast	Hot	Yes
Rainy	Mild	Yes
Rainy	Cool	Yes
Rainy	Cool	No
Overcast	Cool	Yes
Sunny	Mild	No
Sunny	Cool	Yes
Rainy	Mild	Yes
Sunny	Mild	Yes
Overcast	Mild	Yes
Overcast	Hot	Yes
Rainy	Mild	No

01 CALCULATE PRIOR PROBABILITY FOR GIVEN CLASS LABELS

02 CALCULATE CONDITIONAL PROBABILITY WITH EACH ATTRIBUTE FOR EACH CLASS

03 MULTIPLY SAME CLASS CONDITIONAL PROBABILITY.

04 MULTIPLY PRIOR PROBABILITY WITH STEP 3 PROBABILITY.

05 SEE WHICH CLASS HAS HIGHER PROBABILITY, HIGHER PROBABILITY CLASS BELONGS TO GIVEN INPUT SET STEP.

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9

Likelihood table				
Weather	No	Yes		
Overcast		4	=4/14	0.29
Rainy	3	2	=5/14	0.36
Sunny	2	3	=5/14	0.36
All	5	9		
	=5/14	=9/14		
	0.36	0.64		

https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/