Naivní Bayesův klasifikátor

- klasifikace mnohorozměrných veličin za předpokladu podmíněné nezávislosti

Příklad:

$y_t = \begin{bmatrix} y_{1;t} \\ y_{2;t} \end{bmatrix}$ – dvourozměrná, $c_t \in \{1, 2, \ldots, n_c\}$ – ukazovátko

Model:

$$f(y_{1;t}, y_{2;t}|\Theta, c_t) = \prod_{i=1}^{N=2} f(y_{i;t}|\Theta_i, c_t)$$

pro podmíněně nezávislé platí:
$f(y_1, y_2|c) = f(y_1|\cancel{y_2}, c)f(y_2|c) =$

$= f(y_1|c)f(y_2|c)$

- jednotlivé veličiny $y_{i;t}$ se modelují zvlášť
- významné zjednodušení – pouze skalární modely

Princip naivního Bayesova klasifikátoru:

$$f(c_t|y_t) = f(c_t) \prod_{i=1}^{N=2} f(y_{i;t}|\Theta_i, c_t)$$

Bayesovo pravidlo:
$f(c, y) = f(c|y)f(y) = f(y|c)f(c)$
$f(c|y) \propto f(y|c)f(c)$

$f(c|y_1, y_2) \propto f(y_1|c)f(y_2|c)f(c)$

Příklad - klasifikace výběru dopravního prostředku

c_t – dopravní prostředek při cestě do práce (0 - auto, 1 - MHD, 2 - kolo)

$y_{1;t}$ – doba cesty do práce (minuty)
$y_{2;t}$ – náklady na cestu (Kč/km)
$y_{3;t}$ – příjem (Kč/měsíc)
$y_{4;t}$ – věk (roky)
$y_{5;t}$ – vzdálenost od zastávky MHD (m)
$y_{6;t}$ – dostupnost cyklostezek (1 - ano, 0 - ne)
$y_{7;t}$ – ekologické ohledy (1 - ano, 0 - ne)

$$y_t = \begin{bmatrix} y_{1;t} \\ y_{2;t} \\ \dots \\ \\ y_{7;t} \end{bmatrix}$$

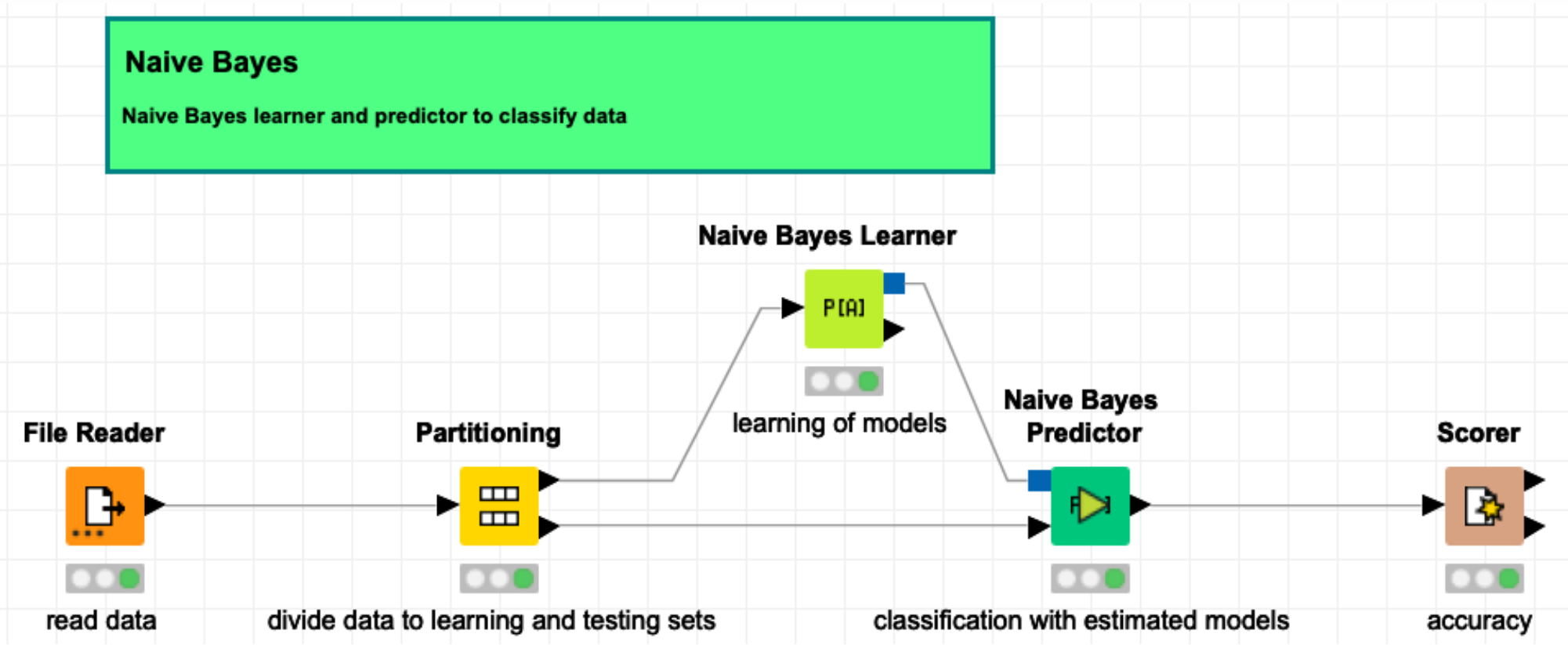
Naivní Bayes – postup:

1. Fáze trénování modelů – trénovací data (s učitelem – měřené ukazovátko)
   - odhad parametrů skalárních komponent (filtrace každé veličiny + offline odhad)

2. Fáze testování (klasifikace) – testovací data (bez ukazovátka)
   - odhad ukazovátka z odhadnutých skalárních komponent

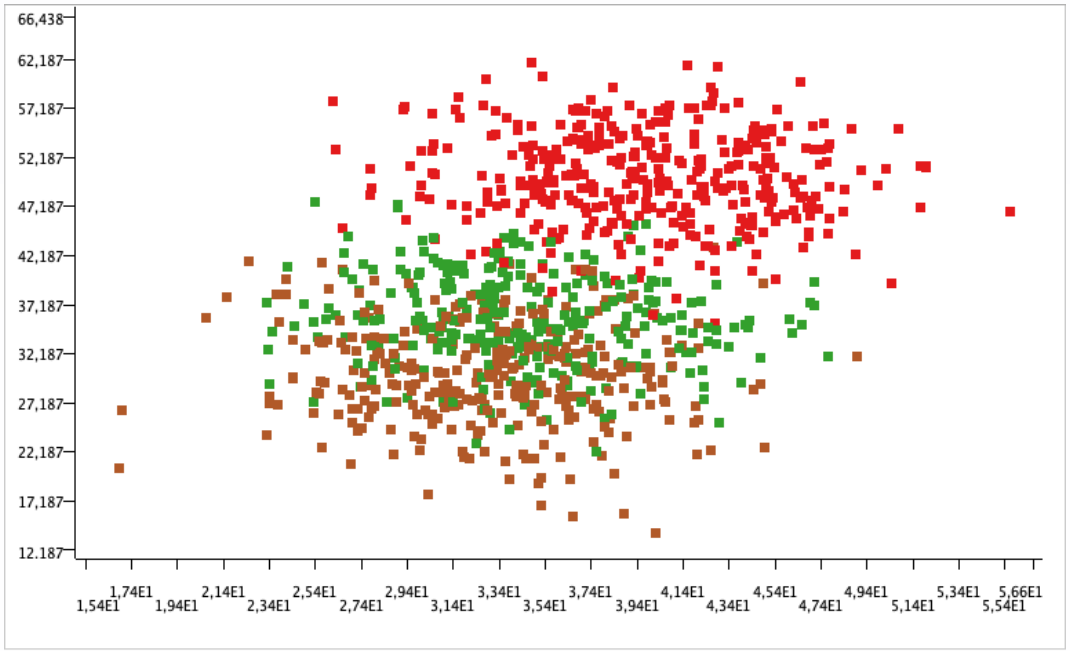
$$f(c_t|y_t) = \underbrace{\text{normovaný histogram ukazovátka}}_{f(c_t)} \times \underbrace{\text{součin komponent}}_{\prod_{i=1}^{N=7} f(y_{i;t}|\hat{\Theta}_i, c_t)}$$

Program v KNIME



Výsledky klasifikace s naivním Bayesem

Nalezené shluky: doba cesty vs věk



Chybová matice

Column7 \...	0	1	2
0	282	14	9
1	14	351	18
2	4	13	295

Correct classified: 928 Wrong classified: 72

Accuracy: 92,8% Error: 7,2%

Cohen's kappa (κ): 0,891%

Table "default" – Rows: 4 | Spec – Columns: 11 | Properties | Flow Variables

Columns: 11	Column Type	Column In...	Color Han...	Size Han...	Shape Ha...	Filter Han...	Lower Bo...	Upper Bo...
TruePositives	Number (integer)	0					282	351
FalsePositives	Number (integer)	1					18	27
TrueNegatives	Number (integer)	2					590	677
FalseNegatives	Number (integer)	3					17	32
Recall	Number (double)	4					0.916	0.946
Precision	Number (double)	5					0.916	0.94
Sensitivity	Number (double)	6					0.916	0.946
Specificity	Number (double)	7					0.956	0.974
F-measure	Number (double)	8					0.922	0.932
Accuracy	Number (double)	9					0.928	0.928
Cohen's kappa	Number (double)	10					0.891	0.891

Přesnost klasifikace

Příklad: binární klasifikace (pozitivní, negativní)

- Chybová matice (confusion matrix)

	Skutečně pozitivní	Skutečně negativní
Pozitivní predikce	TP	FP
Negativní predikce	FN	TN

- Přesnost (accuracy)

Procento správných predikcí ze všech dat

$$Acc = \frac{TP + TN}{TP + FP + FN + TN}$$

- Preciznost (precision)

Procento správných pozitivních predikcí mezi pozitivními predikcemi

$$Pre = \frac{TP}{TP + FP}$$

- Výtěžnost (recall)

Procento správných pozitivních predikcí mezi skutečně pozitivními

$$Rec = \frac{TP}{TP + FN}$$

- Skóre F1 $\in (0, 1)$

$$F1 = 2\frac{Pre \times Rec}{Pre + Rec}$$

Logistická regrese (binární - $c_t \in \{0, 1\}$)

Obecný tvar pro $c_t \in \{0, 1\}$

$$\frac{\exp\{c_t z_t\}}{1 + \exp\{z_t\}},$$

$$\underbrace{\frac{\exp\{z_t\}}{1 + \exp\{z_t\}}}_{P(c_t=1)} \qquad \underbrace{\frac{1}{1 + \exp\{z_t\}}}_{P(c_t=0)}$$

$$z_t = \underbrace{y_{1;t}\theta_1 + y_{2;t}\theta_2 + \ldots + k}_{\text{lineární regrese}}$$

$$y_t = \underbrace{[y_{1;t} \quad y_{2;t} \quad \cdots \quad y_{n;t}]}_{\text{data}}$$

Příklad: úsek silnice

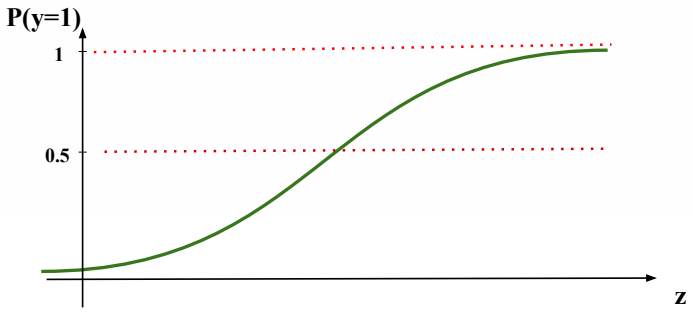
$c_t \in \{0, 1\}$ – nehoda (lehká, těžká)

$y_{1;t}$ – rychlost 1.auta (spojitá)

$y_{2;t}$ – rychlost 2.auta (spojitá)

$y_{3;t}$ – povrch vozovky (diskrétní)

$y_{4;t}$ – zkušenost řidiče (diskrétní)



$$z_t \in \mathbb{R} \Rightarrow P \in (0, 1)$$

Odhad parametrů logistické regrese

Metoda maximální věrohodnosti

Věrohodnostní funkce

$$L_t(\theta) = \prod_{\tau=1}^{t} \frac{e^{c_\tau z_\tau}}{1 + e^{z_\tau}}$$

$$\hat{\theta} = \arg \max_{\theta} \ln L_t(\theta)$$

Výhody – známé derivace $\ln L_t(\theta)$

max – Newtonova metoda hledání extrémů

$$x_{i+1} = x_i - \frac{f'(x_i)}{f''(x_i)}$$

Odkaz na web

Multinomiální logistická regrese – $c_t \in \{0, 1, 2, \cdots, N\}$

Inverzní forma logistické regrese:

$$\text{logit}\,(p) = \ln \frac{p}{1-p}, \quad p = \frac{\exp\{c_t z_t\}}{1 + \exp\{z_t\}}$$

$$\ln \frac{p_1}{p_0} = z_{t(1)}, \quad \ln \frac{p_2}{p_0} = z_{t(2)}, \quad \cdots \quad , \ln \frac{p_N}{p_0} = z_{t(N)}$$

Příklad - klasifikace výběru dopravního prostředku

c_t – dopravní prostředek při cestě do práce (0 - auto, 1 - MHD, 2 - kolo)

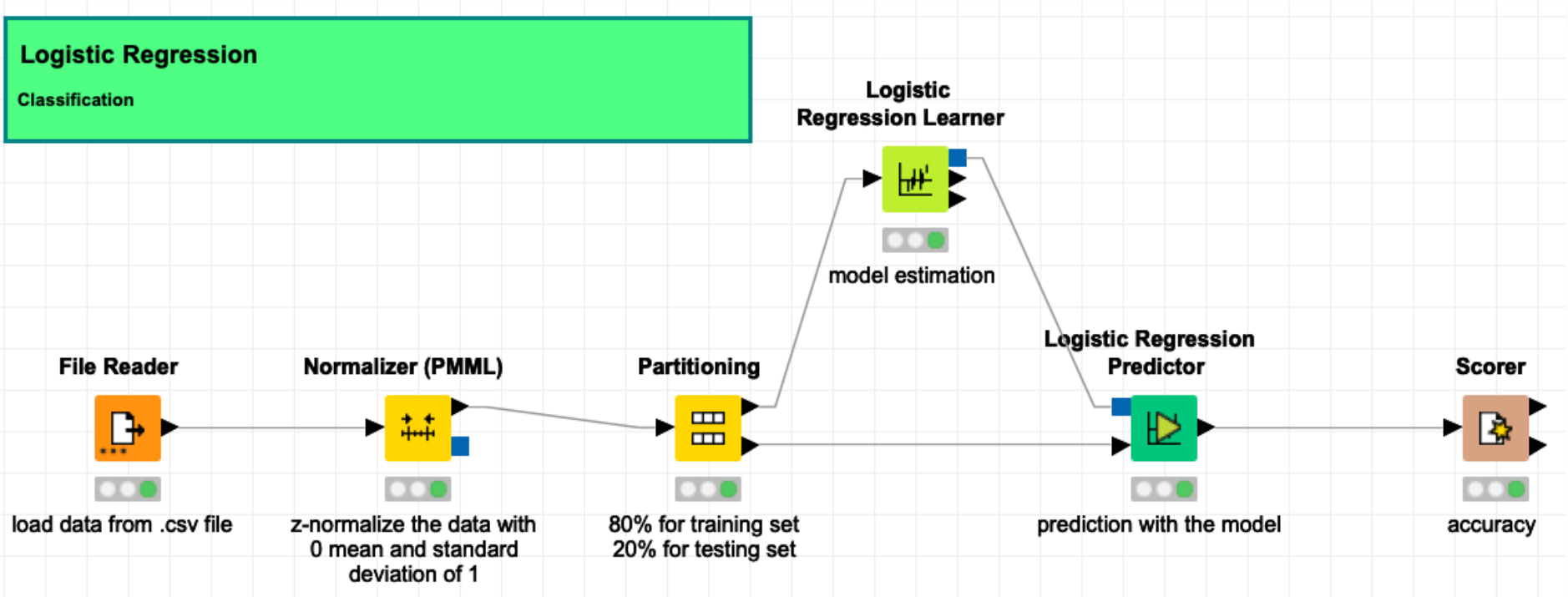
$y_{1;t}$ – doba cesty do práce (minuty)
$y_{2;t}$ – náklady na cestu (Kč/km)
$y_{3;t}$ – příjem (Kč/měsíc)
$y_{4;t}$ – věk (roky)
$y_{5;t}$ – vzdálenost od zastávky MHD (m)
$y_{6;t}$ – dostupnost cyklostezek (1 - ano, 0 - ne)
$y_{7;t}$ – ekologické ohledy (1 - ano, 0 - ne)

$$y_t = \begin{bmatrix} y_{1;t} \\ y_{2;t} \\ \dots \\ \\ y_{7;t} \end{bmatrix}$$

Logistická regrese – postup:

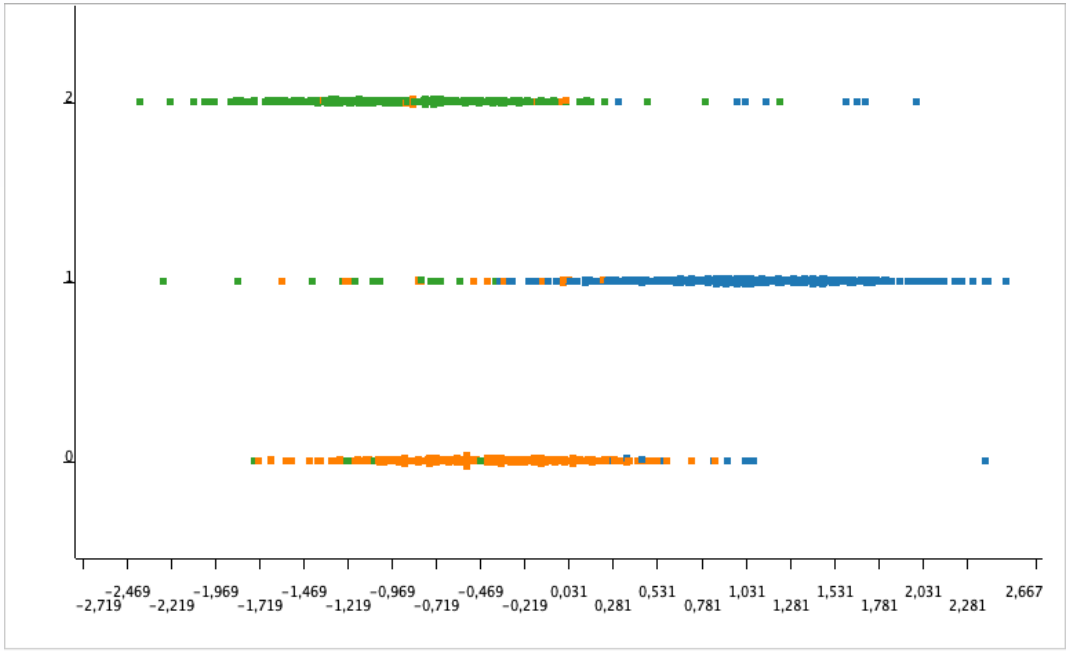
1. Fáze trénování modelů – trénovací data (s učitelem – měřené ukazovátko)
   - odhad regresních koeficientů
2. Fáze testování (klasifikace) – testovací data (bez ukazovátka)
   - odhad ukazovátka

Program v KNIME



Výsledky klasifikace s logistickou regresí

Doba cesty vs dopravní prostředek (norm.data)



Default Settings | Column Selection | Appearance

X Column: D Column0 | Y Column: S Column7

X attribute	-2,719	2,667
Y attribute	-2,507	2,479

Chybová matice

Column7 \...	0	1	2
0	284	10	5
1	16	374	15
2	7	8	281

Correct classified: 939 | Wrong classified: 61

Accuracy: 93,9% | Error: 6,1%

Cohen's kappa (κ): 0,908%

Table "default" – Rows: 4 | Spec – Columns: 11 | Properties | Flow Variables

Columns: 11	Column Type	Column In...	Color Han...	Size Han...	Shape Ha...	Filter Han...	Lower Bo...	Upper Bo...
TruePositives	Number (integer)	0					281	374
FalsePositives	Number (integer)	1					18	23
TrueNegatives	Number (integer)	2					577	684
FalseNegatives	Number (integer)	3					15	31
Recall	Number (double)	4					0.923	0.95
Precision	Number (double)	5					0.925	0.954
Sensitivity	Number (double)	6					0.923	0.95
Specificity	Number (double)	7					0.967	0.972
F-measure	Number (double)	8					0.937	0.941
Accuracy	Number (double)	9					0.939	0.939
Cohen's kappa	Number (double)	10					0.908	0.908