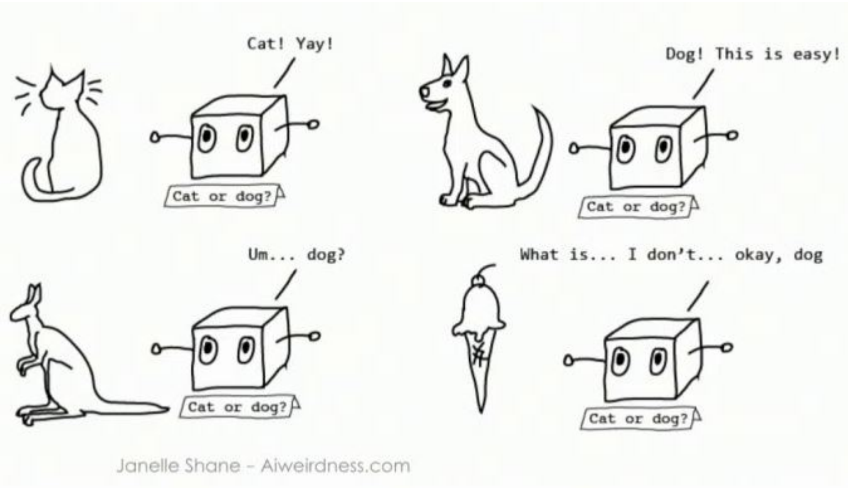

Cat! Yay!
Cat or dog?
Dog! This is easy!
Cat or dog?
Um... dog?
Cat or dog?
What is... I don't... okay, dog
Cat or dog?
Janelle Shane - Aiweirdness.com

Uncertainty reasoning and machine learning

Uncertainty, Decision and Evaluation in Machine Learning

Vu-Linh Nguyen

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

AOS4 master courses

Who is more reliable?

An example: Assume we travel to a small village

- There are **two doctors** who can give suggestion on whether a patient suffers from at least one type of serious cancers.
- Either "yes (y)" or "don't know (y/n)" $\longrightarrow$ go to the closest hospital for further diagnosis
- People ask you "who is more reliable?" given historical record on 1000 patients.

True situations	50 y	50 y	400 n	500 n
Dr. A's predictions	50 y	**50 n**	400 n	400 n + 100 y
Dr. B's predictions	50 y	**40 y/n + 10 n**	400 n	400 n + 100 y

Which model is more reliable?

Another example: Assume we travel to another village

- There are **3 pre-trained models** which can give suggestion on whether a patient suffers from at least one type of serious cancers.
- Either "yes (y)" or "don't know (y/n)" ⟶ go to the closest hospital for further diagnosis
- People ask you "which model is more reliable?" given historical record on 1000 patients.

True situations	50 y	50 y	400 n	500 n
C's predictions	50 y	**50 n**	400 n	**400 n + 100 y**
D's predictions	50 y	**40 y/n + 10 n**	400 n	**400 n + 100 y**
E's predictions	50 y	**40 y/n + 10 n**	400 n	**450 n + 50 y/n**

Go beyond the predictive performance?

It might be safer to defer our answer until we know more about

- how the **models** were learned and make their predictions
- how robust their predictions are (under the presence of noise)
- the decision-making process (cost, consequence, etc.)
- ...

Objectives

After this lecture students should be able to

- conceptually describe the Imprecise Dirichlet model (IDM) [1]
- use IDM in K-nn classifiers with fixed windows [6]
- evaluate classifiers based on IDM and related models [2, 7]

Outline

- Inference from Multinomial Data
- Imprecise Dirichlet Model (IDM)
- Applications in Classification Tasks
- Evaluate Classifiers
- Summary and Outlook

Basic setup:

- Univariate discrete variable V
- A finite set of possible outcomes $v \in \mathcal{V}$
- Each possible outcome is assigned a **probability value** $\theta_v := P(V = v) = P(\{v\})$

Basic setup:

- Univariate discrete variable V
- A finite set of possible outcomes $v \in \mathcal{V}$
- Each possible outcome is assigned a **probability value** $\theta_v := P(V = v) = P(\{v\})$

Questions

- How to model and estimate θ_v?
- How to do inference?
- How to handle small data?
- How to handle missing/partial data?

Frequentist, Bayesian and Imprecise approaches

Axioms

1. Positive: $\theta_v \geq 0$ for all outcomes $v \in \mathcal{V}$
2. Additive: $P(S) = \sum_{v \in S} \theta_v$ for all events $S \subseteq \mathcal{V}$
3. Normed: $P(\mathcal{V}) = 1$

Frequentist, Bayesian and Imprecise approaches

Axioms

1. Positive: $\theta_v \geq 0$ for all outcomes $v \in \mathcal{V}$
2. Additive: $P(S) = \sum_{v \in S} \theta_v$ for all events $S \subseteq \mathcal{V}$
3. Normed: $P(\mathcal{V}) = 1$

Three approaches (discussed in this lecture):

4F. **Frequentist**: $\boldsymbol{\theta} = \{\theta_v | v \in \mathcal{V}\}$ is **not** a random variable (VR).

4B. **Bayesian**: $\boldsymbol{\theta} = \{\theta_v | v \in \mathcal{V}\}$ is a RV $\longleftarrow$ **prior uncertainty** (PU) is described by **a distribution**.

4I. **Imprecise**: $\boldsymbol{\theta} = \{\theta_v | v \in \mathcal{V}\}$ is a RV $\longleftarrow$ PU is described by **a set of distribution** $\boldsymbol{\theta} \in \Theta$.

Some Inference Problems

Multinomial data:

- Given the observed data $\boldsymbol{D}$ where v appear n_v times, $v \in \mathcal{V}$:
- Let $n = \sum_v n_v$ and $\boldsymbol{n} = \{n_v | v \in \mathcal{V}\}$

Multinomial likelihood:

- $\propto$: is proportional to.
- $L(\boldsymbol{\theta}|\boldsymbol{D}) \propto \prod_{v \in \mathcal{V}} (\theta_v)^{n_v}$.

Make inferences about

- the **unknown $\boldsymbol{\theta}$**
- some derived parameter of interest $g(\boldsymbol{\theta})$
- future observations $\boldsymbol{D}'$

(Few) Potential Applications

Multinomial data:

- Given the observed data $\boldsymbol{D}$ where v appear n_v times, $v \in \mathcal{V}$:
- Let $n = \sum_v n_v$ and $\boldsymbol{n} = \{n_v | v \in \mathcal{V}\}$
- Multinomial likelihood: $L(\boldsymbol{\theta}|\boldsymbol{D}) \propto \prod_{x \in \mathcal{V}} (\theta_v)^{n_v}$.

Make inferences about

- the **unknown** $\boldsymbol{\theta}$, e.g., its best estimate $\boldsymbol{\theta}^*$
- some derived parameter of interest $g(\boldsymbol{\theta})$

(Few) Potential Applications

Multinomial data:

- Given the observed data $\boldsymbol{D}$ where v appear n_v times, $v \in \mathcal{V}$:
- Let $n = \sum_v n_v$ and $\boldsymbol{n} = \{n_v | v \in \mathcal{V}\}$
- Multinomial likelihood: $L(\boldsymbol{\theta}|\boldsymbol{D}) \propto \prod_{x \in \mathcal{V}} (\theta_v)^{n_v}$.

Make inferences about

- the **unknown** $\boldsymbol{\theta}$, e.g., its best estimate $\boldsymbol{\theta}^*$
- some derived parameter of interest $g(\boldsymbol{\theta})$

You would find such a problem in

- **Parzen window classifiers**
- (Credal) Decision trees, Naive Bayesian/credal Classifier (Lecture 4)
- Ensembles (Trees, Neural Nets, etc.)
- Bayesian Neural Nets

Outline

Outline

Frequentist (Recap)

Axioms

1. Positive: $\theta_v \geq 0$ for all outcomes $v \in \mathcal{V}$
2. Additive: $P(S) = \sum_{v \in V} \theta_v$ for all events $S \subseteq \mathcal{V}$
3. Normed: $P(\mathcal{V}) = 1$

4F. **Frequentist**: $\boldsymbol{\theta} = \{\theta_v | v \in \mathcal{V}\}$ is **not** a random variable (VR).

Frequentist (Recap)

Axioms

1. Positive: $\theta_v \geq 0$ for all outcomes $v \in \mathcal{V}$
2. Additive: $P(S) = \sum_{v \in V} \theta_v$ for all events $S \subseteq \mathcal{V}$
3. Normed: $P(\mathcal{V}) = 1$

4F. **Frequentist**: $\boldsymbol{\theta} = \{\theta_v | v \in \mathcal{V}\}$ is **not** a random variable (VR).

Estimate $\boldsymbol{\theta}$:

- **Frequencies**: Maximum likelihood estimation (MLE) gives $\theta_v^* = {}^{n_v}/_{n}$

Frequentist: Comments

- Does not take into account the **importance of sample size** $\longleftarrow$ **Sources of uncertainty**!

Frequentist: Comments

- Does not take into account the **importance of sample size** $\longleftarrow$ **Sources of uncertainty**!

Coin	Small	Large
Flips	2	$2 \cdot 10^6$
Heads	50%	50%
Tails	50%	50%

- For both coins, a frequentist says

$$\theta^*_{\text{Head}} = \theta^*_{\text{Tail}} = 1/2$$

Frequentist: Comments

- Does not take into account the **importance of sample size** $\longleftarrow$ **Sources of uncertainty**!

Coin	Small	Large
Flips	2	$2 \cdot 10^6$
Heads	50%	50%
Tails	50%	50%

- For both coins, a frequentist says

$$\theta^*_{\text{Head}} = \theta^*_{\text{Tail}} = 1/2$$

- What can you say about the reliability of the estimate for each coin?

Frequentist: Comments

- Does not take into account the **importance of sample size** $\longleftarrow$ **Sources of uncertainty**!

Coin	Small	Large
Flips	2	$2 \cdot 10^6$
Heads	50%	50%
Tails	50%	50%

- For both coins, a frequentist says
$$\theta^*_{\text{Head}} = \theta^*_{\text{Tail}} = 1/2$$
- What can you say about the reliability of the estimate for each coin?

Coin	Small	Large
Flips	2	$2 \cdot 10^6$
Heads	0%	0%
Tails	100%	100%

- For both coins, a frequentist says
$$\theta^*_{\text{Head}} = 0 \text{ and } \theta^*_{\text{Tail}} = 1$$

Frequentist: Comments

- Does not take into account the **importance of sample size** $\longleftarrow$ **Sources of uncertainty**!

Coin	Small	Large
Flips	2	$2 \cdot 10^6$
Heads	50%	50%
Tails	50%	50%

- For both coins, a frequentist says
$$\theta^*_{\text{Head}} = \theta^*_{\text{Tail}} = {}^1\!/_2$$
- What can you say about the reliability of the estimate for each coin?

Coin	Small	Large
Flips	2	$2 \cdot 10^6$
Heads	0%	0%
Tails	100%	100%

- For both coins, a frequentist says
$$\theta^*_{\text{Head}} = 0 \text{ and } \theta^*_{\text{Tail}} = 1$$
- What can you say about the reliability of the estimate for each coin?

Frequentist: Comments (Cont.)

- Does not (naturally) take into account missing/partial data

Frequentist: Comments (Cont.)

- Does not (naturally) take into account missing/partial data

Coin	Small	Large
Flips	2	$2 \cdot 10^6$
Heads	$[0,1]$	$[5,10]$
Tails	$[1,2]$	$[5, 2 \cdot 10^6]$

- Can we use frequencies to estimate θ^*_{Head} and θ^*_{Tail}?
- What can you say about the reliability of the estimate for each coin?

Bayesian (Recap)

Axioms

1. Positive: $\theta_v \geq 0$ for all outcomes $v \in \mathcal{V}$
2. Additive: $P(S) = \sum_{v \in V} \theta_v$ for all events $S \subseteq \mathcal{V}$
3. Normed: $P(\mathcal{V}) = 1$

4B. **Bayesian**: $\boldsymbol{\theta} = \{\theta_v | v \in \mathcal{V}\}$ is a RV $\longleftarrow$ **prior uncertainty** (PU) is described by **a distribution**.

Bayesian (Recap)

Axioms

1. Positive: $\theta_v \geq 0$ for all outcomes $v \in \mathcal{V}$
2. Additive: $P(S) = \sum_{v \in V} \theta_v$ for all events $S \subseteq \mathcal{V}$
3. Normed: $P(\mathcal{V}) = 1$

4B. **Bayesian**: $\boldsymbol{\theta} = \{\theta_v | v \in \mathcal{V}\}$ is a RV $\longleftarrow$ **prior uncertainty** (PU) is described by **a distribution**.

Bayesian estimates:

- **posterior mean** θ_v^* of θ_v: $E(\theta_v)$
- **posterior mean** $\theta_v^* | \boldsymbol{D}$ of $\theta_v | \boldsymbol{D}$: $E(\theta_v | \boldsymbol{D})$

Bayesian (Recap)

Axioms

1. Positive: $\theta_v \geq 0$ for all outcomes $v \in \mathcal{V}$
2. Additive: $P(S) = \sum_{v \in V} \theta_v$ for all events $S \subseteq \mathcal{V}$
3. Normed: $P(\mathcal{V}) = 1$

4B. **Bayesian**: $\boldsymbol{\theta} = \{\theta_v | v \in \mathcal{V}\}$ is a RV $\longleftarrow$ **prior uncertainty** (PU) is described by **a distribution**.

Bayesian estimates:

- **posterior mean** θ_v^* of θ_v: $E(\theta_v)$
- **posterior mean** $\theta_v^* | \boldsymbol{D}$ of $\theta_v | \boldsymbol{D}$: $E(\theta_v | \boldsymbol{D})$
- We can also use **posterior mode**

Dirichlet Model

Prior uncertainty: $\boldsymbol{\theta} \sim \text{Diri}(\boldsymbol{\alpha}) = \text{Diri}(s\boldsymbol{f})$

- Prior strengths (hyperparameter): α_v, $v \in \mathcal{V}$
- Total strength (hyperparameter): $s := \sum_{v \in \mathcal{V}} \alpha_v$
- Prior frequencies: $\boldsymbol{f} := \{f_v | v \in \mathcal{V}\}$ with $f_v := {\alpha_v}/{s}$, $v \in \mathcal{V}$
- $\theta_v \sim \text{Beta}(sf_v, s\sum_{v' \neq v} f_{v'})$
- $\boldsymbol{\theta}|\boldsymbol{D} \sim \text{Diri}(\boldsymbol{n} + \boldsymbol{\alpha}) = \text{Diri}(\boldsymbol{n} + s\boldsymbol{f})$
- $\theta_x|\boldsymbol{D} \sim \text{Beta}(n_v + sf_v, \sum_{v' \neq v} n_{v'} + s\sum_{v' \neq v} f_{v'})$

Dirichlet Model

Prior uncertainty: $\boldsymbol{\theta} \sim \text{Diri}(\boldsymbol{\alpha}) = \text{Diri}(s\boldsymbol{f})$

- Prior strengths (hyperparameter): α_v, $v \in \mathcal{V}$
- Total strength (hyperparameter): $s := \sum_{v \in \mathcal{V}} \alpha_v$
- Prior frequencies: $\boldsymbol{f} := \{f_v | v \in \mathcal{V}\}$ with $f_v := \alpha_v / s$, $v \in \mathcal{V}$
- $\theta_v \sim \text{Beta}(sf_v, s\sum_{v' \neq v} f_{v'})$
- $\boldsymbol{\theta}|\boldsymbol{D} \sim \text{Diri}(\boldsymbol{n} + \boldsymbol{\alpha}) = \text{Diri}(\boldsymbol{n} + s\boldsymbol{f})$
- $\theta_x|\boldsymbol{D} \sim \text{Beta}(n_v + sf_v, \sum_{v' \neq v} n_{v'} + s\sum_{v' \neq v} f_{v'})$

Bayesian estimates:

- **posterior mean** θ_v^* of θ_v: $E(\theta_v) = f_v$
- **posterior mean** $\theta_v^*|\boldsymbol{D}$ of $\theta_v|\boldsymbol{D}$:

$$E(\theta_k|\boldsymbol{D}) = (n_v + \alpha_v)/(n+s) = (n_v + sf_v)/(n+s)$$

Dirichlet Model: Hyperparameters

Solutions for fixed n are usually **symmetric Dirichlet priors**

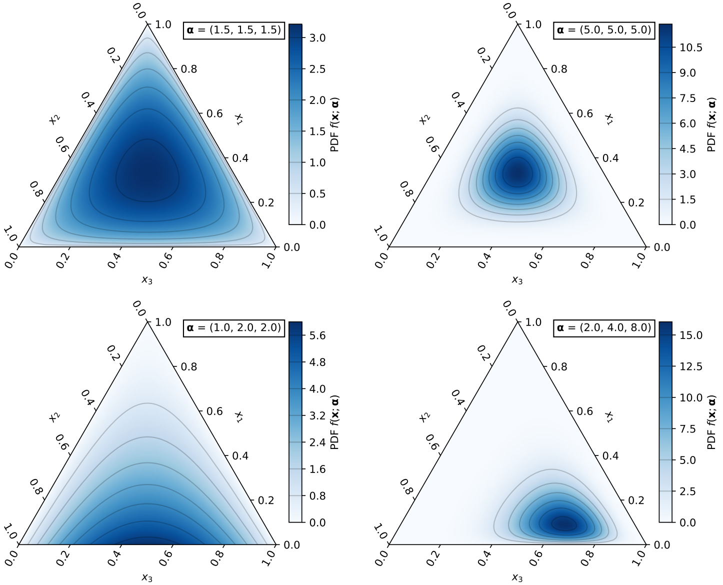
- Prior frequencies: $f_v = 1/|\mathcal{V}|$, $v \in \mathcal{V}$
- Total strength: $s = g'(|\mathcal{V}|)$

Dirichlet Model: Hyperparameters

Solutions for fixed *n* are usually **symmetric Dirichlet priors**

- Prior frequencies: $f_v = 1/|\mathcal{V}|$, $v \in \mathcal{V}$
- Total strength: $s = g'(|\mathcal{V}|)$

Advocators	α_v	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	$\lvert\mathcal{V}\rvert$


α = (1.5, 1.5, 1.5)
α = (5.0, 5.0, 5.0)
α = (1.0, 2.0, 2.0)
α = (2.0, 4.0, 8.0)
PDF f(x; α)
x_1
x_2
x_3

The Importance of Sample Size

Coin	Small	Large
Flips	2	$2\cdot10^6$
Heads	50%	50%
Tails	50%	50%

- For both coins, a frequentist says

$$p_{\text{Heads}} = p_{\text{Tails}} = 1/2$$

- Do Bayesians say the same thing? ⟵
  Yes!

Coin	Small	Large
Flips	4	$4\cdot10^6$
Heads	25%	25%
Tails	75%	75%

- For both coins, a frequentist says

$$p_{\text{Heads}} = 0.25, p_{\text{Tails}} = 0.75$$

- Do Bayesians say the same thing?

The Importance of Sample Size

Coin	Small	Large
Flips	2	$2 \cdot 10^6$
Heads	50%	50%
Tails	50%	50%

- For both coins, a frequentist says

$$p_{\text{Heads}} = p_{\text{Tails}} = 1/2$$

- Do Bayesians say the same thing? ⟵
  Yes!

Coin	Small	Large
Flips	4	$4 \cdot 10^6$
Heads	25%	25%
Tails	75%	75%

- For both coins, a frequentist says

$$p_{\text{Heads}} = 0.25, p_{\text{Tails}} = 0.75$$

- Do Bayesians say the same thing?

Advocators	α_v	s	p^S_{H}	p^S_{T}	p^L_{H}	p^L_{T}
Haldane (1948)	0	0	0.25	0.75	0.25	0.75
Perks (1947)	$1/\lvert\mathcal{V}\rvert$	1	0.3	0.7	0.25	0.75
Jeffreys (1946, 1961)	$1/2$	$\lvert\mathcal{V}\rvert/2$	0.3	0.7	0.25	0.75
Bayes-Laplace	1	$\lvert\mathcal{V}\rvert$	0.33	0.67	0.25	0.75

The Importance of Sample Size (Cont.)

Coin	Small	Large
Flips	2	$2 \cdot 10^6$
Heads	0%	0%
Tails	100%	100%

- For both coins, a frequentist says

$$p_{\text{Heads}} = 0\,, p_{\text{Tails}} = 1$$

- Do Bayesians say the same thing?

The Importance of Sample Size (Cont.)

Coin	Small	Large
Flips	2	$2 \cdot 10^6$
Heads	0%	0%
Tails	100%	100%

- For both coins, a frequentist says

$$p_{\text{Heads}} = 0, p_{\text{Tails}} = 1$$

- Do Bayesians say the same thing?

Advocators	α_x	s	p_{H}^S	p_{T}^S	p_{H}^L	p_{T}^L
Haldane (1948)	0	0	0	1	0	1
Perks (1947)	$1/\lvert\mathcal{V}\rvert$	1	0.17	0.83	$3 \cdot 10^{-7}$	$1 - 3 \cdot 10^{-7}$
Jeffreys	$1/\lvert\mathcal{V}\rvert$	1	0.17	0.83	$3 \cdot 10^{-7}$	$1 - 3 \cdot 10^{-7}$
Bayes-Laplace	1	$\lvert\mathcal{V}\rvert$	0.25	0.75	$5 \cdot 10^{-7}$	$1 - 5 \cdot 10^{-7}$

Dirichlet Model (DM): Comments

- Does not (naturally) take into account missing/partial data

Dirichlet Model (DM): Comments

- Does not (naturally) take into account missing/partial data

Coin	Small	Large
Flips	2	$2 \cdot 10^6$
Heads	$[0,1]$	$[5,10]$
Tails	$[1,2]$	$[5, 2 \cdot 10^6]$

- Can we use DM to estimate θ^*_{Head} and θ^*_{Tail}?
- What can you say about the reliability of the estimate for each coin?

Outline

- Inference from Multinomial Data
- Imprecise Dirichlet Model (IDM)
  - Frequentist and Bayesian Approaches
  - Imprecise Dirichlet Model
- Applications in Classification Tasks
- Evaluate Classifiers
- Summary and Outlook

Imprecise (Recap)

Axioms

1. Positive: $\theta_v \geq 0$ for all outcomes $v \in \mathcal{V}$
2. Additive: $P(S) = \sum_{v \in V} \theta_v$ for all events $S \subseteq \mathcal{V}$
3. Normed: $P(\mathcal{V}) = 1$

4I. **Imprecise**: $\boldsymbol{\theta} = \{\theta_v | v \in \mathcal{V}\}$ is a RV $\longleftarrow$ **prior uncertainty** (PU) is described by **a set of distribution $\boldsymbol{\theta} \in \boldsymbol{\Theta}$**.

Imprecise (Recap)

Axioms

1. Positive: $\theta_v \geq 0$ for all outcomes $v \in \mathcal{V}$
2. Additive: $P(S) = \sum_{v \in V} \theta_v$ for all events $S \subseteq \mathcal{V}$
3. Normed: $P(\mathcal{V}) = 1$

4I. **Imprecise**: $\boldsymbol{\theta} = \{\theta_v | v \in \mathcal{V}\}$ is a RV $\longleftarrow$ **prior uncertainty** (PU) is described by **a set of distribution** $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

Interval estimates:

- **posterior mean** θ_v^* of θ_v:

$$E(\theta_v) \in [\underline{E}(\theta_v), \overline{E}(\theta_v)]$$

- **posterior mean** $\theta_v^* | \boldsymbol{D}$ of $\theta_v | \boldsymbol{D}$:

$$E(\theta_v | \boldsymbol{D}) \in [\underline{E}(\theta_v | \boldsymbol{D}), \overline{E}(\theta_v | \boldsymbol{D})]$$

Imprecise Dirichlet Model

Prior uncertainty: $\boldsymbol{\Theta} = \{\boldsymbol{\theta} \sim \text{Diri}(\boldsymbol{\alpha}) = \text{Diri}(s\boldsymbol{f}) | \sum_{v \in \mathcal{V}} \alpha_v = s\}$

- Hyperparameter: s = **degree of imprecision** in the inferences
- Prior frequencies: $\boldsymbol{f} := \{f_v | v \in \mathcal{V}\}$ with $f_v := \alpha_v / s$, $v \in \mathcal{V}$
- $\theta_v \sim \text{Beta}(sf_v, s\sum_{v' \neq v} f_{v'})$
- $\boldsymbol{\theta} | \boldsymbol{D} \sim \text{Diri}(\boldsymbol{n} + \boldsymbol{\alpha}) = \text{Diri}(\boldsymbol{n} + s\boldsymbol{f})$
- $\theta_x | \boldsymbol{D} \sim \text{Beta}(n_v + sf_v, \sum_{v' \neq v} n_{v'} + s\sum_{v' \neq v} f_{v'})$

Imprecise Dirichlet Model

Prior uncertainty: $\boldsymbol{\Theta} = \{\boldsymbol{\theta} \sim \text{Diri}(\boldsymbol{\alpha}) = \text{Diri}(s\boldsymbol{f}) | \sum_{v \in \mathcal{V}} \alpha_v = s\}$

- Hyperparameter: s = **degree of imprecision** in the inferences
- Prior frequencies: $\boldsymbol{f} := \{f_v | v \in \mathcal{V}\}$ with $f_v := \alpha_v / s$, $v \in \mathcal{V}$
- $\theta_v \sim \text{Beta}(sf_v, s\sum_{v' \neq v} f_{v'})$
- $\boldsymbol{\theta}|\boldsymbol{D} \sim \text{Diri}(\boldsymbol{n} + \boldsymbol{\alpha}) = \text{Diri}(\boldsymbol{n} + s\boldsymbol{f})$
- $\theta_x|\boldsymbol{D} \sim \text{Beta}(n_v + sf_v, \sum_{v' \neq v} n_{v'} + s\sum_{v' \neq v} f_{v'})$

Posterior mean $\theta_v^*|\boldsymbol{D}$ of $\theta_v|\boldsymbol{D}$:

$$E(\theta_v|\boldsymbol{D}) \in [\underline{E}(\theta_v|\boldsymbol{D}), \overline{E}(\theta_v|\boldsymbol{D})], \quad (1)$$

$$\underline{E}(\theta_v|\boldsymbol{D}) = n_v / (n+s), \quad (2)$$

$$\overline{E}(\theta_v|\boldsymbol{D}) = (n_v+s) / (n+s). \quad (3)$$

The Importance of Sample Size

Coin	Small	Large
Flips	2	$2 \cdot 10^6$
Heads	50%	50%
Tails	50%	50%

- For both coins, a frequentist says

$$\theta_{\text{Heads}} = \theta_{\text{Tails}} = {}^1\!/_2$$

- Bayesians would say the same thing
- Would IDM say the same thing?

The Importance of Sample Size

Coin	Small	Large
Flips	2	$2 \cdot 10^6$
Heads	50%	50%
Tails	50%	50%

- For both coins, a frequentist says

$$\theta_{\text{Heads}} = \theta_{\text{Tails}} = {}^1\!/_2$$

- Bayesians would say the same thing
- Would IDM say the same thing?

	$\underline{P}^S_{\text{H}}$	$\overline{P}^S_{\text{H}}$	$\underline{P}^L_{\text{H}}$	$\overline{P}^L_{\text{H}}$
$s = 1$	0.33	0.67	$0.5 - 3 \cdot 10^{-7}$	$0.5 + 3 \cdot 10^{-7}$
$s = 2$	0.25	0.75	$0.5 - 5 \cdot 10^{-7}$	$0.5 + 5 \cdot 10^{-7}$

The Importance of Sample Size (Cont.)

Coin	Small	Large
Flips	2	$2 \cdot 10^6$
Heads	0%	0%
Tails	100%	100%

- For both coins, a frequentist says

$$\theta_{\text{Heads}} = 0, \theta_{\text{Tails}} = 1$$

- Bayesians would say different things
- What would IDM say?

Advocators	α_x	s	p_{H}^S	p_{T}^S	p_{H}^L	p_{T}^L
Haldane (1948)	0	0	0	1	0	1
Perks (1947)	$1/\lvert\mathcal{V}\rvert$	1	0.17	0.83	$3 \cdot 10^{-7}$	$1 - 3 \cdot 10^{-7}$
Jeffreys	$1/\lvert\mathcal{V}\rvert$	1	0.17	0.83	$3 \cdot 10^{-7}$	$1 - 3 \cdot 10^{-7}$
Bayes-Laplace	1	$\lvert\mathcal{V}\rvert$	0.25	0.75	$5 \cdot 10^{-7}$	$1 - 5 \cdot 10^{-7}$

The Importance of Sample Size (Cont.)

Coin	Small	Large
Flips	2	$2 \cdot 10^6$
Heads	0%	0%
Tails	100%	100%

- For both coins, a frequentist says

$$\theta_{\text{Heads}} = 0, \theta_{\text{Tails}} = 1$$

- Bayesians would say different things
- What would IDM say?

Advocators	α_x	s	p_{H}^S	p_{T}^S	p_{H}^L	p_{T}^L
Haldane (1948)	0	0	0	1	0	1
Perks (1947)	$1/\lvert\mathcal{V}\rvert$	1	0.17	0.83	$3 \cdot 10^{-7}$	$1 - 3 \cdot 10^{-7}$
Jeffreys	$1/\lvert\mathcal{V}\rvert$	1	0.17	0.83	$3 \cdot 10^{-7}$	$1 - 3 \cdot 10^{-7}$
Bayes-Laplace	1	$\lvert\mathcal{V}\rvert$	0.25	0.75	$5 \cdot 10^{-7}$	$1 - 5 \cdot 10^{-7}$

IDM	$\underline{P}_{\text{H}}^S$	$\overline{P}_{\text{H}}^S$	$\underline{P}_{\text{H}}^L$	$\overline{P}_{\text{H}}^L$
$s = 1$	0	0.33	0	$5 \cdot 10^{-7}$
$s = 2$	0	0.50	0	10^{-6}

The case of Partial/Missing Data

What if we only know $n_v \in \boldsymbol{n}_v \subset \{0, 1, \ldots, n\}$?

The case of Partial/Missing Data

What if we only know $n_v \in \boldsymbol{n}_v \subset \{0, 1, \dots, n\}$?

- Imprecise approaches provide nice tools to handle such data sets [6]
- Uncertainty (due to the incompleteness) is described by a set of **possible** precise data sets $\mathscr{D} = \{\boldsymbol{D} | n_v \in \boldsymbol{n}_v, \sum_{v \in \mathcal{V}} n_v = n\}$

The case of Partial/Missing Data

What if we only know $n_v \in \boldsymbol{n}_v \subset \{0, 1, \ldots, n\}$?

- Imprecise approaches provide nice tools to handle such data sets [6]
- Uncertainty (due to the incompleteness) is described by a set of **possible** precise data sets $\mathscr{D} = \{\boldsymbol{D} | n_v \in \boldsymbol{n}_v, \sum_{v \in \mathcal{V}} n_v = n\}$

Interval posterior mean $\theta^*_v | \mathscr{D}$ of $\theta_v | \mathscr{D}$:

$$E(\theta_v|\mathscr{D}) \in [\underline{E}(\theta_v|\mathscr{D}), \overline{E}(\theta_v|\mathscr{D})], \quad (4)$$

$$\underline{E}(\theta_v|\mathscr{D}) = \min_{\boldsymbol{D} \in \mathscr{D}} \underline{E}(\theta_v|\boldsymbol{D}) = \min_{\boldsymbol{D} \in \mathscr{D}} n_v / (n+s), \quad (5)$$

$$\overline{E}(\theta_v|\mathscr{D}) = \max_{\boldsymbol{D} \in \mathscr{D}} \overline{E}(\theta_v|\boldsymbol{D}) = \max_{\boldsymbol{D} \in \mathscr{D}} (n_v+s) / (n+s). \quad (6)$$

Determine $\mathscr{D}$

Coin	Small	Large
Flips	2	$2 \cdot 10^6$
Heads	$[0,1]$	$[5,10]$
Tails	$[1,2]$	$[5, 2 \cdot 10^6]$

- Recap: $\mathscr{D} = \{\boldsymbol{D} | n_v \in \boldsymbol{n}_v, \sum_{v \in \mathcal{V}} n_v = n\}$
- What is $\mathscr{D}^S$ for the first coin?
- What is $\mathscr{D}^L$ for the second coin?

Determine $\mathscr{D}$

Coin	Small	Large
Flips	2	$2 \cdot 10^6$
Heads	$[0,1]$	$[5,10]$
Tails	$[1,2]$	$[5, 2 \cdot 10^6]$

- Recap: $\mathscr{D} = \{\boldsymbol{D} | n_v \in \boldsymbol{n}_v, \sum_{v \in \mathscr{V}} n_v = n\}$
- What is $\mathscr{D}^S$ for the first coin?
- What is $\mathscr{D}^L$ for the second coin?

Coin	Small	$\boldsymbol{D}_1$	$\boldsymbol{D}_2$
Flips	2	2	2
Heads	$[0,1]$	0	1
Tails	$[1,2]$	2	1

Determine $\mathscr{D}$

Coin	Small	Large
Flips	2	$2 \cdot 10^6$
Heads	$[0,1]$	$[5,10]$
Tails	$[1,2]$	$[5,2\cdot 10^6]$

- Recap: $\mathscr{D} = \{\boldsymbol{D} | n_v \in \boldsymbol{n}_v, \sum_{v \in \mathcal{V}} n_v = n\}$
- What is $\mathscr{D}^S$ for the first coin?
- What is $\mathscr{D}^L$ for the second coin?

Coin	Small	$\boldsymbol{D}_1$	$\boldsymbol{D}_2$
Flips	2	2	2
Heads	$[0,1]$	0	1
Tails	$[1,2]$	2	1

Coin	Large	$\boldsymbol{D}_1$	$\boldsymbol{D}_2$	$\boldsymbol{D}_3$	$\boldsymbol{D}_4$	$\boldsymbol{D}_5$	$\boldsymbol{D}_6$
Flips	$n = 2\cdot 10^6$	n	n	n	n	n	n
Heads	$[5,10]$	5	6	7	8	9	10
Tails	$[5,n]$	$n-5$	$n-6$	$n-7$	$n-8$	$n-9$	$n-10$

Compute Lower and Upper Expectations

Interval posterior mean $\theta_v^*|\mathscr{D}$ of $\theta_v|\mathscr{D}$:

$$\underline{E}(\theta_v|\mathscr{D}) = \min_{\boldsymbol{D}\in\mathscr{D}} \underline{E}(\theta_v|\boldsymbol{D}) = \min_{\boldsymbol{D}\in\mathscr{D}} n_v/(n+s)\,, \tag{7}$$

$$\overline{E}(\theta_v|\mathscr{D}) = \max_{\boldsymbol{D}\in\mathscr{D}} \overline{E}(\theta_v|\boldsymbol{D}) = \max_{\boldsymbol{D}\in\mathscr{D}} (n_v+s)/(n+s)\,. \tag{8}$$

Compute Lower and Upper Expectations

Interval posterior mean $\theta_v^*|\mathscr{D}$ of $\theta_v|\mathscr{D}$:

$$\underline{E}(\theta_v|\mathscr{D}) = \min_{\boldsymbol{D}\in\mathscr{D}} \underline{E}(\theta_v|\boldsymbol{D}) = \min_{\boldsymbol{D}\in\mathscr{D}} {}^{n_v}\!/\!_{(n+s)}, \quad (7)$$

$$\overline{E}(\theta_v|\mathscr{D}) = \max_{\boldsymbol{D}\in\mathscr{D}} \overline{E}(\theta_v|\boldsymbol{D}) = \max_{\boldsymbol{D}\in\mathscr{D}} {}^{(n_v+s)}\!/\!_{(n+s)}. \quad (8)$$

Coin	Small	$\boldsymbol{D}_1$	$\boldsymbol{D}_2$	$\underline{E}(\theta_v\|\mathscr{D})$	$\overline{E}(\theta_v\|\boldsymbol{D})$
Flips	2	2	2		
Heads	[0,1]	0	1	$^0/_{(2+s)}$	$^{(1+s)}/_{(2+s)}$
Tails	[1,2]	2	1	$^1/_{(2+s)}$	$^{(2+s)}/_{(2+s)}$

Compute Lower and Upper Expectations

Interval posterior mean $\theta_v^*|\mathscr{D}$ of $\theta_v|\mathscr{D}$:

$$\underline{E}(\theta_v|\mathscr{D}) = \min_{\boldsymbol{D}\in\mathscr{D}} \underline{E}(\theta_v|\boldsymbol{D}) = \min_{\boldsymbol{D}\in\mathscr{D}} {}^{n_v}\!/\!_{(n+s)}, \tag{7}$$

$$\overline{E}(\theta_v|\mathscr{D}) = \max_{\boldsymbol{D}\in\mathscr{D}} \overline{E}(\theta_v|\boldsymbol{D}) = \max_{\boldsymbol{D}\in\mathscr{D}} {}^{(n_v+s)}\!/\!_{(n+s)}. \tag{8}$$

Coin	Small	$\boldsymbol{D}_1$	$\boldsymbol{D}_2$	$\underline{E}(\theta_v\|\mathscr{D})$	$\overline{E}(\theta_v\|\boldsymbol{D})$
Flips	2	2	2		
Heads	$[0,1]$	0	1	$^0/_{(2+s)}$	$^{(1+s)}/_{(2+s)}$
Tails	$[1,2]$	2	1	$^1/_{(2+s)}$	$^{(2+s)}/_{(2+s)}$

Coin	Large	$\boldsymbol{D}_1$	…	$\boldsymbol{D}_6$	$\underline{E}(\theta_v\|\mathscr{D})$	$\overline{E}(\theta_v\|\boldsymbol{D})$
Flips	$n = 2\cdot 10^6$	n	…	n		
Heads	$[5,10]$	5	…	10	$^5/_{(n+s)}$	$^{(10+s)}/_{(n+s)}$
Tails	$[5,n]$	$n-5$	…	$n-10$	$^{(n-10)}/_{(n+s)}$	$^{(n-5+s)}/_{(n+s)}$

Outline

- Inference from Multinomial Data
- Imprecise Dirichlet Model (IDM)
- Applications in Classification Tasks
  - Optimal decision rules
  - Pazen Window Classifiers
- Evaluate Classifiers
- Summary and Outlook

Outline

- Inference from Multinomial Data
- Imprecise Dirichlet Model (IDM)
- Applications in Classification Tasks
  - Optimal decision rules
  - Pazen Window Classifiers
- Evaluate Classifiers
- Summary and Outlook

Frequentist approach

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 $\boldsymbol{x}$, $\boldsymbol{\theta}|\boldsymbol{x}$

Frequentist approach

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

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

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

Frequentist approach

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

- Let $\ell : \mathscr{Y} \times \mathscr{Y} \longmapsto \mathbb{R}_+$ be any loss function.
- The Bayes-optimal prediction of ℓ on $\boldsymbol{x}$ is
$$y_\ell^{\boldsymbol{\theta}} = \underset{\overline{y} \in \mathscr{Y}}{\operatorname{argmin}} \sum_{y \in \mathscr{Y}} \ell(\overline{y}, y) \theta_y | \boldsymbol{x}$$
- If ℓ is 0/1 loss, i.e. $\ell(\overline{y}, y) = \mathbb{1}(\overline{y} \neq y)$, then (Check!)
$$y_\ell^{\boldsymbol{\theta}} = \underset{\overline{y} \in \mathscr{Y}}{\operatorname{argmax}} \theta_{\overline{y}} | \boldsymbol{x}$$

Frequentist approach (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 $\boldsymbol{x}$, $\boldsymbol{\theta}|\boldsymbol{x}$

Frequentist approach (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 $\boldsymbol{x}$, $\boldsymbol{\theta}|\boldsymbol{x}$

Generalized optimal decision rules

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

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

Frequentist approach (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 $\boldsymbol{x}$, $\boldsymbol{\theta}|\boldsymbol{x}$

Generalized optimal decision rules

- Let $\mathscr{L} : 2^{\mathscr{Y}} \setminus \{\emptyset\} \times \mathscr{Y} \longmapsto \mathbb{R}_+$ be any loss function.
- The Bayes-optimal prediction of $\mathscr{L}$ on $\boldsymbol{x}$ is
$$Y_{\mathscr{L}}^{\boldsymbol{\theta}} = \operatorname*{argmin}_{\overline{Y} \subset \mathscr{Y}} \sum_{y \in \mathscr{Y}} \mathscr{L}(\overline{Y}, y)\theta_y|\boldsymbol{x}$$
- If $\mathscr{L}$ is the loss version of a utility-discounted accuracy
$$u_\alpha(\overline{Y}, y) = \mathbb{1}(y \in \overline{Y}) g_\alpha(|\overline{Y}|)$$

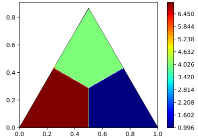
then (Check!) $Y_{\mathscr{L}}^{\boldsymbol{\theta}}$ consists of the most probable outcomes $y \in \mathscr{Y}$.

Illustrations

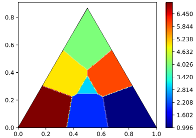
Basic setup and assumption

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

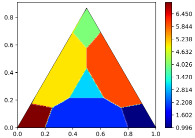
Optimal decision rules



0/1 loss



$u_{1.6}$



$u_{2.2}$

Imprecise approach

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 $\boldsymbol{x}$, $\boldsymbol{\Theta}|\boldsymbol{x}$

Imprecise approach

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 [4, 5]:

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

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

- Computation: Solving linear programs, etc.

Imprecise approach (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 $\boldsymbol{x}$, $\boldsymbol{\Theta}|\boldsymbol{x}$

Maximality Rule [4, 5]:

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

$$Y^M_{\ell,\boldsymbol{\Theta}|\boldsymbol{x}} = \{y \in \mathscr{Y} \mid \nexists y' \text{ s.t. } y' \succ_{\ell,\boldsymbol{\Theta}|\boldsymbol{x}} y\}.$$

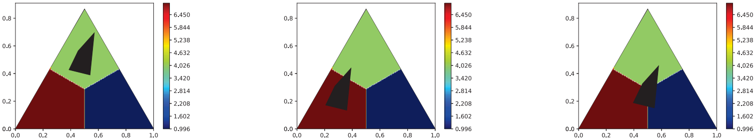
- Computation: Solving linear programs, Iterating over the extreme points of $\boldsymbol{\Theta}|\boldsymbol{x}$.

Illustrations

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 with 0/1 loss



Outline

- Inference from Multinomial Data
- Imprecise Dirichlet Model (IDM)
- Applications in Classification Tasks
  - Optimal decision rules
  - Pazen Window Classifiers
- Evaluate Classifiers
- Summary and Outlook

Pazen Window Classifiers [3]

Basic setup and assumption

- Given training data $\boldsymbol{D} \subset \mathscr{X} \times \mathscr{Y}$, a distance $d(\boldsymbol{x}, \boldsymbol{x}')$, and a threshold ϵ
- For each instance $\boldsymbol{x}$, determine $\boldsymbol{D}_\epsilon(\boldsymbol{x}) = \{\boldsymbol{x}' \in \boldsymbol{D} | d(\boldsymbol{x}, \boldsymbol{x}') \leq \epsilon\}$
- $\boldsymbol{D}_\epsilon(\boldsymbol{x})$ can be used to estimate $\boldsymbol{\theta}|\boldsymbol{x} := \boldsymbol{\theta}|\boldsymbol{D}_\epsilon(\boldsymbol{x})$

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

Learning Problem (Cont.)

Given $\boldsymbol{D}_\epsilon(\boldsymbol{x})$, we can

- Count $n = |\boldsymbol{D}_\epsilon(\boldsymbol{x})|$ and n_y, for any $y \in \mathscr{Y} \longleftarrow \sum_{y \in \mathscr{Y}} n_y = n$
- Estimate $\boldsymbol{\theta}|\boldsymbol{x}$ using MLE, DM, etc.

Learning Problem (Cont.)

Given $\boldsymbol{D}_\epsilon(\boldsymbol{x})$, we can

- Count $n = |\boldsymbol{D}_\epsilon(\boldsymbol{x})|$ and n_y, for any $y \in \mathscr{Y} \longleftarrow \sum_{y \in \mathscr{Y}} n_y = n$
- Estimate $\boldsymbol{\theta}|\boldsymbol{x}$ using MLE, DM, etc.

Optimal decision rules (recap)

- 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

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

Learning Problem (Cont.)

Given $\boldsymbol{D}_\epsilon(\boldsymbol{x})$, we can

- Count $n = |\boldsymbol{D}_\epsilon(\boldsymbol{x})|$ and n_y, for any $y \in \mathscr{Y} \longleftarrow \sum_{y \in \mathscr{Y}} n_y = n$
- Estimate $\boldsymbol{\theta}|\boldsymbol{x}$ using MLE, DM, etc.

What would we do if $\boldsymbol{D}$ contains

- a small number of instances
- and/or missing/partial data?

Learning Problem (Cont.)

Given $\boldsymbol{D}_\epsilon(\boldsymbol{x})$, we can

- Count $n = |\boldsymbol{D}_\epsilon(\boldsymbol{x})|$ and n_y, for any $y \in \mathscr{Y} \longleftarrow \sum_{y \in \mathscr{Y}} n_y = n$
- Estimate $\boldsymbol{\theta}|\boldsymbol{x}$ using MLE, DM, etc.

What would we do if $\boldsymbol{D}$ contains

- a small number of instances
- and/or missing/partial data?

$\boldsymbol{x}' \in \boldsymbol{D}_\epsilon(\boldsymbol{x})$	$Y' \subset \mathscr{Y} = \{\text{Apple}, \text{Banana}, \text{Tomato}\}$
$\boldsymbol{x}'_1$	Apple or Banana, but not Tomato
$\boldsymbol{x}'_2$	Banana or Tomato, but not Apple
$\boldsymbol{x}'_3$	Apple or Tomato, but not Banana
$\boldsymbol{x}'_4$	Tomato
$\boldsymbol{x}'_5$	Tomato
$\boldsymbol{x}'_6$	Banana
$\boldsymbol{x}'_7$	Banana

Imprecise Pazen Window Classifiers

Basic setup and assumption

- Given training data $\boldsymbol{D} \subset \mathscr{X} \times \mathscr{Y}$, a distance $d(\boldsymbol{x}, \boldsymbol{x}')$, and a threshold ϵ
- For each instance $\boldsymbol{x}$, determine $\boldsymbol{D}_\epsilon(\boldsymbol{x}) = \{\boldsymbol{x}' \in \boldsymbol{D} | d(\boldsymbol{x}, \boldsymbol{x}') \leq \epsilon\}$
- $\boldsymbol{D}_\epsilon(\boldsymbol{x})$ can be used to estimate $\boldsymbol{\Theta}|\boldsymbol{x} := \boldsymbol{\Theta}|\boldsymbol{D}_\epsilon(\boldsymbol{x})$

E-admissibility Rule [4, 5]:

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

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

- Computation: Solving linear programs, etc.

Learning Problem (Cont.)

Given $\boldsymbol{D}_\epsilon(\boldsymbol{x})$, we can

- Count $n = |\boldsymbol{D}_\epsilon(\boldsymbol{x})|$ and $\boldsymbol{n}_y$ for $y \in \mathscr{Y}$
- Determine $\mathscr{D} = \{\boldsymbol{D} | n_y \in \boldsymbol{n}_y, \sum_{y \in \mathscr{Y}} n_y = n\}$

Learning Problem (Cont.)

Given $\boldsymbol{D}_\epsilon(\boldsymbol{x})$, we can

- Count $n = |\boldsymbol{D}_\epsilon(\boldsymbol{x})|$ and $\boldsymbol{n}_y$ for $y \in \mathscr{Y}$
- Determine $\mathscr{D} = \{\boldsymbol{D} | n_y \in \boldsymbol{n}_y, \sum_{y \in \mathscr{Y}} n_y = n\}$

Using IDM to estimate **interval posterior mean** $\theta_y^* | \mathscr{D}$ of $\theta_y | \mathscr{D}$:

$$\underline{E}(\theta_y | \boldsymbol{x}) = \min_{\boldsymbol{D} \in \mathscr{D}} \underline{E}(\theta_y | \boldsymbol{x}) = \min_{\boldsymbol{D} \in \mathscr{D}} {}^{n_y}\!/\!_{(n+s)}, \tag{9}$$

$$\overline{E}(\theta_y | \boldsymbol{x}) = \max_{\boldsymbol{D} \in \mathscr{D}} \overline{E}(\theta_y | \boldsymbol{D}) = \max_{\boldsymbol{D} \in \mathscr{D}} {}^{(n_y+s)}\!/\!_{(n+s)}. \tag{10}$$

Determine Possible Precise Data Set

$\boldsymbol{x}' \in \boldsymbol{D}_\epsilon(\boldsymbol{x})$	$Y \subset \mathscr{Y} = \{\text{Apple}, \text{Banana}, \text{Tomato}\}$
$\boldsymbol{x}'_1$	Apple or Banana, but not Tomato
$\boldsymbol{x}'_2$	Banana or Tomato, but not Apple
$\boldsymbol{x}'_3$	Apple or Tomato, but not Banana
$\boldsymbol{x}'_4$	Tomato
$\boldsymbol{x}'_5$	Tomato
$\boldsymbol{x}'_6$	Banana
$\boldsymbol{x}'_7$	Banana

$$n = 7\,, \boldsymbol{n}_A = \{0, 1, 2\}\,, \boldsymbol{n}_B = \{2, 3, 4\}\,, \boldsymbol{n}_T = \{2, 3, 4\} \qquad (11)$$

Determine Possible Precise Data Set

$\boldsymbol{x}' \in \boldsymbol{D}_\epsilon(\boldsymbol{x})$	$Y \subset \mathscr{Y} = \{\text{Apple}, \text{Banana}, \text{Tomato}\}$
$\boldsymbol{x}'_1$	Apple or Banana, but not Tomato
$\boldsymbol{x}'_2$	Banana or Tomato, but not Apple
$\boldsymbol{x}'_3$	Apple or Tomato, but not Banana
$\boldsymbol{x}'_4$	Tomato
$\boldsymbol{x}'_5$	Tomato
$\boldsymbol{x}'_6$	Banana
$\boldsymbol{x}'_7$	Banana

$$n = 7\,, \boldsymbol{n}_A = \{0,1,2\}\,, \boldsymbol{n}_B = \{2,3,4\}\,, \boldsymbol{n}_T = \{2,3,4\} \tag{11}$$

	$\boldsymbol{D}_1$	$\boldsymbol{D}_2$	$\boldsymbol{D}_3$	$\boldsymbol{D}_4$	$\boldsymbol{D}_5$	$\boldsymbol{D}_6$	$\boldsymbol{D}_7$	$\boldsymbol{D}_8$
n_A	0	0	1	1	1	2	2	2
n_B	3	4	2	3	4	2	3	4
n_T	4	3	4	3	2	4	3	3

Compute Lower and Upper Expectations

	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8
n_A	0	0	1	1	1	2	2	2
n_B	3	4	2	3	4	2	3	4
n_T	4	3	4	3	2	4	3	3

Compute Lower and Upper Expectations

	$\boldsymbol{D}_1$	$\boldsymbol{D}_2$	$\boldsymbol{D}_3$	$\boldsymbol{D}_4$	$\boldsymbol{D}_5$	$\boldsymbol{D}_6$	$\boldsymbol{D}_7$	$\boldsymbol{D}_8$
n_A	0	0	1	1	1	2	2	2
n_B	3	4	2	3	4	2	3	4
n_T	4	3	4	3	2	4	3	3

Using IDM to estimate **interval posterior mean** $\theta_y^*|\mathscr{D}$ of $\theta_y|\mathscr{D}$:

$$\underline{E}(\theta_y|\boldsymbol{x}) = \min_{\boldsymbol{D}\in\mathscr{D}} \underline{E}(\theta_y|\boldsymbol{x}) = \min_{\boldsymbol{D}\in\mathscr{D}} {}^{n_y}/_{(n+s)}, \quad (12)$$

$$\overline{E}(\theta_y|\boldsymbol{x}) = \max_{\boldsymbol{D}\in\mathscr{D}} \overline{E}(\theta_y|\boldsymbol{D}) = \max_{\boldsymbol{D}\in\mathscr{D}} {}^{(n_y+s)}/_{(n+s)}. \quad (13)$$

	$\underline{E}(\theta_y\|\boldsymbol{x})$	$\overline{E}(\theta_y\|\boldsymbol{x})$
A	$^0/_{(7+s)}$	$^{(2+s)}/_{(7+s)}$
B	$^2/_{(7+s)}$	$^{(4+s)}/_{(7+s)}$
T	$^2/_{(7+s)}$	$^{(4+s)}/_{(7+s)}$

Compute Lower and Upper Expectations (cont.)

- For any $y \in \mathscr{Y}$, let

$$\underline{n}_y = \sum_{\boldsymbol{x}' \in \boldsymbol{D}} \mathbb{1}(y = Y'), \tag{14}$$

$$\overline{n}_y = \sum_{\boldsymbol{x}' \in \boldsymbol{D}} \mathbb{1}(y \in Y'). \tag{15}$$

- Compute **interval posterior mean** $\theta^*_y|\mathscr{D}$ of $\theta_y|\mathscr{D}$:

$$\underline{E}(\theta_y|\boldsymbol{x}) = \underline{n}_y/(n+s), \tag{16}$$

$$\overline{E}(\theta_y|\boldsymbol{x}) = (\overline{n}_y+s)/(n+s). \tag{17}$$

Compute Lower and Upper Bound Expectation (Again)

$\boldsymbol{x}' \in \boldsymbol{D}_\epsilon(\boldsymbol{x})$	$Y \subset \mathscr{Y} = \{\text{Apple}, \text{Banana}, \text{Tomato}\}$
$\boldsymbol{x}'_1$	Apple or Banana, but not Tomato
$\boldsymbol{x}'_2$	Banana or Tomato, but not Apple
$\boldsymbol{x}'_3$	Apple or Tomato, but not Banana
$\boldsymbol{x}'_4$	Tomato
$\boldsymbol{x}'_5$	Tomato
$\boldsymbol{x}'_6$	Banana
$\boldsymbol{x}'_7$	Banana

Compute Lower and Upper Bound Expectation (Again)

$\boldsymbol{x}' \in \boldsymbol{D}_\epsilon(\boldsymbol{x})$	$Y \subset \mathscr{Y} = \{\text{Apple}, \text{Banana}, \text{Tomato}\}$
$\boldsymbol{x}'_1$	Apple or Banana, but not Tomato
$\boldsymbol{x}'_2$	Banana or Tomato, but not Apple
$\boldsymbol{x}'_3$	Apple or Tomato, but not Banana
$\boldsymbol{x}'_4$	Tomato
$\boldsymbol{x}'_5$	Tomato
$\boldsymbol{x}'_6$	Banana
$\boldsymbol{x}'_7$	Banana

	$\underline{n}_y$	$\overline{n}_y$	$\underline{E}(\theta_y \mid \boldsymbol{x})$	$\overline{E}(\theta_y \mid \boldsymbol{x})$
A	0	2	$0/(7+s)$	$(2+s)/(7+s)$
B	2	4	$2/(7+s)$	$(4+s)/(7+s)$
T	2	4	$2/(7+s)$	$(4+s)/(7+s)$

Outline

Outline

- Inference from Multinomial Data
- Imprecise Dirichlet Model (IDM)
- Applications in Classification Tasks
- Evaluate Classifiers
  - The cases of Singleton Prediction
  - The cases of Set-Valued Predictions
- Summary and Outlook

(Few) Commonly Used Criteria

Predictive ability (on a test set):

- Let $\ell : \mathscr{Y} \times \mathscr{Y} \longmapsto \mathbb{R}_+$ be any loss function.
- Compute (average) loss on the test set

(Few) Commonly Used Criteria

Predictive ability (on a test set):

- Let $\ell : \mathscr{Y} \times \mathscr{Y} \longmapsto \mathbb{R}_+$ be any loss function.
- Compute (average) loss on the test set

(Few) Other criteria:

- Calibration errors (See Lecture 3)
- Model complexity (Storage memory)
- Training and/or Inference time
- Robustness: Under the presence of noise
- Trustworthiness: Explainability, interpretability, etc.

Outline

- Inference from Multinomial Data
- Imprecise Dirichlet Model (IDM)
- Applications in Classification Tasks
- Evaluate Classifiers
  - The cases of Singleton Prediction
  - The cases of Set-Valued Predictions
- Summary and Outlook

(Few) Commonly Used Criteria

Predictive ability (on a test set):

- We can use any loss function $\ell : 2^{\mathscr{Y}} \times \mathscr{Y} \longmapsto \mathbb{R}_+$.
- If we use utility metric $u = 1 - \ell$, replacing $\min$ by $\max$.
- Set-based utility functions [7]: $u(Y, y) = \mathbb{1}(y \in Y)g(|Y|)$
- Few commonly used utility function [2]:

$$g_\alpha(|Y|) = \frac{\alpha}{|Y|} + \frac{\alpha - 1}{|Y|^2},.$$

(Few) Commonly Used Criteria

Predictive ability (on a test set):

- We can use any loss function $\ell : 2^{\mathscr{Y}} \times \mathscr{Y} \longmapsto \mathbb{R}_+$.
- If we use utility metric $u = 1 - \ell$, replacing $\min$ by $\max$.
- Set-based utility functions [7]: $u(Y, y) = \mathbb{1}(y \in Y) g(|Y|)$
- Few commonly used utility function [2]:

$$g_\alpha(|Y|) = \frac{\alpha}{|Y|} + \frac{\alpha - 1}{|Y|^2},.$$

(Few) Other criteria:

- Calibration errors (See Lecture 3)
- Model complexity (Storage memory)
- Training and/or Inference time
- Robustness: Under the presence of noise
- Trustworthiness: Explainability, interpretability, etc.

Set-Based Utility Functions

Few commonly used **utility functions**:

$$g_\alpha(|Y|) = \frac{\alpha}{|Y|} - \frac{\alpha - 1}{|Y|^2}.$$

Set-Based Utility Functions

Few commonly used **utility functions**:

$$g_\alpha(|Y|) = \frac{\alpha}{|Y|} - \frac{\alpha - 1}{|Y|^2}.$$

Reward to cautiousness:

- u_{50}: $\alpha = 1 \longleftarrow$ no reward.
- u_{65}: $\alpha = 1.6$, moderate reward.
- u_{80}: $\alpha = 2.2$, big reward.
- higher α, higher reward

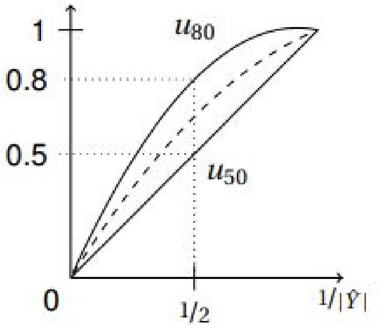
Set-Based Utility Functions

Few commonly used **utility functions**:

$$g_\alpha(|Y|) = \frac{\alpha}{|Y|} - \frac{\alpha - 1}{|Y|^2}.$$

Reward to cautiousness:

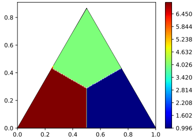
- u_{50}: $\alpha = 1 \longleftarrow$ no reward.
- u_{65}: $\alpha = 1.6$, moderate reward.
- u_{80}: $\alpha = 2.2$, big reward.
- higher α, higher reward



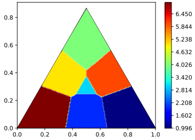
Outline

Optimal Decision Rules

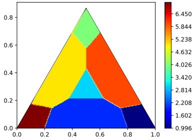
Frequentist approaches



0/1 loss



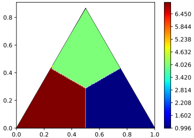
$u_{1.6}$



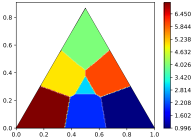
$u_{2.2}$

Optimal Decision Rules

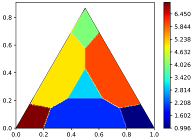
Frequentist approaches



0/1 loss

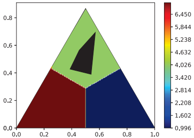


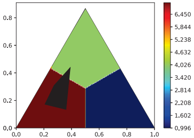
$u_{1.6}$

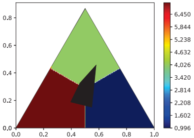


$u_{2.2}$

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 [4, 5]:

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

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

- 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

References I

[1] J.-M. Bernard.
An introduction to the imprecise dirichlet model for multinomial data.
International Journal of Approximate Reasoning, 39(2-3):123–150, 2005.

[2] T. Mortier, M. Wydmuch, K. Dembczyński, E. Hüllermeier, and W. Waegeman.
Efficient set-valued prediction in multi-class classification.
Data Mining and Knowledge Discovery, 35(4):1435–1469, 2021.

[3] V.-L. Nguyen, M. H. Shaker, and E. Hüllermeier.
How to measure uncertainty in uncertainty sampling for active learning.
Machine Learning, 111(1):89–122, 2022.

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

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

[6] L. V. Utkin and T. Augustin.
Decision making under incomplete data using the imprecise dirichlet model.
International Journal of Approximate Reasoning, 44(3):322–338, 2007.

[7] M. Zaffalon, G. Corani, and D. Mauá.
Evaluating credal classifiers by utility-discounted predictive accuracy.
International Journal of Approximate Reasoning, 53(8):1282–1301, 2012.