IPML

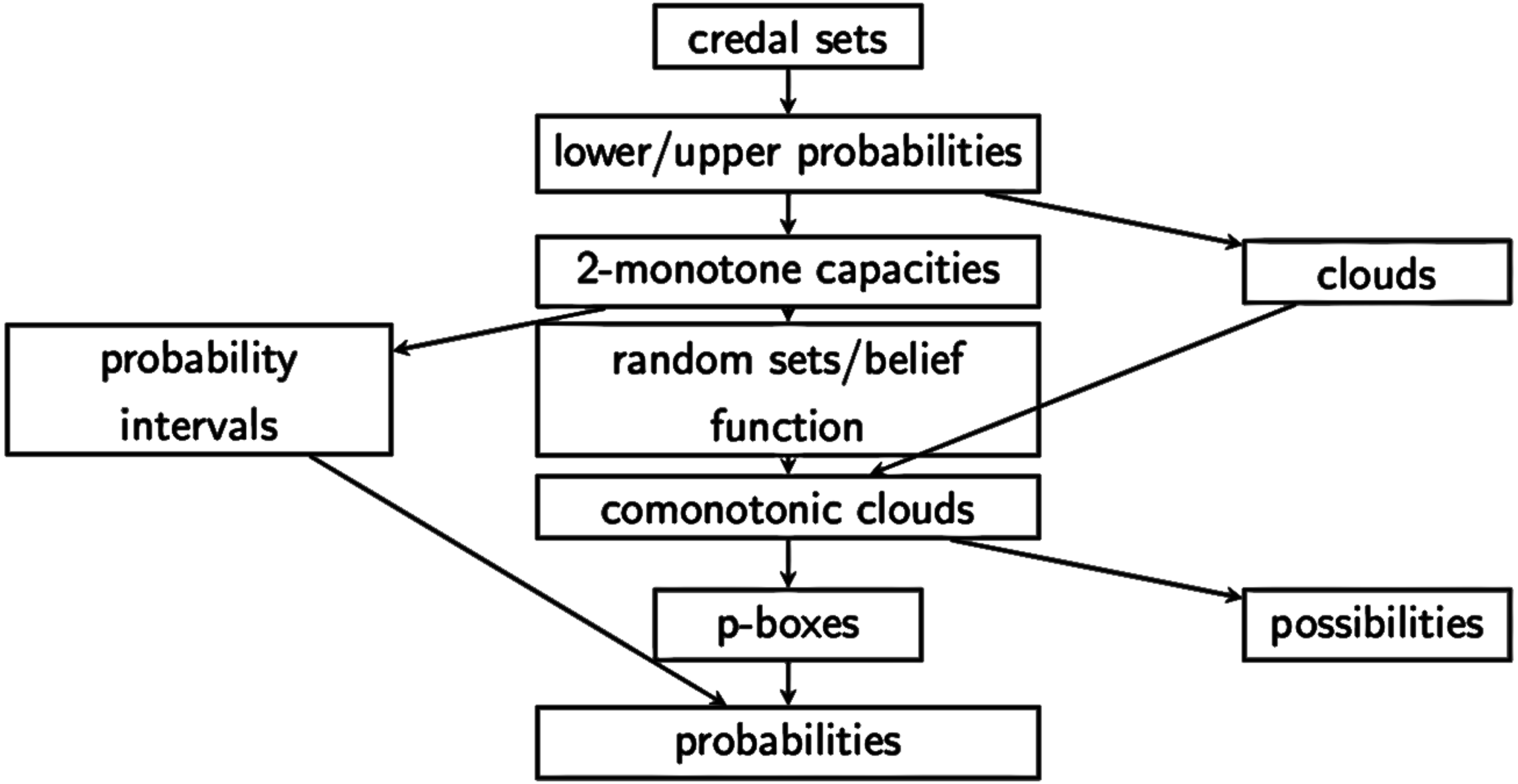
IMPRECISE PROBABILISTIC MACHINE LEARNING

Lecture 7: Imprecise Classification and Regression

Krikamol Muandet

12 December 2025

Overview


credal sets
lower/upper probabilities
2-monotone capacities
clouds
probability intervals
random sets/belief function
comonotonic clouds
p-boxes
possibilities
probabilities

Outline

1. Naïve Credal Classifier

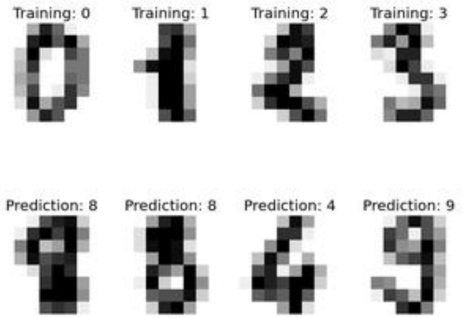
2. Computation

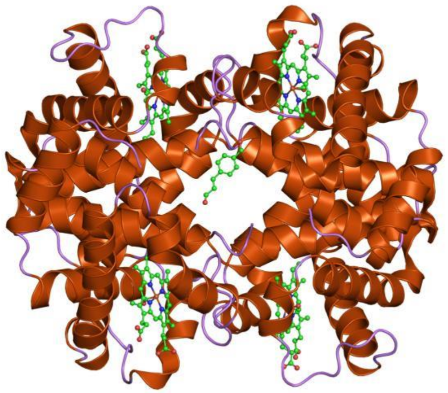
3. Real-world Example

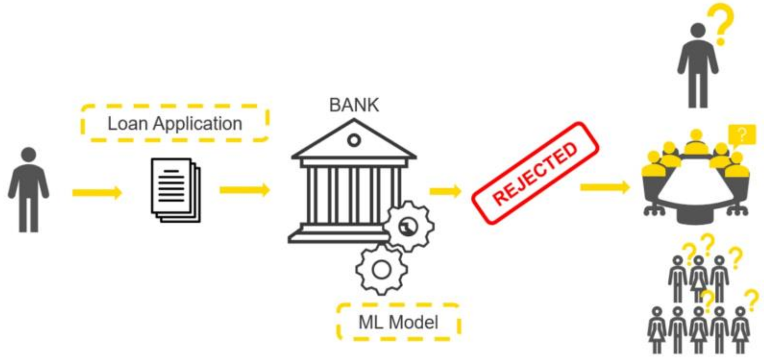
Naïve Credal Classifier

Classification

- Many real-world problems can be casted as a classification problem.







Problem Setup

- Y is the **target** variable taking values in $\mathcal{Y} = \{y_1, \dots, y_m\}, m \geq 2$.
- $X_1, X_2 \dots, X_d$ are d **features** taking values $x_1, x_2, \dots, x_d$ in $\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_d$.
- The discrete joint probability distribution

$$P(Y, X_1, X_2 \dots, X_d)$$

- The classification of a new pattern $(x_1, x_2, \dots, x_d)$ is realised by selecting a class $y \in \mathcal{Y}$ that maximises

$$P(y \mid x_1, x_2, \dots, x_d)$$

- This classification rule minimises the expected cost of misclassification.

Independence Assumption

- The number of probabilities grows **exponentially** with the number of attributes, i.e., k^d where k is the number of values X can take.
- Duda and Hart (1973) proposed an **independence** assumption:

$$P(X_1, X_2 \dots, X_d \mid Y) = \prod_{i=1}^{d} P(X_i \mid Y), \qquad (k^d \to kd)$$

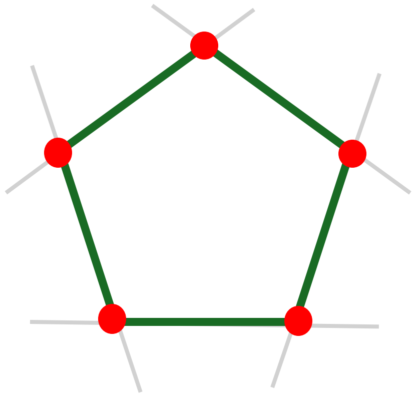
- This results in the **naïve Bayes classifier (NBC)**.
- While unrealistic, several work in the literature often claim that this assumption is not critical for classificatioin.

Naïve Credal Classifier

- A **naïve credal classifier (NCC)** is characterised by a credal set:

$$\mathcal{P} = \left\{ P(Y) \prod_{i=1}^{d} P(X_i|Y) \,\middle|\, P(Y) \in \mathcal{P}_y, P(X_i \mid y) \in \mathcal{P}_{X_i}^{y} \right\}$$

- Here, $\mathcal{P}_y$ is a **local** credal set of the probability distributions $P(Y)$ and $\mathcal{P}_{X_i}^{y}$ is a **local** credal set of the conditional distributions $P(X_i \mid y)$.
- How to specify the local credal sets $\mathcal{P}_y$ and $\mathcal{P}_{X_i}^{y}$?
    1. A finitely generated credal set (FGCS)
    2. Linear constraints on the unknown probabilities



Naïve Credal Classifier

- A **naïve credal classifier (NCC)** is characterised by a credal set:

$$\mathcal{P} = \left\{ P(Y) \prod_{i=1}^{d} P(X_i|Y) \,\middle|\, P(Y) \in \mathcal{P}_y, P(X_i \mid y) \in \mathcal{P}_{X_i}^{y} \right\}$$

- From the credal sets, we can compute the lower and upper probabilities:

$$\begin{aligned} \underline{P}(y) &= \min_{P \in \mathcal{P}_y} P(y), & \overline{P}(y) &= \max_{P \in \mathcal{P}_y} P(y) \\ \underline{P}(x_i \mid y) &= \min_{P \in \mathcal{P}_{X_i}^{y}} P(x_i \mid y), & \overline{P}(x_i \mid y) &= \max_{P \in \mathcal{P}_{X_i}^{y}} P(x_i \mid y) \end{aligned}$$

- NCC assumes that the local credal sets can be specified **separately**.

Credal Classification Rules

- NBC classifies a pattern $(x_1, x_2, \dots, x_d)$ by selecting the class $y \in \mathcal{Y}$ of **maximum** posterior probability $P(y \mid x_1, x_2, \dots, x_d)$.
- To compare intervals, consider **strong dominance** (Luce and Raiffa, 1957):

$$[a, b] \text{ dominates } [c, d] \Leftrightarrow \; a > d$$

- For each class $y \in \mathcal{Y}$, we only need to compute:

$$\underline{P}(y \mid x_1, x_2, \dots, x_d) = \min_{P(y, x_1, x_2, \dots, x_d) \in \mathcal{P}} P(y \mid x_1, x_2, \dots, x_d)$$

$$\overline{P}(y \mid x_1, x_2, \dots, x_d) = \max_{P(y, x_1, x_2, \dots, x_d) \in \mathcal{P}} P(y \mid x_1, x_2, \dots, x_d)$$

Duda and Hart (1973)

Computation

Computation

- Let's focus on $\underline{P}(y \mid x_1, x_2, \ldots, x_d) = \min_{P(y,x_1,x_2,\ldots,x_d) \in \mathcal{P}} P(y \mid x_1, x_2, \ldots, x_d)$

$$P(y \mid x_1, x_2, \ldots, x_d) = \frac{P(y,x_1,x_2,\ldots,x_d)}{\sum_{\acute{y}} P(\acute{y},x_1,x_2,\ldots,x_d)} = \left(1 + \frac{\sum_{\acute{y} \neq y} P(\acute{y},x_1,x_2,\ldots,x_d)}{P(y,x_1,x_2,\ldots,x_d)}\right)^{-1}$$

$$= \left(1 + \frac{\sum_{\acute{y} \neq y} P(\acute{y}) \prod_{i=1}^{d} P(x_i \mid \acute{y})}{P(y) \prod_{i=1}^{d} P(x_i \mid y)}\right)^{-1}$$

- The optimisation becomes

$$\min_{P(Y) \in \mathcal{P}_y} \min_{P(X_i \mid y) \in \mathcal{P}_{X_i}^{y}} \left(1 + \frac{\sum_{\acute{y} \neq y} P(\acute{y}) \prod_{i=1}^{d} P(x_i \mid \acute{y})}{P(y) \prod_{i=1}^{d} P(x_i \mid y)}\right)^{-1}$$

Computation

- The optimisation becomes

$$\min_{P(Y)\in\mathcal{P}_{\mathcal{Y}}} \min_{P(X_i\,|\,y)\in\mathcal{P}^{y}_{X_i}} \left(1+\frac{\sum_{\acute{y}\neq y} P(\acute{y}) \prod_{i=1}^{d} P(x_i\,|\,\acute{y})}{P(y)\prod_{i=1}^{d} P(x_i\,|\,y)}\right)^{-1}$$

$$= \min_{P(Y)\in\mathcal{P}_{\mathcal{Y}}} \left(1+\frac{\sum_{\acute{y}\neq y} P(\acute{y}) \prod_{i=1}^{d} \overline{P}(x_i\,|\,\acute{y})}{P(y)\prod_{i=1}^{d} \underline{P}(x_i\,|\,y)}\right)^{-1}$$

- Similar formular for the upper probability:

$$\overline{P}(y\,|\,x_1,x_2,\dots,x_d) = \max_{P(Y)\in\mathcal{P}_{\mathcal{Y}}} \left(1+\frac{\sum_{\acute{y}\neq y} P(\acute{y}) \prod_{i=1}^{d} \underline{P}(x_i\,|\,\acute{y})}{P(y)\prod_{i=1}^{d} \overline{P}(x_i\,|\,y)}\right)^{-1}$$

Interpretation

- The lower and upper posterior probabilities:

$$\underline{P}(y \mid x_1, x_2, \dots, x_d) = \min_{P(Y)\in\mathcal{P}_Y}\left(1 + \frac{\sum_{\acute{y}\neq y} P(\acute{y}) \prod_{i=1}^{d} \overline{P}(x_i \mid \acute{y})}{P(y) \prod_{i=1}^{d} \underline{P}(x_i \mid y)}\right)^{-1}$$

$$\overline{P}(y \mid x_1, x_2, \dots, x_d) = \max_{P(Y)\in\mathcal{P}_Y}\left(1 + \frac{\sum_{\acute{y}\neq y} P(\acute{y}) \prod_{i=1}^{d} \underline{P}(x_i \mid \acute{y})}{P(y) \prod_{i=1}^{d} \overline{P}(x_i \mid y)}\right)^{-1}$$

- We can think of $P(y)$ as **prior probabilities** and $\prod_{i=1}^{d} \overline{P}(x_i \mid y)$ and $\prod_{i=1}^{d} \underline{P}(x_i \mid y)$ as upper and lower **likelihood functions**.

Combinatorial Procedure

- Based on the **extreme points** of $\mathcal{P}_{\mathcal{Y}}$, we can rewrite the problems as

$$\underline{P}(y \mid x_1, x_2, \ldots, x_d) = \min_{P(Y) \in \text{ext}[\mathcal{P}_{\mathcal{Y}}]} \left(1 + \frac{\sum_{\acute{y} \neq y} P(\acute{y}) \prod_{i=1}^{d} \overline{P}(x_i \mid \acute{y})}{P(y) \prod_{i=1}^{d} \underline{P}(x_i \mid y)}\right)^{-1}$$

$$\overline{P}(y \mid x_1, x_2, \ldots, x_d) = \max_{P(Y) \in \text{ext}[\mathcal{P}_{\mathcal{Y}}]} \left(1 + \frac{\sum_{\acute{y} \neq y} P(\acute{y}) \prod_{i=1}^{d} \underline{P}(x_i \mid \acute{y})}{P(y) \prod_{i=1}^{d} \overline{P}(x_i \mid y)}\right)^{-1}$$

- To compute the lower and upper probabilities of $P(x_i \mid y)$ for $y \in \mathcal{Y}$:

$$\underline{P}(x_i \mid y) = \min_{P(x_i \mid y) \in \text{ext}[\mathcal{P}_{X_i}^{y}]} P(x_i \mid y)$$
$$\overline{P}(x_i \mid y) = \max_{P(x_i \mid y) \in \text{ext}[\mathcal{P}_{X_i}^{y}]} P(x_i \mid y)$$

Computational Complexity

- Let K be the **maximum of the number of extreme distributions**, taken over all the local credal sets.
- Computing the extremes of $P(x_i \mid y)$ takes $O(K)$ time in the worst case, which must be repeated for all classes and features, yielding $O(dK|\mathcal{Y}|)$.
- When the values of the conditional upper and lower probabilities are known, the final steps require $O(d|\mathcal{Y}|)$ time. Repearing this for each extreme distribution in $\mathrm{ext}[\mathcal{P}_{\mathcal{Y}}]$ yields $O(dK|\mathcal{Y}|)$.
- The overall worst-case complexity is

$$O(dK|\mathcal{Y}|).$$

Linear Programming

- The number of extreme points K can grow **exponentially**.
- Alternatively, consider a set of probability intervals

$$I_X = \{[l_i, u_i] | 0 \leq l_i \leq u_i \leq 1,\ i = 1, \ldots, T\}$$

- Let L be the worst-case complexity to solve a linear program. Then, the overall worst-case complexity is

$$O(dL|\mathcal{Y}|).$$

- L only grows **polynomially** even when K grows exponentially.

Real-World Example

Risk Assessment

- An insurance company wants to assess the risk (R) about the car insurance for a *new* customer:

$$R \in \{\text{low}, \text{medium}, \text{high}\}$$

- The company infers the risk on the basis of two attributes: the age (A) and the city (C) where the customer lives:
  - $A \in \{\text{Young},\ \text{Middle}-\text{aged}, \text{Old}\}$
  - $C \in \{\text{Tübingen,Saarbrücken, Berlin}\}$

Risk Assessment

Recall that the lower and upper probabilities can be defined as

$$\underline{P}(S) \coloneqq \max\left\{\sum_{x\in S}\underline{p}_x\,,\,1-\sum_{x\in S^{\mathrm{c}}}\overline{p}_x\right\}, \qquad \overline{P}(S) \coloneqq \min\left\{\sum_{x\in S}\overline{p}_x\,,\,1-\sum_{x\in S^{\mathrm{c}}}\underline{p}_x\right\}$$

$P(R)$

R	**Intervals**
Low	[0.77, 0.85]
Medium	[0.10, 0.15]
High	[0.05, 0.08]

$P(A \mid R)$

A	***R***		
	Low	Medium	High
Young	[0.15,0.22]	[0.27,0.32]	[0.60,0.70]
Middle-aged	[0.50,0.55]	[0.33,0.38]	[0.05,0.15]
Old	[0.28,0.34]	[0.34,0.38]	[0.20,0.30]

$P(C \mid R)$

C	***R***		
	Low	Medium	High
Tü	[0.70,0.72]	[0.15,0.20]	[0.02,0.06]
Saar	[0.18,0.20]	[0.60,0.65]	[0.22,0.28]
Berlin	[0.08,0.10]	[0.20,0.25]	[0.66,0.72]

Risk Assessment

$P(R \mid A = \text{old}, C = \text{Tü})$

R	Intervals
Low	[0.922, 0.975]
Medium	[0.024, 0.070]
High	[0.001, 0.009]

$P(R \mid A = \text{young}, C = \text{Ber})$

R	Intervals
Low	[0.150, 0.426]
Medium	[0.085, 0.290]
High	[0.435, 0.693]

$P(R \mid A = \text{young}, C = \text{Saar})$

R	Intervals
Low	[0.307, 0.621]
Medium	[0.238, 0.525]
High	[0.100, 0.267]

$$\underline{P}(y \mid x_1, x_2, \dots, x_d) = \min_{P(Y) \in \text{ext}[\mathcal{P}_{\mathcal{Y}}]} \left(1 + \frac{\sum_{\acute{y} \neq y} P(\acute{y}) \prod_{i=1}^{d} \overline{P}(x_i \mid \acute{y})}{P(y) \prod_{i=1}^{d} \underline{P}(x_i \mid y)}\right)^{-1}$$

$$\overline{P}(y \mid x_1, x_2, \dots, x_d) = \max_{P(Y) \in \text{ext}[\mathcal{P}_{\mathcal{Y}}]} \left(1 + \frac{\sum_{\acute{y} \neq y} P(\acute{y}) \prod_{i=1}^{d} \underline{P}(x_i \mid \acute{y})}{P(y) \prod_{i=1}^{d} \overline{P}(x_i \mid y)}\right)^{-1}$$

Summary

- Unlike standard naïve Bayes classifier, naïve credal classifier may produce a set of prediction $y \in \mathcal{Y}_{\mathrm{pred}} \subset \mathcal{Y}$.
- Interval dominance as a classification rule:

$$\left[\underline{P}(y \mid x_1, x_2, \dots, x_d),\ \overline{P}(y \mid x_1, x_2, \dots, x_d)\right]$$

- But the credal sets typically contain more information than the intervals.
  - Other classification rules such as **credal dominance** can be applied.
  - If there is a utility/loss function associated with the prediction $y \in \mathcal{Y}_{\mathrm{pred}}$, we can apply imprecise decision rules to obtain final prediction.
- A fundamental issue is **how to learn** the local credal sets from data.

Other credal classifiers

- Credal decision trees
- Imprecise Dirichlet model (IDM)
- Evidential k-NN
- Evidential Neural Network (ENN)

Regression

- In regression, the target is typically a real value $y \in \mathbb{R}$.
- To capture the uncertainty, we can instead output an interval $[y^{\min}, y^{\max}]$:

$$[y] = [\underline{y}, \overline{y}] = \left\{ y \in \mathbb{R} \middle| \underline{y} \leq y \leq \overline{y} \right\}$$

- For intervals $[a] = [\underline{a}, \overline{a}]$, $[b] = [\underline{b}, \overline{b}]$ basic operations are:

$$\begin{aligned} [\mathrm{a}] + [\mathrm{b}] &= [\underline{a} + \underline{b}, \overline{a} + \overline{b}] \\ [\mathrm{a}] - [\mathrm{b}] &= [\underline{a} - \overline{b}, \overline{a} - \underline{b}] \end{aligned}$$

- The second operation is problematic! Assume $a = b = [1,2]$. Then, we have

$$[\mathrm{a}] - [\mathrm{b}] = [1,2] - [1,2] = [-1,1].$$

dependency problem

Precise Regression

- Assume a linear dependence between input and output parametrised by w_* with an additive Gaussian noise:

$$y = {w_*}^{\mathrm{T}} x + \epsilon$$

- Here, $x \in \mathbb{R}^d$ is a d-dimensional input feature sampled from some distribution $\mathcal{P}$ and $y \in \mathbb{R}$ is the true label. The MSE loss over n data pair $\{(x_i, y_i)\}_{\mathrm{i}=1,\ldots,n}$ in matrix form is defined as

$$\widehat{\boldsymbol{w}} = \mathrm{argmin}_{\boldsymbol{w} \in \mathbb{R}^d} \frac{1}{2n} || \boldsymbol{y} - \boldsymbol{X}\boldsymbol{w} ||^2$$

- Assuming $\boldsymbol{X}$ is full rank, we have the unique solution:

$$\widehat{\boldsymbol{w}} = (\boldsymbol{X}^{\mathrm{T}} \boldsymbol{X})^{-1} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{y}$$

Imprecise Regression

- Assume that the labels $y_i s$ are intervals $y_i = [\underline{y_i}, \overline{y_i}]$. This means the parameters of our model are interval $\boldsymbol{w} = [\underline{\boldsymbol{w}}, \overline{\boldsymbol{w}}]$.
- For predictor $\boldsymbol{h} = X[\boldsymbol{w}]$ and interval labels $[\boldsymbol{y}]$, there are two main approaches:

$$L_{\text{INN}} = \tfrac{1}{2}\|[\boldsymbol{h}] - [\boldsymbol{y}]\|^2$$

$$L_{\text{ISH}} = \frac{1}{2}\left\|\underline{\boldsymbol{h}} - \underline{\boldsymbol{y}}\right\|^2 + \frac{1}{2}\|\overline{\boldsymbol{h}} - \overline{\boldsymbol{y}}\|^2$$

- The first approach involves interval operations making it prone to **dependency problem**.

Recommended Reading

- **The Naïve Credal Classifier** by Marco Zaffalon

Exercises