
THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

Uncertainty reasoning and machine learning

Some first probabilistic and credal classifiers

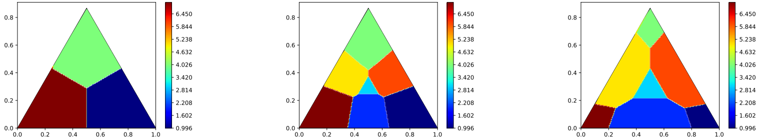
Vu-Linh Nguyen

Chaire de Professeur Junior, Laboratoire Heudiasyc
Université de technologie de Compiègne

AOS4 master courses

Optimal Decision Rules

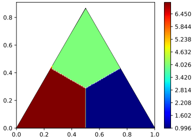
Frequentist approaches



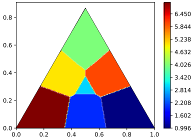
0/1 loss $u_{1.6}$ $u_{2.2}$

Optimal Decision Rules

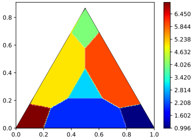
Frequentist approaches



0/1 loss

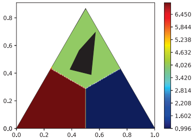


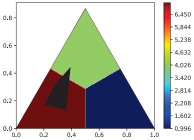
$u_{1.6}$

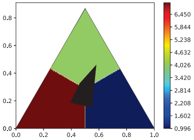


$u_{2.2}$

Credal approaches







Computational Aspects

Basic setup and assumption

- Given training data $\boldsymbol{D} \subset \mathscr{X} \times \mathscr{Y}$
- $\boldsymbol{D}$ is used to estimate a classifier, which predicts, for each $\boldsymbol{x}$, $\boldsymbol{\theta}|\boldsymbol{x}$

Optimal decision rules

- The Bayes-optimal prediction of any $\ell : \mathscr{Y} \times \mathscr{Y} \longmapsto \mathbb{R}_+$ on $\boldsymbol{x}$ is

$$y_\ell^{\boldsymbol{\theta}} = \operatorname*{argmin}_{\overline{y} \in \mathscr{Y}} \sum_{y \in \mathscr{Y}} \ell(\overline{y}, y)\theta_y|\boldsymbol{x}$$

- The Bayes-optimal prediction of any $\mathscr{L} : 2^{\mathscr{Y}} \setminus \{\emptyset\} \times \mathscr{Y} \longmapsto \mathbb{R}_+$ on $\boldsymbol{x}$ is

$$Y_{\mathscr{L}}^{\boldsymbol{\theta}} = \operatorname*{argmin}_{\overline{Y} \subset \mathscr{Y}} \sum_{y \in \mathscr{Y}} \mathscr{L}(\overline{Y}, y)\theta_y|\boldsymbol{x}$$

Computational Aspects (Cont.)

Basic setup and assumption

- Given training data $\boldsymbol{D} \subset \mathscr{X} \times \mathscr{Y}$
- $\boldsymbol{D}$ is used to estimate a classifier, which predicts, for each $\boldsymbol{x}$, $\boldsymbol{\Theta}|\boldsymbol{x}$

E-admissibility Rule [11, 13]:

- Let $\ell : \mathscr{Y} \times \mathscr{Y} \longmapsto \mathbb{R}_+$ be a loss. An optimal prediction is

$$Y^{E}_{\ell,\boldsymbol{\Theta}|\boldsymbol{x}} = \{y \in \mathscr{Y} \,|\, \exists \boldsymbol{\theta}|\boldsymbol{x} \in \boldsymbol{\Theta}|\boldsymbol{x} \text{ s.t. } y = y^{\boldsymbol{\theta}|\boldsymbol{x}}_{\ell}\}.$$

- Computation: Solving linear programs, etc.

Beyond Multi-Class Classification

Other predictive tasks:

- Multi-Label Classification
- Multi-Dimensional Classification
- Multi-Target Prediction

Beyond Multi-Class Classification

Other predictive tasks:

- Multi-Label Classification
- Multi-Dimensional Classification
- Multi-Target Prediction

Practical Challenges:

- Mixed features (e.g., Multimodal inputs)
- Insufficient training data: Imbalance, Scarce, Incomplete, Noise
- Incomplete test inputs

Multi-label classification with partial abstention

- **Precise predictions**: $\mathscr{Y} = \{0,1\}^K$
- **Set-valued predictions**: $\mathscr{Y}_{\text{set}} = 2^{\mathscr{Y}}$
- **Predictions with partial abstention**: $\mathscr{Y}_{\text{par}} = \{0,1,\perp\}^K \subsetneq \mathscr{Y}_{\text{set}}$

Multi-label classification with partial abstention

- **Precise predictions**: $\mathscr{Y} = \{0,1\}^K$
- **Set-valued predictions**: $\mathscr{Y}_{\text{set}} = 2^{\mathscr{Y}}$
- **Predictions with partial abstention**: $\mathscr{Y}_{\text{par}} = \{0,1,\perp\}^K \subsetneq \mathscr{Y}_{\text{set}}$

Journal of Artificial Intelligence Research 72 (2021) 613-665   Submitted 01/2021; published 11/2021

Multilabel Classification with Partial Abstention: Bayes-Optimal Prediction under Label Independence

Vu-Linh Nguyen   V.L.NGUYEN@TUE.NL
Department of Mathematics and Computer Science
Eindhoven University of Technology, The Netherlands

Eyke Hüllermeier   EYKE@LMU.DE
Department of Computer Science
University of Munich (LMU), Germany

Objectives

After this lecture, students should be able to

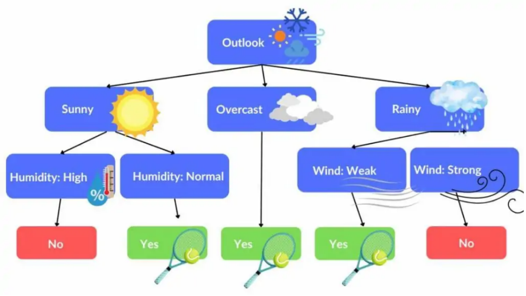
- use IDM and related models in Naïve credal classifier (NCC) [3]
- use IDM and related models in decision trees [9]

Outline

- Graphical Interpretation of Probabilistic Models
- Naïve Bayesian/Credal classifiers
- Decision Trees
- Bayesian Neural Networks
- Summary and Outlook

How to interpret a decision tree?

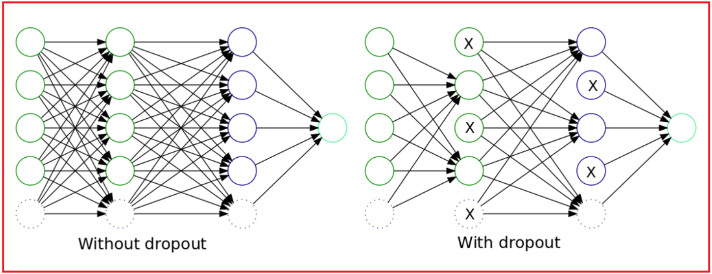
How to interpret a decision tree?



Source: `https://spotintelligence.com/2024/05/22/decision-trees-in-ml/`

How to interpret a (feedforward) neural network?

How to interpret a (feedforward) neural network?



Probabilistic Models: Graphical Interpretation [6, 10]

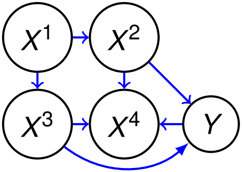
Basic setup

- A set of features $\boldsymbol{X} = \{X^1, \dots, X^M\}$; $[M] := \{1, \dots, M\}$
- A class variable Y whose outcome $y \in \mathscr{Y}$

Probabilistic Models: Graphical Interpretation [6, 10]

Basic setup

- A set of features $\boldsymbol{X} = \{X^1, \ldots, X^M\}$; $[M] := \{1, \ldots, M\}$
- A class variable Y whose outcome $y \in \mathscr{Y}$
- A directed acyclic graph (DAG) connecting Y and X^m



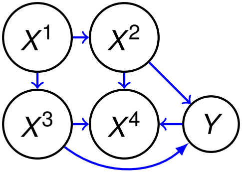
This DAG (model structure) tells us:

- $\text{pa}(Y) = \{X^2, X^3\}$, $\text{pa}(X^1) = \emptyset$
- $\text{pa}(X^2) = \{X^1\}$, $\text{pa}(X^3) = \{X^1\}$
- $\text{pa}(X^4) = \{Y, X^2, X^3\}$

Probabilistic Models: Graphical Interpretation [6, 10]

Basic setup

- A set of features $\boldsymbol{X} = \{X^1, \ldots, X^M\}$; $[M] := \{1, \ldots, M\}$
- A class variable Y whose outcome $y \in \mathscr{Y}$
- A directed acyclic graph (DAG) connecting Y and X^m



This DAG (model structure) tells us:

- $\text{pa}(Y) = \{X^2, X^3\}$, $\text{pa}(X^1) = \emptyset$
- $\text{pa}(X^2) = \{X^1\}$, $\text{pa}(X^3) = \{X^1\}$
- $\text{pa}(X^4) = \{Y, X^2, X^3\}$

Probabilistic Models:

- Expressing $P(Y, \boldsymbol{X})$ using the **chain rule** (probability):

$$P(Y, \boldsymbol{X}) = P(Y|\text{pa}(Y)) \prod_{m=1}^{M} P(X^m|\text{pa}(X^m)).$$

Probabilistic Models: Model Families [10]

Probabilistic Models:

- Estimate $P(Y, \boldsymbol{X})$
- Chain rule (probability):

$$P(Y, \boldsymbol{X}) = P(Y|\text{pa}(Y)) \prod_{m=1}^{M} P(X^m|\text{pa}(X^m)).$$

Extreme Cases:

- Discriminative models: $Y \notin \text{pa}(X^m)$, $m \in [M]$
- Generative models: $\text{pa}(Y) = \emptyset$ and $Y \in \text{pa}(X^p)$, $m \in [M]$.

Probabilistic Models: Model Families [10]

Probabilistic Models:

- Estimate $P(Y, \boldsymbol{X})$
- Chain rule (probability):

$$P(Y, \boldsymbol{X}) = P(Y|\text{pa}(Y)) \prod_{m=1}^{M} P(X^m|\text{pa}(X^m)).$$

Extreme Cases:

- Discriminative models: $Y \notin \text{pa}(X^m)$, $m \in [M]$
- Generative models: $\text{pa}(Y) = \emptyset$ and $Y \in \text{pa}(X^p)$, $m \in [M]$.

Model Families:

- How to encode/parametrize $P(Y|\text{pa}(Y))$ and $P(X^m|\text{pa}(X^m))$.
- How to estimate $P(Y, \boldsymbol{X})$ from training data.

Credal (Imprecise Probability) Models [5]

Basic setup

- A set of features $\boldsymbol{X} = \{X^1, \ldots, X^M\}$
- A class variable Y whose outcome $y \in \mathscr{Y}$

Credal Models:

- $\mathscr{P} := \{P(Y, \boldsymbol{X}) | P$ is compatible with knowledge/data$\}$
- Chain rule (probability):

$$P(Y, \boldsymbol{X}) = P(Y|\text{pa}(Y)) \prod_{m=1}^{M} P(X^m|\text{pa}(X^m)).$$

Extreme Cases:

- Discriminative models: $Y \notin \text{pa}(X^m)$, $m \in [M] := \{1, \ldots, M\}$
- Generative models: $\text{pa}(Y) = \emptyset$ and $Y \in \text{pa}(X^m)$, $m \in [M]$.

Model Families:

- How to encode/parametrize $P(Y|\text{pa}(Y))$ and $P(X^m|\text{pa}(X^m))$.
- How to estimate $\mathscr{P}(Y, \boldsymbol{X})$ from training data.

Assumptions and Questions

Assumption and desirable property:

A1. X^m, $m \in [M] := \{1, \ldots, M\}$, are always made available

P1. Best estimates of $P(Y|\mathrm{pa}(Y))$ and $P(X^m|\mathrm{pa}(X^m))$ can be found given (training) data.

Assumptions and Questions

Assumption and desirable property:

A1. X^m, $m \in [M] := \{1, \ldots, M\}$, are always made available

P1. Best estimates of $P(Y|\text{pa}(Y))$ and $P(X^m|\text{pa}(X^m))$ can be found given (training) data.

Questions (Exercise):

- Does the P1 hold for Naïve Bayes Classifier?
- Does the P1 hold for Decision trees?

Assumptions and Questions

Assumption and desirable property:

A1. X^m, $m \in [M] := \{1, \ldots, M\}$, are always made available

P1. Best estimates of $P(Y|\text{pa}(Y))$ and $P(X^m|\text{pa}(X^m))$ can be found given (training) data.

Questions (Exercise):

- Does the P1 hold for Naïve Bayes Classifier?
- Does the P1 hold for Decision trees?

Questions (which will not be discussed in this lecture):

- What may happen if X^m, $m \in [M]$, can be partially given?
- What may happen if best estimates of $P(Y|\text{pa}(Y))$ and $P(X^m|\text{pa}(X^m))$ may not be found?

The Next Slides

We shall elaborate on how to solve classification task using

- Naïve Bayesian classifier (NBC) (an example of generative model)
- Decision trees (DTs) (examples of discriminative model)

The Next Slides

We shall elaborate on how to solve classification task using

- Naïve Bayesian classifier (NBC) (an example of generative model)
- Decision trees (DTs) (examples of discriminative model)

How IDM (Lecture 1) can be used to generalize NBC and DTs to

- cope with the case of small and partial/missing data
- make set-valued predictions under the presence of uncertainty

The Next Slides

We shall elaborate on how to solve classification task using

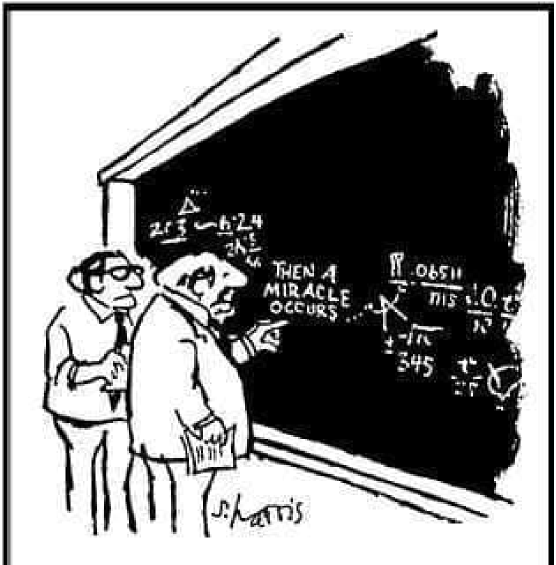
- Naïve Bayesian classifier (NBC) (an example of generative model)
- Decision trees (DTs) (examples of discriminative model)

How IDM (Lecture 1) can be used to generalize NBC and DTs to

- cope with the case of small and partial/missing data
- make set-valued predictions under the presence of uncertainty

We would also discuss (if we have time) the cases of

- Ensembles (Trees, Neural Nets, etc.)
- Bayesian Neural Nets



"I THINK YOU SHOULD BE MORE EXPLICIT HERE IN STEP TWO."

Outline

- Graphical Interpretation of Probabilistic Models
- Naïve Bayesian/Credal classifiers
  - Naïve Bayesian classifier
  - Naïve Credal classifiers
- Decision Trees
- Bayesian Neural Networks
- Summary and Outlook

Outline

Generative Models

Probabilistic Models:

- Estimate $P(Y, \boldsymbol{X})$
- Chain rule (probability):

$$P(Y, \boldsymbol{X}) = P(Y|\text{pa}(Y)) \prod_{m=1}^{M} P(X^m|\text{pa}(X^m)).$$

Extreme Cases:

- Discriminative models: $Y \notin \text{pa}(X^m)$, $m \in [M]$
- Generative models: $\text{pa}(Y) = \emptyset$ and $Y \in \text{pa}(X^p)$, $m \in [M]$.

Model Families:

- How to encode/parametrize $P(Y|\text{pa}(Y))$ and $P(X^m|\text{pa}(X^m))$.
- How to estimate $P(Y, \boldsymbol{X})$ from training data.

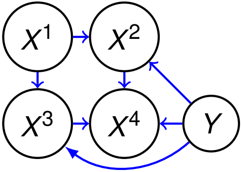
Generative Models: Structure

Let's start with an example where one wishes to model

$$P(Y,\boldsymbol{X}) = P(Y, X^1, X^2, X^3, X^4).$$

Chain rule (probability):

$$P(Y,\boldsymbol{X}) = P(Y|\text{pa}(Y)) \prod_{m=1}^{M} P(X^m|\text{pa}(X^m)).$$



- $\text{pa}(Y) = \emptyset$, $\text{pa}(X^1) = \emptyset$
- $\text{pa}(X^2) = \{Y, X^1\}$
- $\text{pa}(X^3) = \{Y, X^1\}$
- $\text{pa}(X^4) = \{Y, X^2, X^3\}$

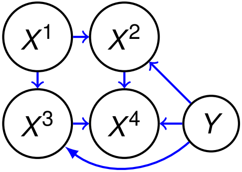
Generative Models: Structure

Let's start with an example where one wishes to model

$$P(Y,\boldsymbol{X}) = P(Y,X^1,X^2,X^3,X^4).$$

Chain rule (probability):

$$P(Y,\boldsymbol{X}) = P(Y|\text{pa}(Y))\prod_{m=1}^{M} P(X^m|\text{pa}(X^m)).$$



- $\text{pa}(Y) = \emptyset$, $\text{pa}(X^1) = \emptyset$
- $\text{pa}(X^2) = \{Y, X^1\}$
- $\text{pa}(X^3) = \{Y, X^1\}$
- $\text{pa}(X^4) = \{Y, X^2, X^3\}$

The chain rule gives us

$$P(Y,\boldsymbol{X}) = P(Y)P(X^1)P(X^2|Y,X^1)P(X^3|Y,X^1)P(X^4|Y,X^2,X^3).$$

Naïve Bayesian classifier (NBC)

Comments:

- NBC is a generative model with no arc $X' \longrightarrow X$
- Chain rule gives us

$$P(Y, \boldsymbol{X}) = P(Y) \prod_{m=1}^{M} P(X^m | Y).$$

Naïve Bayesian classifier (NBC)

Comments:

- NBC is a generative model with no arc $X' \longrightarrow X$
- Chain rule gives us

$$P(Y, \boldsymbol{X}) = P(Y) \prod_{m=1}^{M} P(X^m | Y).$$

To solve the **classification task**,

- joint probability distribution $P(Y, \boldsymbol{X})$ is learn from training data $\boldsymbol{D}$
- conditional distribution $P(Y|\boldsymbol{X})$ is extracted using **Bayes' theorem**

$$P(y|\boldsymbol{x}) = \frac{P(y, \boldsymbol{x})}{\sum_{y' \in \mathscr{Y}} P(y', \boldsymbol{x})} = \frac{P(y) \prod_{m=1}^{M} P(x^m|y)}{\sum_{y' \in \mathscr{Y}} P(y') \prod_{m=1}^{M} P(x^m|y')}. \tag{1}$$

Estimate Parameters of NBC

Basic setup:

- A class variable Y with K possible values: $\mathscr{Y} = \{y^1, \ldots y^K\}$
- M discrete features: $\boldsymbol{X} = (X^1, \ldots, X^M)$
- Feature X^m has Q_m possible values: $\mathscr{X}^m = \{x^{m,1}, \ldots x^{m,Q_m}\}$

Estimate Parameters of NBC

Basic setup:

- A class variable Y with K possible values: $\mathscr{Y} = \{y^1, \ldots y^K\}$
- M discrete features: $\boldsymbol{X} = (X^1, \ldots, X^M)$
- Feature X^m has Q_m possible values: $\mathscr{X}^m = \{x^{m,1}, \ldots x^{m,Q_m}\}$

Task: Finding the best estimate of

- $\theta_k := P(y^k)$, $k \in [K]$
- $\theta_k^{m,q_m} := P(x^{q_m,m} | y^k)$, $q_m \in [Q_m]$, $k \in [K]$, $m \in [M]$

Estimate Parameters of NBC

Basic setup:

- A class variable Y with K possible values: $\mathscr{Y} = \{y^1, \dots y^K\}$
- M discrete features: $\boldsymbol{X} = (X^1, \dots, X^M)$
- Feature X^m has Q_m possible values: $\mathscr{X}^m = \{x^{m,1}, \dots x^{m,Q_m}\}$

Task: Finding the best estimate of

- $\theta_k := P(y^k)$, $k \in [K]$
- $\theta_k^{m,q_m} := P(x^{q_m,m} | y^k)$, $q_m \in [Q_m]$, $k \in [K]$, $m \in [M]$

Probability axioms:

- $\sum_{k=1}^{K} \theta_k = 1$
- $\sum_{q_m=1}^{Q_m} \theta_k^{m,q_m} = 1$ when fixing k and m

Maximum Likelihood Estimate

Basic setup: Given training data $\boldsymbol{D} = \{(y_1, \boldsymbol{x}_1), \ldots, (y_N, \boldsymbol{x}_N)\}$, count

- n_k: Number of training instances with label y^k
- n_k^{m,q_m}: Number of training instances with label y^k and feature X^m takes value x^{m,q_m}

Maximum Likelihood Estimate

Basic setup: Given training data $\boldsymbol{D} = \{(y_1, \boldsymbol{x}_1), \ldots, (y_N, \boldsymbol{x}_N)\}$, count

- n_k: Number of training instances with label y^k
- n_k^{m,q_m}: Number of training instances with label y^k and feature X^m takes value x^{m,q_m}

MLE gives us the best estimates

$$\theta_k := n_k / N \quad (2)$$

$$\theta_k^{m,q_m} := n^{m,qm} / n_k \quad (3)$$

MLE with Examples

- $\mathscr{Y} = \{A, B, C\}$
- $\mathscr{X}^1 = \{d, e\}$
- $\mathscr{X}^2 = \{f, g, h\}$

n	Y	X^1	X^2
1	A	d	f
2	A	d	g
3	A	e	g
4	B	d	f
5	B	e	g
6	C	d	f
7	C	e	f
8	C	e	g

MLE with Examples

- $\mathcal{Y} = \{A, B, C\}$
- $\mathcal{X}^1 = \{d, e\}$
- $\mathcal{X}^2 = \{f, g, h\}$

n	Y	X^1	X^2
1	A	d	f
2	A	d	g
3	A	e	g
4	B	d	f
5	B	e	g
6	C	d	f
7	C	e	f
8	C	e	g

$$n_A = 3 \quad n_B = 2 \quad n_C = 3$$
$$\theta_A = 3/8 \quad \theta_B = 1/4 \quad \theta_C = 3/8$$

$n_A^{1,d} = 2$	$n_A^{1,e} = 1$	$\theta_A^{1,d} = 2/3$	$\theta_A^{1,e} = 1/3$
$n_B^{1,d} = 1$	$n_B^{1,e} = 1$	$\theta_B^{1,d} = 1/2$	$\theta_B^{1,e} = 1/2$
$n_C^{1,d} = 1$	$n_C^{1,e} = 2$	$\theta_C^{1,d} = 1/3$	$\theta_C^{1,e} = 2/3$
$n_A^{2,f} = 1$	$n_A^{2,g} = 2$	$\theta_A^{2,f} = 1/3$	$\theta_A^{2,g} = 2/3$
$n_B^{2,f} = 1$	$n_B^{2,g} = 1$	$\theta_B^{2,f} = 1/2$	$\theta_B^{2,g} = 1/2$
$n_C^{2,f} = 2$	$n_C^{2,g} = 1$	$\theta_C^{2,f} = 2/3$	$\theta_C^{2,g} = 1/3$

$$n_A^{2,h} = 0 \quad n_B^{2,h} = 0 \quad n_C^{2,h} = 0$$
$$\theta_A^{2,h} = 0 \quad \theta_B^{2,h} = 0 \quad \theta_C^{2,h} = 0$$

Conditional Probabilities

Given $\boldsymbol{x} = (x^{1,q_1}, \ldots, x^{M,q_M})$, for any $y^k \in \mathscr{Y}$:

$$P(y^k|\boldsymbol{x}) = \frac{\theta_k \prod_{m=1}^{M} \theta_k^{m,q_m}}{\sum_{y_{k'} \in \mathscr{Y}} \theta_{k'} \prod_{m=1}^{M} \theta_{k'}^{m,q_m}} \propto P'(y^k|\boldsymbol{x}) = \theta_k \prod_{m=1}^{M} \theta_k^{m,q_m}. \quad (4)$$

Conditional Probabilities

Given $\boldsymbol{x} = (x^{1,q_1}, \dots, x^{M,q_M})$, for any $y^k \in \mathscr{Y}$:

$$P(y^k|\boldsymbol{x}) = \frac{\theta_k \prod_{m=1}^M \theta_k^{m,q_m}}{\sum_{y_{k'} \in \mathscr{Y}} \theta_{k'} \prod_{m=1}^M \theta_{k'}^{m,q_m}} \propto P'(y^k|\boldsymbol{x}) = \theta_k \prod_{m=1}^M \theta_k^{m,q_m}. \qquad (4)$$

$$\theta_A = 3/8 \quad \theta_B = 1/4 \quad \theta_C = 3/8$$

$\theta_A^{1,d} = 2/3$	$\theta_A^{1,e} = 1/3$	$\theta_A^{2,f} = 1/3$	$\theta_A^{2,g} = 2/3$
$\theta_B^{1,d} = 1/2$	$\theta_B^{1,e} = 1/2$	$\theta_B^{2,f} = 1/2$	$\theta_B^{2,g} = 1/2$
$\theta_C^{1,d} = 1/3$	$\theta_C^{1,e} = 2/3$	$\theta_C^{2,f} = 2/3$	$\theta_C^{2,g} = 1/3$

$$\theta_A^{2,h} = 0 \quad \theta_B^{2,h} = 0 \quad \theta_C^{2,h} = 0$$

$\boldsymbol{x}$	$P'(A\|\boldsymbol{x})$	$P'(B\|\boldsymbol{x})$	$P'(C\|\boldsymbol{x})$
(d,f)	$1/12$	$1/16$	$1/12$
(e,h)	0	0	0

Optimal Decision Rules

If $\ell(y^{k'}, y^k) = \mathbb{1}(y^{k'} \neq y^k)$, then (Check!)

$$y_\ell^{\boldsymbol{\theta}}(\boldsymbol{x}) = \underset{y^k \in \mathscr{Y}}{\operatorname{argmax}}\, P'(y^k|\boldsymbol{x})$$

Optimal Decision Rules

If $\ell(y^{k'}, y^k) = \mathbb{1}(y^{k'} \neq y^k)$, then (Check!)

$$y_\ell^{\boldsymbol{\theta}}(\boldsymbol{x}) = \underset{y^k \in \mathcal{Y}}{\operatorname{argmax}}\, P'(y^k|\boldsymbol{x})$$

$\boldsymbol{x}$	$P'(A\|\boldsymbol{x})$	$P'(B\|\boldsymbol{x})$	$P'(C\|\boldsymbol{x})$
(d, f)	$1/12$	$1/16$	$1/12$
(e, h)	0	0	0

- If $\boldsymbol{x} = (d, f)$, then

$$y_\ell^{\boldsymbol{\theta}}(\boldsymbol{x}) = \text{ either } A \text{ or } C, \tag{5}$$

- If $\boldsymbol{x} = (e, h)$, then

$$y_\ell^{\boldsymbol{\theta}}(\boldsymbol{x}) = \text{ not well-defined }, \tag{6}$$

NBC + MLE: Comments

- May lead to **indecision** and **not well-defined** $P(y|\boldsymbol{x})$

NBC + MLE: Comments

- May lead to **indecision** and **not well-defined** $P(y|\boldsymbol{x})$

$\boldsymbol{x}$	$P'(A\|\boldsymbol{x})$	$P'(B\|\boldsymbol{x})$	$P'(C\|\boldsymbol{x})$
(d,f)	$1/12$	$1/16$	$1/12$
(e,h)	0	0	0

- May suffer from small numbers of observations
  - n_k: Number of training instances with label y^k
  - n_k^{m,q_m}: Number of training instances with label y^k and feature X^m takes value x^{m,q_m}

NBC + MLE: Comments

- May lead to **indecision** and **not well-defined** $P(y|\boldsymbol{x})$

$\boldsymbol{x}$	$P'(A\|\boldsymbol{x})$	$P'(B\|\boldsymbol{x})$	$P'(C\|\boldsymbol{x})$
(d, f)	$^1/_{12}$	$^1/_{16}$	$^1/_{12}$
(e, h)	0	0	0

- May suffer from small numbers of observations
  - n_k: Number of training instances with label y^k
  - n_k^{m,q_m}: Number of training instances with label y^k and feature X^m takes value x^{m,q_m}
- Does not (naturally) take into account missing/partial data

NBC + Dirichlet Model (DM)

Basic setup: Given training data $\boldsymbol{D} = \{(y_1, \boldsymbol{x}_1), \ldots, (y_N, \boldsymbol{x}_N)\}$, count

- n_k: Number of training instances with label y^k
- n_k^{m,q_m}: Number of instances with $Y = y^k$ and feature $X^m = x^{m,q_m}$

NBC + Dirichlet Model (DM)

Basic setup: Given training data $\boldsymbol{D} = \{(y_1, \boldsymbol{x}_1), \ldots, (y_N, \boldsymbol{x}_N)\}$, count

- n_k: Number of training instances with label y^k
- n_k^{m,q_m}: Number of instances with $Y = y^k$ and feature $X^m = x^{m,q_m}$

DM gives Bayesian estimates

$$\theta_k := {}^{(n_k+\alpha_k)}\!/_{(N+s)} = {}^{(n_k+sf_k)}\!/_{(N+s)} \quad (7)$$

$$\theta_k^{m,q_m} := {}^{(n_k^{m,q_m}+\alpha_k^{m,q_m})}\!/_{(n_k+s)} = {}^{(n_k^{m,q_m}+sf_k^{m,q_m})}\!/_{(n_k+s)} \quad (8)$$

NBC + Dirichlet Model (DM)

Basic setup: Given training data $\boldsymbol{D} = \{(y_1, \boldsymbol{x}_1), \ldots, (y_N, \boldsymbol{x}_N)\}$, count

- n_k: Number of training instances with label y^k
- n_k^{m,q_m}: Number of instances with $Y = y^k$ and feature $X^m = x^{m,q_m}$

DM gives Bayesian estimates

$$\theta_k := (n_k + \alpha_k) / (N + s) = (n_k + sf_k) / (N + s) \tag{7}$$

$$\theta_k^{m,q_m} := (n_k^{m,q_m} + \alpha_k^{m,q_m}) / (n_k + s) = (n_k^{m,q_m} + sf_k^{m,q_m}) / (n_k + s) \tag{8}$$

Advocators	α_v (= y^k or x^{m,q_m})	s
Haldane (1948)	0	0
Perks (1947)	$1/\lvert\mathcal{V}\rvert$	1
Jeffreys (1946, 1961)	$1/2$	$\lvert\mathcal{V}\rvert/2$
Bayes-Laplace	1	$\lvert\mathcal{V}\rvert$

NBC + DM with Examples

$$\theta_k := {(n_k+1/3)}/{(N+1)}, \qquad \theta_k^{m,q_m} := {(n_k^{m,q_m}+1/|\mathscr{X}^m|)}/{(n_k+1)}.$$

NBC + DM with Examples

$$\theta_k := {}^{(n_k+1/3)}\!/_{(N+1)}, \qquad \theta_k^{m,q_m} := {}^{(n_k^{m,q_m}+1/|\mathscr{X}^m|)}\!/_{(n_k+1)}.$$

- $\mathscr{Y} = \{A, B, C\}$
- $\mathscr{X}^1 = \{d, e\}$
- $\mathscr{X}^2 = \{f, g, h\}$

n	Y	X^1	X^2
1	A	d	f
2	A	d	g
3	A	e	g
4	B	d	f
5	B	e	g
6	C	d	f
7	C	e	f
8	C	e	g

$$n_A = 3 \qquad n_B = 2 \qquad n_C = 3$$
$$\theta_A = {}^{10}\!/_{27} \qquad \theta_B = {}^{7}\!/_{27} \qquad \theta_C = {}^{10}\!/_{27}$$

$n_A^{1,d} = 2$	$n_A^{1,e} = 1$	$\theta_A^{1,d} = 5/8$	$\theta_A^{1,e} = 3/8$
$n_B^{1,d} = 1$	$n_B^{1,e} = 1$	$\theta_B^{1,d} = 3/6$	$\theta_B^{1,e} = 3/6$
$n_C^{1,d} = 1$	$n_C^{1,e} = 2$	$\theta_C^{1,d} = 3/8$	$\theta_C^{1,e} = 5/8$
$n_A^{2,f} = 1$	$n_A^{2,g} = 2$	$\theta_A^{2,f} = 4/12$	$\theta_A^{2,g} = 7/12$
$n_B^{2,f} = 1$	$n_B^{2,g} = 1$	$\theta_B^{2,f} = 4/9$	$\theta_B^{2,g} = 4/9$
$n_C^{2,f} = 2$	$n_C^{2,g} = 1$	$\theta_C^{2,f} = 7/12$	$\theta_C^{2,g} = 4/12$

$$n_A^{2,h} = 0 \qquad n_B^{2,h} = 0 \qquad n_C^{2,h} = 0$$
$$\theta_A^{2,h} = {}^{1}\!/_{12} \qquad \theta_B^{2,h} = {}^{1}\!/_{9} \qquad \theta_C^{2,h} = {}^{1}\!/_{12}$$

Conditional Probabilities

Given $\boldsymbol{x} = (x^{1,q_1}, \ldots, x^{M,q_M})$, for any $y^k \in \mathcal{Y}$:

$$P(y^k|\boldsymbol{x}) = \frac{\theta_k \prod_{m=1}^{M} \theta_k^{m,q_m}}{\sum_{y_{k'} \in \mathcal{Y}} \theta_{k'} \prod_{m=1}^{M} \theta_{k'}^{m,q_m}} \propto P'(y^k|\boldsymbol{x}) = \theta_k \prod_{m=1}^{M} \theta_k^{m,q_m}. \qquad (9)$$

$$\theta_A = 10/27 \quad \theta_B = 7/27 \quad \theta_C = 10/27$$

$\theta_A^{1,d} = 5/8$	$\theta_A^{1,e} = 3/8$	$\theta_A^{2,f} = 4/12$	$\theta_A^{2,g} = 7/12$
$\theta_B^{1,d} = 3/6$	$\theta_B^{1,e} = 3/6$	$\theta_B^{2,f} = 4/9$	$\theta_B^{2,g} = 4/9$
$\theta_C^{1,d} = 3/8$	$\theta_C^{1,e} = 5/8$	$\theta_C^{2,f} = 7/12$	$\theta_C^{2,g} = 4/12$

$$\theta_A^{2,h} = 1/12 \quad \theta_B^{2,h} = 1/9 \quad \theta_C^{2,h} = 1/12$$

Conditional Probabilities

Given $\boldsymbol{x} = (x^{1,q_1}, \ldots, x^{M,q_M})$, for any $y^k \in \mathcal{Y}$:

$$P(y^k|\boldsymbol{x}) = \frac{\theta_k \prod_{m=1}^{M} \theta_k^{m,q_m}}{\sum_{y_{k'} \in \mathcal{Y}} \theta_{k'} \prod_{m=1}^{M} \theta_{k'}^{m,q_m}} \propto P'(y^k|\boldsymbol{x}) = \theta_k \prod_{m=1}^{M} \theta_k^{m,q_m}. \qquad (9)$$

$$\theta_A = 10/27 \quad \theta_B = 7/27 \quad \theta_C = 10/27$$

$\theta_A^{1,d} = 5/8$	$\theta_A^{1,e} = 3/8$	$\theta_A^{2,f} = 4/12$	$\theta_A^{2,g} = 7/12$
$\theta_B^{1,d} = 3/6$	$\theta_B^{1,e} = 3/6$	$\theta_B^{2,f} = 4/9$	$\theta_B^{2,g} = 4/9$
$\theta_C^{1,d} = 3/8$	$\theta_C^{1,e} = 5/8$	$\theta_C^{2,f} = 7/12$	$\theta_C^{2,g} = 4/12$

$$\theta_A^{2,h} = 1/12 \quad \theta_B^{2,h} = 1/9 \quad \theta_C^{2,h} = 1/12$$

$\boldsymbol{x}$	$P'(A\|\boldsymbol{x})$	$P'(B\|\boldsymbol{x})$	$P'(C\|\boldsymbol{x})$	$y_\ell^{\boldsymbol{\theta}}(\boldsymbol{x})$
(d,f)	$\frac{10}{27}\frac{5}{8}\frac{4}{12}$	$\frac{7}{27}\frac{3}{6}\frac{4}{9}$	$\frac{10}{27}\frac{3}{8}\frac{7}{12}$	C
(e,h)	$\frac{10}{27}\frac{3}{8}\frac{1}{12}$	$\frac{7}{27}\frac{3}{6}\frac{1}{9}$	$\frac{10}{27}\frac{5}{8}\frac{1}{12}$	C

NBC + DM: Comments

- May lead to **indecision**, but can avoid **not well-defined** $P(y|\boldsymbol{x})$

NBC + DM: Comments

- May lead to **indecision**, but can avoid **not well-defined** $P(y|\boldsymbol{x})$

$\boldsymbol{x}$	$P'(A|\boldsymbol{x})$	$P'(B|\boldsymbol{x})$	$P'(C|\boldsymbol{x})$	$y_\ell^{\boldsymbol{\theta}}(\boldsymbol{x})$
(d,f)	$\frac{10}{27}\frac{5}{8}\frac{4}{12}$	$\frac{7}{27}\frac{3}{6}\frac{4}{9}$	$\frac{10}{27}\frac{3}{8}\frac{7}{12}$	C
(e,h)	$\frac{10}{27}\frac{3}{8}\frac{1}{12}$	$\frac{7}{27}\frac{3}{6}\frac{1}{9}$	$\frac{10}{27}\frac{5}{8}\frac{1}{12}$	C

- May suffer from small numbers of observations
  - n_k: Number of training instances with label y^k
  - n_k^{m,q_m}: Number of training instances with label y^k and feature X^m takes value x^{m,q_m}

NBC + DM: Comments

- May lead to **indecision**, but can avoid **not well-defined** $P(y|\boldsymbol{x})$

$\boldsymbol{x}$	$P'(A\|\boldsymbol{x})$	$P'(B\|\boldsymbol{x})$	$P'(C\|\boldsymbol{x})$	$y_\ell^{\boldsymbol{\theta}}(\boldsymbol{x})$
(d,f)	$\frac{10}{27}\frac{5}{8}\frac{4}{12}$	$\frac{7}{27}\frac{3}{6}\frac{4}{9}$	$\frac{10}{27}\frac{3}{8}\frac{7}{12}$	C
(e,h)	$\frac{10}{27}\frac{3}{8}\frac{1}{12}$	$\frac{7}{27}\frac{3}{6}\frac{1}{9}$	$\frac{10}{27}\frac{5}{8}\frac{1}{12}$	C

- May suffer from small numbers of observations
  - n_k: Number of training instances with label y^k
  - n_k^{m,q_m}: Number of training instances with label y^k and feature X^m takes value x^{m,q_m}
- Does not (naturally) take into account missing/partial data

Outline

- Graphical Interpretation of Probabilistic Models
- Naïve Bayesian/Credal classifiers
  - Naïve Bayesian classifier
  - Naïve Credal classifiers
- Decision Trees
- Bayesian Neural Networks
- Summary and Outlook

Naïve Credal classifiers (NCC)

Basic setup: Given training data $\boldsymbol{D} = \{(y_1, \boldsymbol{x}_1), \ldots, (y_N, \boldsymbol{x}_N)\}$, count

- n_k: Number of training instances with label y^k
- n_k^{m,q_m}: Number of instances with $Y = y^k$ and feature $X^m = x^{m,q_m}$

Naïve Credal classifiers (NCC)

Basic setup: Given training data $\boldsymbol{D} = \{(y_1, \boldsymbol{x}_1), \ldots, (y_N, \boldsymbol{x}_N)\}$, count

- n_k: Number of training instances with label y^k
- n_k^{m,q_m}: Number of instances with $Y = y^k$ and feature $X^m = x^{m,q_m}$

Imprecise Dirichlet model (IDM) gives

$$\underline{\theta}_k := {n_k}/{(N+s)} \quad (10) \qquad \overline{\theta}_k := {(n_k+s)}/{(N+s)} \quad (12)$$

$$\underline{\theta}_k^{m,q_m} := {n_k^{m,q_m}}/{(n_k+s)} \quad (11) \qquad \overline{\theta}_k^{m,q_m} := {(n_k^{m,q_m}+s}/{(n_k+s)} \quad (13)$$

Naïve Credal classifiers (NCC)

Basic setup: Given training data $\boldsymbol{D} = \{(y_1, \boldsymbol{x}_1), \ldots, (y_N, \boldsymbol{x}_N)\}$, count

- n_k: Number of training instances with label y^k
- n_k^{m,q_m}: Number of instances with $Y = y^k$ and feature $X^m = x^{m,q_m}$

Imprecise Dirichlet model (IDM) gives

$$\underline{\theta}_k := n_k/(N+s) \quad (10)$$

$$\underline{\theta}_k^{m,q_m} := n_k^{m,qm}/(n_k+s) \quad (11)$$

$$\overline{\theta}_k := (n_k+s)/(N+s) \quad (12)$$

$$\overline{\theta}_k^{m,q_m} := (n_k^{m,qm}+s/(n_k+s) \quad (13)$$

IDM + ϵ regularization [2]

$$\underline{\theta}_k := (n_k+s\underline{\epsilon}_k)/(N+s) \quad (14)$$

$$\underline{\theta}_k^{m,q_m} := (n_k^{m,qm}+s\underline{\epsilon}_k^{m,qm})/(n_k+s) \quad (15)$$

$$\overline{\theta}_k := (n_k+s\overline{\epsilon}_k)/(N+s) \quad (16)$$

$$\overline{\theta}_k^{m,q_m} := (n_k^{m,qm}+s\overline{\epsilon}_k^{m,qm})/(n_k+s) \quad (17)$$

Interval Conditional Probabilities

Given a query instance $\boldsymbol{x} = (x^{q_1,1}, x^{q_2,2}, \dots, x^{q_M,M})$, we have

$$1/\overline{P}(y^k|\boldsymbol{x}) - 1 = \sum_{k' \neq k} \left(\frac{n_{k'} + s\underline{\epsilon}_k}{n_k + s\overline{\epsilon}_k} \left(\frac{n_k + s}{n_{k'} + s} \right)^M \prod_{m=1}^{M} \frac{n_{k'}^{q_m,m} + s\underline{\epsilon}_k^{m,q_m}}{n_k^{q_m,m} + s\overline{\epsilon}_k^{m,q_m}} \right),$$

$$1/\underline{P}(y^k|\boldsymbol{x}) - 1 = \sum_{k' \neq k} \left(\frac{n_{k'} + s\overline{\epsilon}_k}{n_k + s\underline{\epsilon}_k} \left(\frac{n_k + s}{n_{k'} + s} \right)^M \prod_{m=1}^{M} \frac{n_{k'}^{q_m,m} + s\overline{\epsilon}_k^{m,q_m}}{n_k^{q_m,m} + s\underline{\epsilon}_k^{m,q_m}} \right).$$

Interval Conditional Probabilities

Given a query instance $\boldsymbol{x} = (x^{q_1,1}, x^{q_2,2}, \ldots, x^{q_M,M})$, we have

$$1/\overline{P}(y^k|\boldsymbol{x}) - 1 = \sum_{k' \neq k} \left(\frac{n_{k'} + s\underline{\epsilon}_k}{n_k + s\overline{\epsilon}_k} \left(\frac{n_k + s}{n_{k'} + s} \right)^M \prod_{m=1}^{M} \frac{n_{k'}^{q_m,m} + s\underline{\epsilon}_k^{m,q_m}}{n_k^{q_m,m} + s\overline{\epsilon}_k^{m,q_m}} \right),$$

$$1/\underline{P}(y^k|\boldsymbol{x}) - 1 = \sum_{k' \neq k} \left(\frac{n_{k'} + s\overline{\epsilon}_k}{n_k + s\underline{\epsilon}_k} \left(\frac{n_k + s}{n_{k'} + s} \right)^M \prod_{m=1}^{M} \frac{n_{k'}^{q_m,m} + s\overline{\epsilon}_k^{m,q_m}}{n_k^{q_m,m} + s\underline{\epsilon}_k^{m,q_m}} \right).$$

$$\mathscr{P}(\mathscr{Y}|\boldsymbol{x}) := \left\{ P(\mathscr{Y}|\boldsymbol{x}) | P(y^k|\boldsymbol{x}) \in [\underline{P}(y^k|\boldsymbol{x}), \overline{P}(y^k|\boldsymbol{x})], \sum_{k=1}^{K} P(y^k|\boldsymbol{x}) = 1 \right\}.$$

Interval Conditional Probabilities

$$1/\overline{P}(y^k|\boldsymbol{x}) - 1 = \sum_{k' \neq k} \left(\frac{n_{k'} + s\underline{\epsilon}_k}{n_k + s\overline{\epsilon}_k} \left(\frac{n_k + s}{n_{k'} + s} \right)^M \prod_{m=1}^{M} \frac{n_{k'}^{q_m,m} + s\underline{\epsilon}_k^{m,q_m}}{n_k^{q_m,m} + s\overline{\epsilon}_k^{m,q_m}} \right),$$

$$1/\underline{P}(y^k|\boldsymbol{x}) - 1 = \sum_{k' \neq k} \left(\frac{n_{k'} + s\overline{\epsilon}_k}{n_k + s\underline{\epsilon}_k} \left(\frac{n_k + s}{n_{k'} + s} \right)^M \prod_{m=1}^{M} \frac{n_{k'}^{q_m,m} + s\overline{\epsilon}_k^{m,q_m}}{n_k^{q_m,m} + s\underline{\epsilon}_k^{m,q_m}} \right).$$

Interval Conditional Probabilities

$$1/\overline{P}(y^k|\boldsymbol{x}) - 1 = \sum_{k' \neq k} \left(\frac{n_{k'} + s\underline{\epsilon}_k}{n_k + s\overline{\epsilon}_k} \left(\frac{n_k + s}{n_{k'} + s} \right)^M \prod_{m=1}^{M} \frac{n_{k'}^{q_m,m} + s\underline{\epsilon}_k^{m,q_m}}{n_k^{q_m,m} + s\overline{\epsilon}_k^{m,q_m}} \right),$$

$$1/\underline{P}(y^k|\boldsymbol{x}) - 1 = \sum_{k' \neq k} \left(\frac{n_{k'} + s\overline{\epsilon}_k}{n_k + s\underline{\epsilon}_k} \left(\frac{n_k + s}{n_{k'} + s} \right)^M \prod_{m=1}^{M} \frac{n_{k'}^{q_m,m} + s\overline{\epsilon}_k^{m,q_m}}{n_k^{q_m,m} + s\underline{\epsilon}_k^{m,q_m}} \right).$$

$s = 1 \quad \underline{\epsilon}_k = 0.01 \quad \overline{\epsilon}_k = 0.99 \quad n_A = 3 \quad n_B = 2 \quad n_C = 3$

$n_A^{1,d} = 2$	$n_A^{1,e} = 1$	$n_A^{2,f} = 1$	$n_A^{2,g} = 2$
$n_B^{1,d} = 1$	$n_B^{1,e} = 1$	$n_B^{2,f} = 1$	$n_B^{2,g} = 1$
$n_C^{1,d} = 1$	$n_C^{1,e} = 2$	$n_C^{2,f} = 2$	$n_C^{2,g} = 1$

$n_A^{2,h} = 0 \quad n_B^{2,h} = 0 \quad n_C^{2,h} = 0$

$\boldsymbol{x}$	$\underline{P}(A\|\boldsymbol{x})$	$\underline{P}(B\|\boldsymbol{x})$	$\underline{P}(C\|\boldsymbol{x})$	$\overline{P}(A\|\boldsymbol{x})$	$\overline{P}(B\|\boldsymbol{x})$	$\overline{P}(C\|\boldsymbol{x})$
(d,f)	???	???	???	???	???	???
(e,h)	???	???	???	???	???	???

Making Set-Valued Predictions (Recap)

For **each instance** $\boldsymbol{x}$, let

- $\boldsymbol{\theta} \longleftarrow P(\mathscr{Y}|\boldsymbol{x})$ and $\boldsymbol{\Theta} \longleftarrow \mathscr{P}(\mathscr{Y}|\boldsymbol{x})$

E-admissibility Rule:

- An optimal prediction is

$$\mathbf{Y}^{E}_{\ell,\boldsymbol{\Theta}} = \{y \in \mathscr{Y} | \exists \boldsymbol{\theta} \in \boldsymbol{\Theta} \text{ s.t. } y = y^{\theta}_{\ell}\}.$$

- Computation: Solving linear programs [11] , etc.

Maximality Rule:

- An optimal prediction is

$$\mathbf{Y}^{M}_{\ell,\boldsymbol{\Theta}} = \{y \in \mathscr{Y} | \nexists y' \text{ s.t. } y' \succ_{\ell,\boldsymbol{\Theta}} y\}.$$

- Computation: Solving linear programs [11], Iterating over the extreme points of $\boldsymbol{\Theta}$ [11], exploiting the properties of NCC [3].

NCC: Comments

NCC inherits properties of IDM [3]:

- May lead to **set-valued predictions**
- ϵ-regularization can avoid **not well-defined** $P(y|\boldsymbol{x})$
- May provide reliable interval probabilities when seeing small numbers of observations
  - n_k: Number of training instances with label y^k
  - n_k^{m,q_m}: Number of training instances with label y^k and feature X^m takes value x^{m,q_m}

NCC: Comments

NCC inherits properties of IDM [3]:

- May lead to **set-valued predictions**
- ϵ-regularization can avoid **not well-defined** $P(y|\boldsymbol{x})$
- May provide reliable interval probabilities when seeing small numbers of observations
  - n_k: Number of training instances with label y^k
  - n_k^{m,q_m}: Number of training instances with label y^k and feature X^m takes value x^{m,q_m}
- Provide tools to (naturally) take into account missing/partial data
  - Naive solutions are computationally expensive (in exponential time)
  - More efficient (polynomial-time) procedure exists

NCC: Technical Details + Performance

Journal of Machine Learning Research 9 (2008) 581-621 Submitted 1/07; Revised 2/08; Published 4/08

Learning Reliable Classifiers From Small or Incomplete Data Sets: The Naive Credal Classifier 2

Giorgio Corani GIORGIO@IDSIA.CH
Marco Zaffalon ZAFFALON@IDSIA.CH
IDSIA
Istituto Dalle Molle di Studi sull'Intelligenza Artificiale
CH-6928 Manno (Lugano), Switzerland

Editor: Charles Elkan

Outline

- Graphical Interpretation of Probabilistic Models
- Naïve Bayesian/Credal classifiers
- **Decision Trees**
  - Decision Trees
  - Credal Decision Trees
- Bayesian Neural Networks
- Summary and Outlook

Outline

- Graphical Interpretation of Probabilistic Models
- Naïve Bayesian/Credal classifiers
- Decision Trees
  - Decision Trees
  - Credal Decision Trees
- Bayesian Neural Networks
- Summary and Outlook

Discriminative Models

Probabilistic Models:

- Estimate $P(Y, \boldsymbol{X})$
- Chain rule (probability):

$$P(Y, \boldsymbol{X}) = P(Y|\text{pa}(Y)) \prod_{m=1}^{M} P(X^m|\text{pa}(X^m)).$$

Extreme Cases:

- Discriminative models: $Y \notin \text{pa}(X^m)$, $m \in [M]$
- Generative models: $\text{pa}(Y) = \emptyset$ and $Y \in \text{pa}(X^p)$, $m \in [M]$.

Model Families:

- How to encode/parametrize $P(Y|\text{pa}(Y))$ and $P(X^m|\text{pa}(X^m))$.
- How to estimate $P(Y, \boldsymbol{X}) = P(Y|\boldsymbol{X})P(\boldsymbol{X})$ from training data.

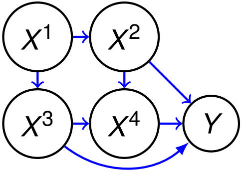
Discriminative Models: Structure

Let's start with an example where one wishes to model

$$P(Y, \boldsymbol{X}) = P(Y, X^1, X^2, X^3, X^4).$$

Chain rule (probability):

$$P(Y, \boldsymbol{X}) = P(Y|\text{pa}(Y)) \prod_{m=1}^{M} P(X^m|\text{pa}(X^m)).$$



- $\text{pa}(Y) = \{X^2, X^3, X^4\}$
- $\text{pa}(X^1) = \emptyset$, $\text{pa}(X^2) = \{X^1\}$
- $\text{pa}(X^3) = \{X^1\}$
- $\text{pa}(X^4) = \{X^2, X^3\}$

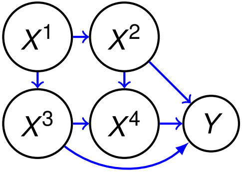
Discriminative Models: Structure

Let's start with an example where one wishes to model

$$P(Y, \boldsymbol{X}) = P(Y, X^1, X^2, X^3, X^4).$$

Chain rule (probability):

$$P(Y, \boldsymbol{X}) = P(Y|\text{pa}(Y)) \prod_{m=1}^{M} P(X^m|\text{pa}(X^m)).$$



- $\text{pa}(Y) = \{X^2, X^3, X^4\}$
- $\text{pa}(X^1) = \emptyset$, $\text{pa}(X^2) = \{X^1\}$
- $\text{pa}(X^3) = \{X^1\}$
- $\text{pa}(X^4) = \{X^2, X^3\}$

The chain rule gives us

$$P(Y, \boldsymbol{X}) = P(Y|X^2, X^3, X^4) P(X^1) P(X^2|X^1) P(X^3|X^1) P(X^4|X^2, X^3).$$

Classification Task

Prove that

$$P(y|\boldsymbol{x}) = P(y|\mathrm{pa}(y)). \tag{18}$$

Classification Task

Prove that

$$P(y|\boldsymbol{x}) = P(y|\mathrm{pa}(y)). \tag{18}$$

We have

$$P(y|\boldsymbol{x}) = \frac{P(y,\boldsymbol{x})}{\sum_{y'\in\mathcal{Y}} P(y',\boldsymbol{x})} \tag{19}$$

$$= \frac{P(y|\mathrm{pa}(y))\prod_{m=1}^{M} P(x^m|\mathrm{pa}(x^m))}{\sum_{y'\in\mathcal{Y}} P(y'|\mathrm{pa}(y'))\prod_{m=1}^{M} P(x^m|\mathrm{pa}(x^m))} \tag{20}$$

$$= \frac{\prod_{m=1}^{M} P(x^m|\mathrm{pa}(x^m))P(y|\mathrm{pa}(y))}{\prod_{m=1}^{M} P(x^m|\mathrm{pa}(x^m))\sum_{y'\in\mathcal{Y}} P(y'|\mathrm{pa}(y'))} \tag{21}$$

$$= \frac{P(y|\mathrm{pa}(y))}{\sum_{y'\in\mathcal{Y}} P(y'|\mathrm{pa}(y'))} \tag{22}$$

$$= P(y|\mathrm{pa}(y)). \tag{23}$$

Classification Task: Comments

$P(Y|\boldsymbol{X})$ is extracted using **Bayes' theorem**

$$P(y|\boldsymbol{x}) = P(y|\text{pa}(y)). \quad (24)$$

Classification Task: Comments

$P(Y|\boldsymbol{X})$ is extracted using **Bayes' theorem**

$$P(y|\boldsymbol{x}) = P(y|\text{pa}(y)). \tag{24}$$

- Features outside $\text{pa}(Y)$ are redundant
- To solve the classification task, we only need $P(Y|\text{pa}(Y))$
- Commonly used assumption: $\text{pa}(Y) = (X^1, \ldots, X^M)$

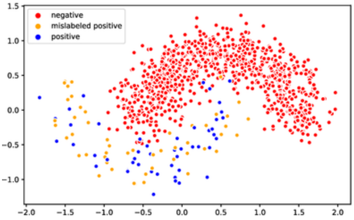
Classification Task: Comments

$P(Y|\boldsymbol{X})$ is extracted using **Bayes' theorem**

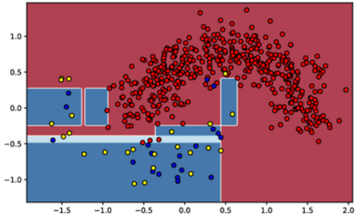
$$P(y|\boldsymbol{x}) = P(y|\text{pa}(y)). \tag{24}$$

- Features outside $\text{pa}(Y)$ are redundant
- To solve the classification task, we only need $P(Y|\text{pa}(Y))$
- Commonly used assumption: $\text{pa}(Y) = (X^1, \ldots, X^M)$
- $P(Y|\text{pa}(Y))$ can be defined either globally or locally:
  - Logistic regression, neural nets, etc., define $P(Y|\text{pa}(Y))$ globally
  - Decision tree, model trees, etc., define $P(Y|\text{pa}(Y))$ locally
  - Decision tree does not require $\text{pa}(Y) = (X^1, \ldots, X^M)$

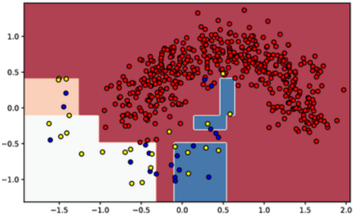
Decision Trees: Example [12]



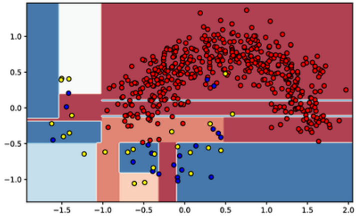
(a) Half-moons data set (ground truth)



(b) PU-Hellinger Decision Tree on the test set



(c) Hellinger Decision Tree on the test set



(d) CART on the test set

Decision Trees: (Informal+Probabilistic) Definition

A decision trees is

- a collection of non-overlapping leaves $L_1, \ldots, L_H$
- where $L_1 \cap \ldots \cap L_H = \mathscr{X}$
- and each leaf L_h has its own $P_h(Y|\text{pa}(Y))$

Decision Trees: (Informal+Probabilistic) Definition

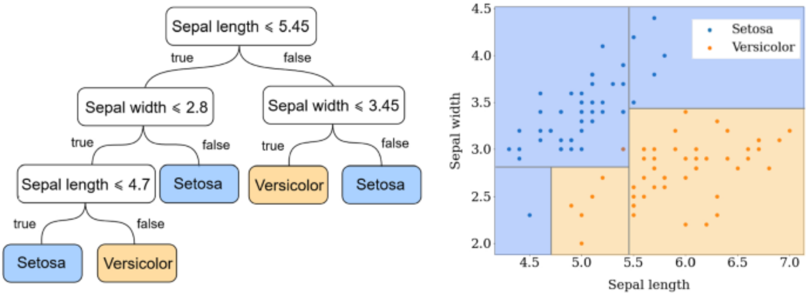
A decision trees is

- a collection of non-overlapping leaves $L_1, \ldots, L_H$
- where $L_1 \cap \ldots \cap L_H = \mathcal{X}$
- and each leaf L_h has its own $P_h(Y|\text{pa}(Y))$

Learning an optimal decision tree from training data

- can be extremely hard (due to huge numbers of possible trees)
- and is often done approximately (top-down induction, bottom-up induction, etc.)

Top-Down Induction (Example) [4]



(a) Tree visualization **(b)** Partitioning visualization

Top-down induction: Steps

Basic Setup:

- Training data $\boldsymbol{D} = \{(y^n, \boldsymbol{x}^n) | n \in [N]\}$
- Local hypothesis space $P(Y|\text{pa}(Y)) \in \mathscr{P}(Y|\text{pa}(Y))$
- An uncertainty measure U or a loss function $\ell \longleftarrow$ assess how good/bad each **local classifier** is

Top-down induction: Steps

Basic Setup:

- Training data $\boldsymbol{D} = \{(y^n, \boldsymbol{x}^n) | n \in [N]\}$
- Local hypothesis space $P(Y|\text{pa}(Y)) \in \mathscr{P}(Y|\text{pa}(Y))$
- An uncertainty measure U or a loss function $\ell \longleftarrow$ assess how good/bad each **local classifier** is

Induction protocol:

- Recursively partition the feature space $\mathscr{X}$
- From the current node, choose the best split which improves the evaluation criterion
- Evaluation criteria: Information gain, entropy, Gini score, etc.,
- Stopping criteria: No more gain on evaluation criterion U or ℓ

Splitting Criteria: Entropy (Frequentist)

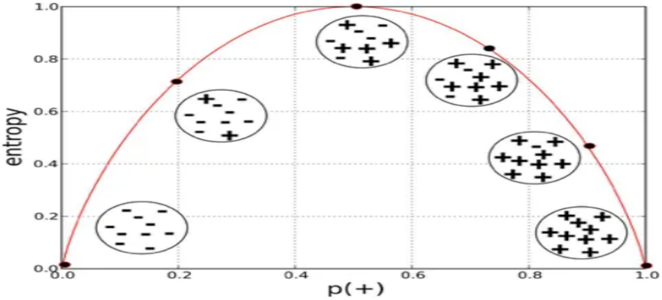
- Entropy of a node $\boldsymbol{D}_h \subset \boldsymbol{D}$ with $P(\mathscr{Y}|\boldsymbol{D}_h)$

$$U_E(P(\mathscr{Y}|\boldsymbol{D}_h)) = -\sum_{y\in\mathscr{Y}} P(y|\boldsymbol{D}_h)\log_2(P(y|\boldsymbol{D}_h)).$$

Splitting Criteria: Entropy (Frequentist)

- Entropy of a node $\boldsymbol{D}_h \subset \boldsymbol{D}$ with $P(\mathscr{Y}|\boldsymbol{D}_h)$

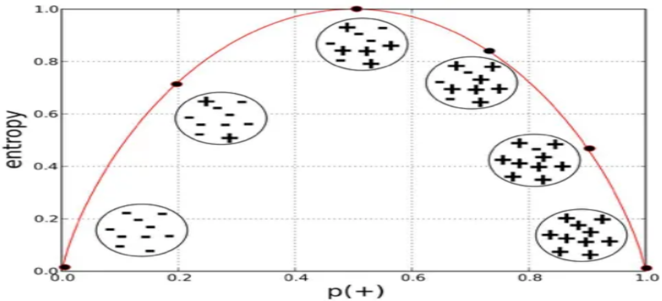
$$U_E(P(\mathscr{Y}|\boldsymbol{D}_h)) = -\sum_{y\in\mathscr{Y}} P(y|\boldsymbol{D}_h)\log_2(P(y|\boldsymbol{D}_h)).$$



Splitting Criteria: Entropy (Frequentist)

- Entropy of a node $\boldsymbol{D}_h \subset \boldsymbol{D}$ with $P(\mathscr{Y}|\boldsymbol{D}_h)$

$$U_E(P(\mathscr{Y}|\boldsymbol{D}_h)) = -\sum_{y\in\mathscr{Y}} P(y|\boldsymbol{D}_h)\log_2(P(y|\boldsymbol{D}_h))\,.$$



- For each possible split $\boldsymbol{D}_h = \boldsymbol{D}_h^1 \cup \boldsymbol{D}_h^2$, its entropy is

$$U_E(\boldsymbol{D}_h^1 \cup \boldsymbol{D}_h^2) = U_E(P(\mathscr{Y}|\boldsymbol{D}_h)) + U_E(P(\mathscr{Y}|\boldsymbol{D}_h))\,.$$

Compute Entropy

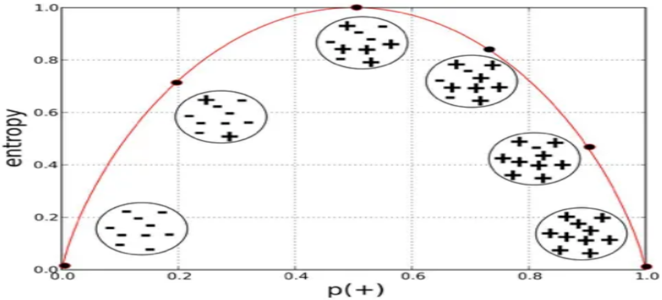
- Entropy of a node $\boldsymbol{D}_h \subset \boldsymbol{D}$ with $P(\mathscr{Y}|\boldsymbol{D}_h)$

$$U_E(P(y|\boldsymbol{D}_h)) = -\sum_{y\in\mathscr{Y}} P(y|\boldsymbol{D}_h)\log_2(P(y|\boldsymbol{D}_h)).$$

Compute Entropy

- Entropy of a node $\boldsymbol{D}_h \subset \boldsymbol{D}$ with $P(\mathscr{Y}|\boldsymbol{D}_h)$

$$U_E(P(y|\boldsymbol{D}_h)) = -\sum_{y\in\mathscr{Y}} P(y|\boldsymbol{D}_h)\log_2(P(y|\boldsymbol{D}_h))\,.$$



- Entropy of the bottom left node is 0
- Entropy of the top node is $-0.5\log_2(0.5) + 0.5\log_2(0.5) = 1$

Splitting Criteria: Bayesian

In principle, we can employ Dirichlet models (DM) to

- derive Bayesian estimates of $P(\mathscr{Y}|\boldsymbol{D}_h)$ and/or $U(P(y|\boldsymbol{D}_h))$
- and modify the top-down induction steps.

Splitting Criteria: Bayesian

In principle, we can employ Dirichlet models (DM) to

- derive Bayesian estimates of $P(\mathscr{Y}|\boldsymbol{D}_h)$ and/or $U(P(y|\boldsymbol{D}_h))$
- and modify the top-down induction steps.

- I haven't seen such decision trees
- I hope I can find some reference soon ...

Outline

- Graphical Interpretation of Probabilistic Models
- Naïve Bayesian/Credal classifiers
- Decision Trees
  - Decision Trees
  - Credal Decision Trees
- Bayesian Neural Networks
- Summary and Outlook

Where and How to be Imprecise?

In principle, we can employ Imprecise Dirichlet models (IDM) to

- derive interval estimates of $P(\mathscr{Y}|\boldsymbol{D}_h)$ and/or $U(P(y|\boldsymbol{D}_h))$
- and modify the top-down induction steps.

Where and How to be Imprecise?

In principle, we can employ Imprecise Dirichlet models (IDM) to

- derive interval estimates of $P(\mathscr{Y}|\boldsymbol{D}_h)$ and/or $U(P(y|\boldsymbol{D}_h))$
- and modify the top-down induction steps.

Credal Decision Trees [1, 8]

- Use IDM to derive interval estimates $\mathscr{P}(\mathscr{Y}|\boldsymbol{D}_h)$ of $P(\mathscr{Y}|\boldsymbol{D}_h)$
- Seek the highest entropy

$$U(\mathscr{P}(\mathscr{Y}|\boldsymbol{D}_h)) = \max_{P \in \mathscr{P}} U(P(\mathscr{Y}|\boldsymbol{D}_h)). \quad (25)$$

- Each leaf is equipped a $\mathscr{P}(\mathscr{Y}|\boldsymbol{L}_h) \rightarrow$ Precise predictions.

Credal Decision Trees: Performance [8]

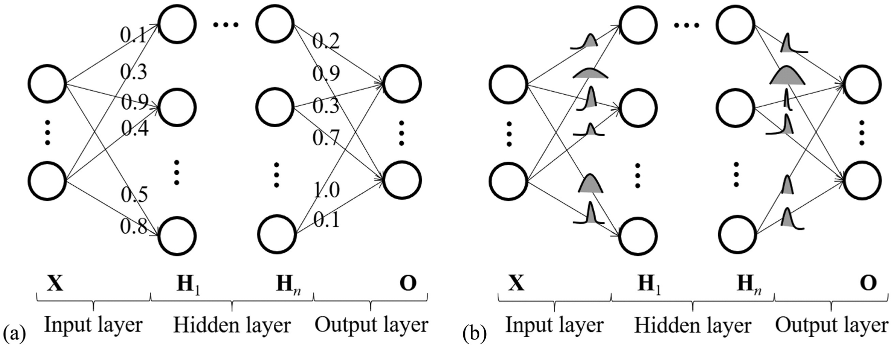
Splitting criterion	0% noise	10% noise	20% noise
Info-Gain (IG)	78.96	77.49	74.76
Info-Gain Ratio (IGR)	78.97	77.66	75.14
Imprecise Info-Gain (IIG)	79.56	78.65	76.72
Complete IIG (CIIG)	**79.63**	**78.66**	**76.74**

Table: 10×10-fold cross-validation procedure: Average accuracies (on 60 data sets) with random noise to the features and the class variable

Outline

- Graphical Interpretation of Probabilistic Models
- Naïve Bayesian/Credal classifiers
- Decision Trees
- Bayesian Neural Networks
- Summary and Outlook

Artificial neural networks vs Bayesian Neural Networks



Graphical interpretation of (a) ANN and (b) BNN

Inference Problems [7]

Algorithm 1 **Inference procedure for a BNN.**

Define $p(\theta|D) = \frac{p(D_y|D_x,\theta)p(\theta)}{\int_\theta p(D_y|D_x,\theta')p(\theta')d\theta'}$;

for $i = 0$ **to** N **do**

  Draw $\theta_i \sim p(\theta|D)$;

  $\mathbf{y}_i = \Phi_{\theta_i}(\mathbf{x})$;

end for

return $Y = \{\mathbf{y}_i | i \in [0, N)\}$, $\Theta = \{\theta_i | i \in [0, N)\}$;

Inference Problems [7]

Algorithm 1 **Inference procedure for a BNN.**

Define $p(\theta|D) = \dfrac{p(D_y|D_x,\theta)\,p(\theta)}{\int_\theta p(D_y|D_x,\theta')\,p(\theta')\,d\theta'}$;

for $i = 0$ **to** N **do**

  Draw $\theta_i \sim p(\theta|D)$;

  $\boldsymbol{y}_i = \Phi_{\theta_i}(\boldsymbol{x})$;

end for

return $Y = \{\boldsymbol{y}_i | i \in [0, N)\}$, $\Theta = \{\theta_i | i \in [0, N)\}$;

- We need some way to aggregate the set of outputs

Aggregation procedures

Predict-then-aggregate (You can try it yourself):

- For each Monte Carlo sample, turn $\Theta_\theta(\mathbf{x})$ into a hard prediction y.
- Aggregate the set of hard predictions into the final hard prediction.
- You might want to try with MLE (Frequentist), DM (Bayesian), IDM (IP), etc.

Aggregation procedures

Predict-then-aggregate (You can try it yourself):

- For each Monte Carlo sample, turn $\Theta_\theta(\mathbf{x})$ into a hard prediction y.
- Aggregate the set of hard predictions into the final hard prediction.
- You might want to try with MLE (Frequentist), DM (Bayesian), IDM (IP), etc.

Aggregation-then-predict (You can try it yourself):

- For each Monte Carlo sample, compute a soft prediction $\Theta_\theta(\mathbf{x})$.
- Aggregate the set of soft predictions into either
  - the final hard prediction
  - or a credal set, from which IP decision rules can be applied.

.

Software Impacts 9 (2021) 100079

Contents lists available at ScienceDirect

Software Impacts

journal homepage: www.journals.elsevier.com/software-impacts

ELSEVIER

Original software publication

BNNpriors: A library for Bayesian neural network inference with different prior distributions

Vincent Fortuin [a,1,*], Adrià Garriga-Alonso [b,1], Mark van der Wilk [c,2], Laurence Aitchison [d,2]

[a] ETH Zürich, Zürich, Switzerland
[b] University of Cambridge, Cambridge, UK
[c] Imperial College London, London, UK
[d] University of Bristol, Bristol, UK

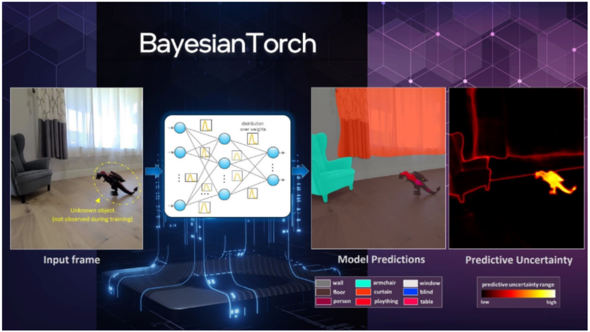
ARTICLE INFO

Keywords:
Machine learning
Bayesian neural networks
Prior distributions

ABSTRACT

Bayesian neural networks have shown great promise in many applications where calibrated uncertainty estimates are crucial and can often also lead to a higher predictive performance. However, it remains challenging to choose a good prior distribution over their weights. While isotropic Gaussian priors are often chosen in practice due to their simplicity, they do not reflect our true prior beliefs well and can lead to suboptimal performance. Our new library, *BNNpriors*, enables state-of-the-art Markov Chain Monte Carlo inference on Bayesian neural networks with a wide range of predefined priors, including heavy-tailed ones, hierarchical ones, and mixture priors. Moreover, it follows a modular approach that eases the design and implementation of new custom priors. It has facilitated foundational discoveries on the nature of the cold posterior effect in Bayesian neural networks and will hopefully catalyze future research as well as practical applications in this area.

Improve Trustworthiness in Deep Learning Models with Bayesian-Torch



Outline

- Graphical Interpretation of Probabilistic Models
- Naïve Bayesian/Credal classifiers
- Decision Trees
- Bayesian Neural Networks
- Summary and Outlook

Probabilistic Models [10]

Probabilistic Graphical Models:

- Estimate $P(Y, \boldsymbol{X})$
- Chain rule (probability):

$$P(Y, \boldsymbol{X}) = P(Y|\mathrm{pa}(Y)) \prod_{m=1}^{M} P(X^m|\mathrm{pa}(X^m)).$$

Extreme Cases:

- Discriminative models: $Y \notin \mathrm{pa}(X^m)$, $m \in [M]$
- Generative models: $\mathrm{pa}(Y) = \emptyset$ and $Y \in \mathrm{pa}(X^p)$, $m \in [M]$.

Probabilistic Models [10]

Probabilistic Graphical Models:

- Estimate $P(Y, \boldsymbol{X})$
- Chain rule (probability):

$$P(Y, \boldsymbol{X}) = P(Y|\mathrm{pa}(Y)) \prod_{m=1}^{M} P(X^m|\mathrm{pa}(X^m)).$$

Extreme Cases:

- Discriminative models: $Y \notin \mathrm{pa}(X^m)$, $m \in [M]$
- Generative models: $\mathrm{pa}(Y) = \emptyset$ and $Y \in \mathrm{pa}(X^p)$, $m \in [M]$.

Model Families:

- How to encode/parametrize $P(Y|\mathrm{pa}(Y))$ and $P(X^m|\mathrm{pa}(X^m))$.
- How to estimate $P(Y, \boldsymbol{X})$ from training data.

Credal (Imprecise) Graphical Models

Basic setup

- A set of features $\boldsymbol{X} = \{X^1, \ldots, X^M\}$
- A class variable Y whose outcome $y \in \mathscr{Y}$

Credal Models:

- $\mathscr{P} := \{P(Y, \boldsymbol{X}) | P$ is compatible with knowledge/data$\}$
- Chain rule (probability):

$$P(Y, \boldsymbol{X}) = P(Y|\text{pa}(Y)) \prod_{m=1}^{M} P(X^m|\text{pa}(X^m)).$$

Extreme Cases:

- Discriminative models: $Y \notin \text{pa}(X^m)$, $m \in [M] := \{1, \ldots, M\}$
- Generative models: $\text{pa}(Y) = \emptyset$ and $Y \in \text{pa}(X^m)$, $m \in [M]$.

Model Families:

- How to encode/parametrize $P(Y|\text{pa}(Y))$ and $P(X^m|\text{pa}(X^m))$.
- How to estimate $\mathscr{P}(Y, \boldsymbol{X})$ from training data.

Beyond Multi-Class Classification

Other predictive tasks:

- Multi-Label Classification
- Multi-Dimensional Classification
- Multi-Target Prediction

Beyond Multi-Class Classification

Other predictive tasks:

- Multi-Label Classification
- Multi-Dimensional Classification
- Multi-Target Prediction

Examples:

- Predict multiple diseases (yes, no) given ChetXray and Demographic information.
- Predict antimicrobial resistance (AMR) phenotypes (susceptible, intermediate, resistant) of multiple drugs given genomic sequences of the strain.
- Predict multiple characteristics of the object that appears in each grid cell: object (no, pedestrian, car, bicycle, and so on), moving (no, forward, backward, left, right), attention (yes, no), and so on.

Probabilistic Models [10]

Probabilistic Graphical Models:

- Estimate $P(\boldsymbol{Y}, \boldsymbol{X})$, where $\boldsymbol{Y} = \{Y^1, \dots, Y^K\}$
- Chain rule (probability):

$$P(\boldsymbol{Y}, \boldsymbol{X}) = \prod_{k=1}^{K} P(Y^k | \mathrm{pa}(Y^k)) \prod_{m=1}^{M} P(X^m | \mathrm{pa}(X^m)).$$

Extreme Cases:

- Discriminative models: $Y^k \notin \mathrm{pa}(X^m)$, $k \in [K]$ and $m \in [M]$.
- Generative models: $\mathrm{pa}(Y^k) \cap \boldsymbol{X} = \emptyset$, $k \in [K]$ and $m \in [M]$.

Probabilistic Models [10]

Probabilistic Graphical Models:

- Estimate $P(\boldsymbol{Y}, \boldsymbol{X})$, where $\boldsymbol{Y} = \{Y^1, \ldots, Y^K\}$
- Chain rule (probability):

$$P(\boldsymbol{Y}, \boldsymbol{X}) = \prod_{k=1}^{K} P(Y^k | \text{pa}(Y^k)) \prod_{m=1}^{M} P(X^m | \text{pa}(X^m)).$$

Extreme Cases:

- Discriminative models: $Y^k \notin \text{pa}(X^m)$, $k \in [K]$ and $m \in [M]$.
- Generative models: $\text{pa}(Y^k) \cap \boldsymbol{X} = \emptyset$, $k \in [K]$ and $m \in [M]$.

Model Families:

- How to encode/parametrize $P(Y^k | \text{pa}(Y^k))$ and $P(X^m | \text{pa}(X^m))$.
- How to estimate $P(\boldsymbol{Y}, \boldsymbol{X})$ from training data.

Credal (Imprecise) Graphical Models [5]

Basic setup

- A set of features $\boldsymbol{X} = \{X^1, \dots, X^M\}$
- A set of class variables Y^k, whose outcome $y \in \mathscr{Y}$, $k = 1, \dots, K$

Credal Models:

- $\mathscr{P} := \{P(\boldsymbol{Y}, \boldsymbol{X}) | P$ is compatible with knowledge/data$\}$
- Chain rule (probability):

$$P(\boldsymbol{Y}, \boldsymbol{X}) = \prod_{k=1}^{K} P(Y^k | \mathrm{pa}(Y^k)) \prod_{m=1}^{M} P(X^m | \mathrm{pa}(X^m)).$$

Extreme Cases:

- Discriminative models: $Y^k \notin \mathrm{pa}(X^m)$, $k \in [K]$ and $m \in [M]$.
- Generative models: $\mathrm{pa}(Y^k) \cap \boldsymbol{X} = \emptyset$, $k \in [K]$ and $m \in [M]$.

Model Families:

- How to encode/parametrize $P(Y^k | \mathrm{pa}(Y^k))$ and $P(X^m | \mathrm{pa}(X^m))$.
- How to estimate $\mathscr{P}(\boldsymbol{Y}, \boldsymbol{X})$ from training data.

Other Families of Graphical Models

Sets of desirable gambles

↓

Interval expectation -·-·- **Credal sets**

↓

Interval probability

↓

2-Monotone capacities -·-·- Probability intervals

↓

Belief functions -·-·- Random sets

↙ ↘

Probability

Possibility -·-·- Fuzzy sets -·-·- Rough sets

References I

[1] J. Abellán and S. Moral.
Building classification trees using the total uncertainty criterion.
International Journal of Intelligent Systems, 18(12):1215–1225, 2003.

[2] A. Antonucci, G. Corani, and S. Bernaschina.
Active learning by the naive credal classifier.
In *Proceedings of the Sixth European Workshop on Probabilistic Graphical Models (PGM)*, pages 3–10, 2012.

[3] G. Corani and M. Zaffalon.
Learning reliable classifiers from small or incomplete data sets: The naive credal classifier 2.
Journal of Machine Learning Research, 9(4), 2008.

[4] V. G. Costa and C. E. Pedreira.
Recent advances in decision trees: An updated survey.
Artificial Intelligence Review, 56(5):4765–4800, 2023.

[5] F. G. Cozman.
Credal networks.
Artificial intelligence, 120(2):199–233, 2000.

[6] N. Friedman, D. Geiger, and M. Goldszmidt.
Bayesian network classifiers.
Machine learning, 29:131–163, 1997.

[7] L. V. Jospin, H. Laga, F. Boussaid, W. Buntine, and M. Bennamoun.
Hands-on bayesian neural networksa tutorial for deep learning users.
IEEE Computational Intelligence Magazine, 17(2):29–48, 2022.

References II

[8] C. J. Mantas and J. Abellán.
Analysis and extension of decision trees based on imprecise probabilities: Application on noisy data.
Expert Systems with Applications, 41(5):2514–2525, 2014.

[9] C. J. Mantas and J. Abellan.
Credal-c4. 5: Decision tree based on imprecise probabilities to classify noisy data.
Expert Systems with Applications, 41(10):4625–4637, 2014.

[10] V.-L. Nguyen, Y. Yang, and C. P. de Campos.
Probabilistic multi-dimensional classification.
In *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 1522–1533, 2023.

[11] V.-L. Nguyen, H. Zhang, and S. Destercke.
Credal ensembling in multi-class classification.
Machine Learning, 114(1):1–62, 2025.

[12] C. Ortega Vázquez, S. vanden Broucke, and J. De Weerdt.
Hellinger distance decision trees for pu learning in imbalanced data sets.
Machine Learning, pages 1–32, 2023.

[13] M. C. Troffaes.
Decision making under uncertainty using imprecise probabilities.
International journal of approximate reasoning, 45(1):17–29, 2007.