Likelihood-based Naive Credal Classifier

A movie (paper) by

Alessandro Antonucci*, Marco Cattaneo†, Giorgio Corani*

* Istituto "Dalle Molle" di Studi sull'Intelligenza Artificiale - Lugano (Switzerland)

† Ludwig-Maximilians-Universität - München (Germany)

ISIPTA'11 - Innsbruck, July 27, 2011

THE CAST

THE BAYESIAN

Giorgio Corani

BIO @IDSIA,
PhD in Engineering,
Bayesian/credal
classification

THE BAYESIAN

Giorgio Corani

BIO @IDSIA,
PhD in Engineering,
Bayesian/credal
classification

THE FREQUENTIST

Marco Cattaneo

BIO @LMU,
PhD Statistics,
Likelihood-based
learning

THE BAYESIAN

Giorgio Corani

BIO @IDSIA,
PhD in Engineering,
Bayesian/credal classification

THE AGNOSTIC

Alessandro Antonucci

BIO @IDSIA,
PhD Comp Science,
Bayesian/credal networks

THE FREQUENTIST

Marco Cattaneo

BIO @LMU,
PhD Statistics,
Likelihood-based learning

EPISODE I

“Crises of Faith”

Learning probabilistic
models from data

Learning from data $\mathcal{D}$ (about random var C, $c \in \mathcal{C}$)

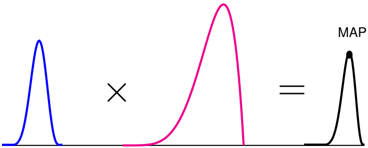
multinomial likelihood $P(\mathcal{D}) = \prod_{c \in \mathcal{C}} \theta_c^{n(c)}$ (counts n)

Learning from data $\mathcal{D}$ (about random var C, $c \in \mathcal{C}$)

multinomial likelihood $P(\mathcal{D}) = \prod_{c \in \mathcal{C}} \theta_c^{n(c)}$ (counts n)

BAYESIAN

prior $\times$ likelihood = posterior

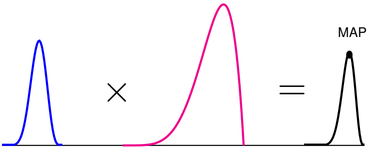


Learning from data $\mathcal{D}$ (about random var C, $c \in \mathcal{C}$)

multinomial likelihood $P(\mathcal{D}) = \prod_{c \in \mathcal{C}} \theta_c^{n(c)}$ (counts n)

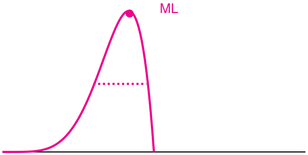
BAYESIAN

prior $\times$ likelihood = posterior



FREQUENTIST

likelihood (only)

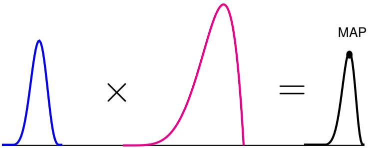


Learning from data $\mathcal{D}$ (about random var C, $c \in \mathcal{C}$)

multinomial likelihood $P(\mathcal{D}) = \prod_{c \in \mathcal{C}} \theta_c^{n(c)}$ (counts n)

BAYESIAN

prior $\times$ likelihood = posterior



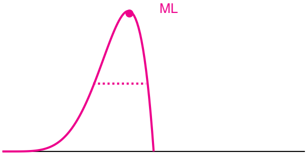
priors $\times$ likelihood = posteriors

IDM prior ignorance with $Dir(s\mathbf{t})$

$\mathbf{t} \in \mathcal{T} := \{\mathbf{t} \,|\, \sum_{c \in \mathcal{C}} t(c) = 1, t(c) > 0\}$

FREQUENTIST

likelihood (only)

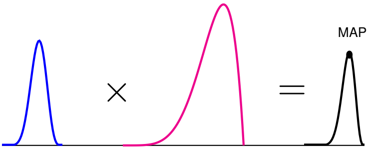


Learning from data $\mathcal{D}$ (about random var C, $c \in \mathcal{C}$)

multinomial likelihood $P(\mathcal{D}) = \prod_{c\in\mathcal{C}} \theta_c^{n(c)}$ (counts n)

BAYESIAN

prior $\times$ likelihood = posterior



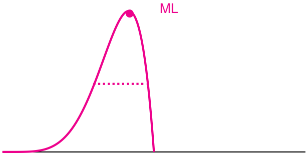
priors $\times$ likelihood = posterior**s**

IDM prior ignorance with $Dir(s\mathbf{t})$

$\mathbf{t} \in \mathcal{T} := \{\mathbf{t} \,|\, \sum_{c\in\mathcal{C}} t(c) = 1, t(c) > 0\}$

FREQUENTIST

likelihood (only)



model**s** with likelihood $\geq$ threshold

LIK refine starting credal set $\mathbf{P}$

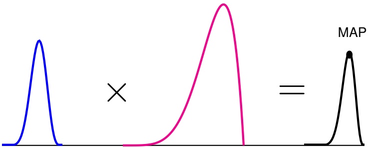
$\mathbf{P}_\alpha := \{P \in \mathbf{P} \,|\, P(\mathcal{D}) \geq \alpha P_{ML}(\mathcal{D})\}$

Learning from data $\mathcal{D}$ (about random var C, $c \in \mathcal{C}$)

multinomial likelihood $P(\mathcal{D}) = \prod_{c \in \mathcal{C}} \theta_c^{n(c)}$ (counts n)

BAYESIAN

prior $\times$ likelihood = posterior



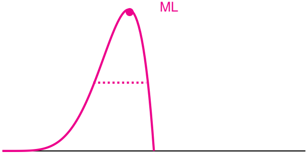
priors $\times$ likelihood = posterior**s**

IDM prior ignorance with $Dir(s\mathbf{t})$

$\mathbf{t} \in \mathcal{T} := \{\mathbf{t} \,|\, \sum_{c \in \mathcal{C}} t(c) = 1, t(c) > 0\}$

FREQUENTIST

likelihood (only)



model**s** with likelihood $\geq$ threshold

LIK refine starting credal set $\mathbf{P}$

$\mathbf{P}_\alpha := \{P \in \mathbf{P} \,|\, P(\mathcal{D}) \geq \alpha P_{ML}(\mathcal{D})\}$

threshold $\alpha \in [0, 1]$ ($\mathbf{P}_\alpha \subseteq \mathbf{P}$)

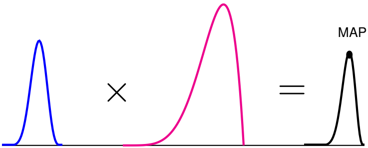
($\mathbf{P}_{\alpha=0} \equiv \mathbf{P}$, $\mathbf{P}_{\alpha=1} = P_{ML}$)

Learning from data $\mathcal{D}$ (about random var C, $c \in \mathcal{C}$)

multinomial likelihood $P(\mathcal{D}) = \prod_{c \in \mathcal{C}} \theta_c^{n(c)}$ (counts n)

BAYESIAN

prior $\times$ likelihood = posterior



priors $\times$ likelihood = posteriors

IDM prior ignorance with $Dir(s\mathbf{t})$

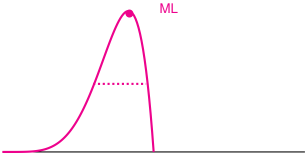
$\mathbf{t} \in \mathcal{T} := \{\mathbf{t} \,|\, \sum_{c \in \mathcal{C}} t(c) = 1, t(c) > 0\}$

s equivalent sample size

($s = 0$ precise, $s \to \infty$ vacuous)

FREQUENTIST

likelihood (only)



models with likelihood $\geq$ threshold

LIK refine starting credal set $\mathbf{P}$

$\mathbf{P}_\alpha := \{P \in \mathbf{P} \,|\, P(\mathcal{D}) \geq \alpha P_{ML}(\mathcal{D})\}$

threshold $\alpha \in [0, 1]$ ($\mathbf{P}_\alpha \subseteq \mathbf{P}$)

($\mathbf{P}_{\alpha=0} \equiv \mathbf{P}$, $\mathbf{P}_{\alpha=1} = P_{ML}$)

EPISODE II

“Becoming adults”

Credal classification

Probabilistic Classifiers

Class C, Features $\mathbf{F} := (F_1, \ldots, F_m)$, $f_i \in \mathcal{F}_i$
complete data $\mathcal{D} := \{(c^{(j)}, \mathbf{f}^{(j)})\}_{j=1}^{d}$

Which class label assign to instance $\mathbf{F} = \tilde{\mathbf{f}}$?

Precise classifiers learn joint $P(C, \mathbf{F})$

Assign to $\tilde{\mathbf{f}}$ most probable class label

$$\arg\max_{c' \in \mathcal{C}} P(c', \tilde{\mathbf{f}})$$

this is a classifier: $(\mathcal{F}_1 \times \ldots \times \mathcal{F}_m) \to \mathcal{C}$

Credal classifiers learn joint credal set $\mathbf{P}(C, \mathbf{F})$

Set of optimal class labels (e.g., maximality)

$$\{c' \in \mathcal{C} | \nexists c'' \in \mathcal{C} : P(c'', \tilde{\mathbf{f}}) > P(c', \tilde{\mathbf{f}}), \forall P \in \mathbf{P}\}$$

A credal classifier: $(\mathcal{F}_1 \times \ldots \times \mathcal{F}_m) \to 2^{\mathcal{C}}$

(may) return multiple classes

naive assumption

"given the class, features are independent"

Probabilistic Classifiers

Class C, Features $\mathbf{F} := (F_1, \ldots, F_m)$, $f_i \in \mathcal{F}_i$
complete data $\mathcal{D} := \{(c^{(j)}, \mathbf{f}^{(j)})\}_{j=1}^{d}$

Which class label assign to instance $\mathbf{F} = \tilde{\mathbf{f}}$?

Precise classifiers learn joint $P(C, \mathbf{F})$

Assign to $\tilde{\mathbf{f}}$ most probable class label

$$\arg\max_{c' \in \mathcal{C}} P(c', \tilde{\mathbf{f}})$$

this is a classifier: $(\mathcal{F}_1 \times \ldots \times \mathcal{F}_m) \to \mathcal{C}$

Credal classifiers learn joint credal set $\mathbf{P}(C, \mathbf{F})$

Set of optimal class labels (e.g., maximality)

$$\{c' \in \mathcal{C} | \nexists c'' \in \mathcal{C} : P(c'', \tilde{\mathbf{f}}) > P(c', \tilde{\mathbf{f}}), \forall P \in \mathbf{P}\}$$

A credal classifier: $(\mathcal{F}_1 \times \ldots \times \mathcal{F}_m) \to 2^{\mathcal{C}}$

(may) return multiple classes

naive assumption

"given the class, features are independent"

Probabilistic Classifiers

Class C, Features $\mathbf{F} := (F_1, \ldots, F_m)$, $f_i \in \mathcal{F}_i$
complete data $\mathcal{D} := \{(c^{(j)}, \mathbf{f}^{(j)})\}_{j=1}^{d}$

Which class label assign to instance $\mathbf{F} = \tilde{\mathbf{f}}$?

Precise classifiers learn joint $P(C, \mathbf{F})$

Assign to $\tilde{\mathbf{f}}$ most probable class label

$$\arg\max_{c' \in \mathcal{C}} P(c', \tilde{\mathbf{f}})$$

this is a classifier: $(\mathcal{F}_1 \times \ldots \times \mathcal{F}_m) \to \mathcal{C}$

Credal classifiers learn joint credal set $\mathbf{P}(C, \mathbf{F})$

Set of optimal class labels (e.g., maximality)

$$\{c' \in \mathcal{C} | \nexists c'' \in \mathcal{C} : P(c'', \tilde{\mathbf{f}}) > P(c', \tilde{\mathbf{f}}), \forall P \in \mathbf{P}\}$$

A credal classifier: $(\mathcal{F}_1 \times \ldots \times \mathcal{F}_m) \to 2^{\mathcal{C}}$

(may) return multiple classes

naive assumption

"given the class, features are independent"

Probabilistic Classifiers

Class C, Features $\mathbf{F} := (F_1, \ldots, F_m)$, $f_i \in \mathcal{F}_i$
complete data $\mathcal{D} := \{(c^{(j)}, \mathbf{f}^{(j)})\}_{j=1}^{d}$

Which class label assign to instance $\mathbf{F} = \tilde{\mathbf{f}}$?

Precise classifiers learn joint $P(C, \mathbf{F})$

Assign to $\tilde{\mathbf{f}}$ most probable class label

$$\arg\max_{c' \in \mathcal{C}} P(c', \tilde{\mathbf{f}})$$

this is a classifier: $(\mathcal{F}_1 \times \ldots \times \mathcal{F}_m) \to \mathcal{C}$

Credal classifiers learn joint credal set $\mathbf{P}(C, \mathbf{F})$

Set of optimal class labels (e.g., maximality)

$$\{c' \in \mathcal{C} | \nexists c'' \in \mathcal{C} : P(c'', \tilde{\mathbf{f}}) > P(c', \tilde{\mathbf{f}}), \forall P \in \mathbf{P}\}$$

A credal classifier: $(\mathcal{F}_1 \times \ldots \times \mathcal{F}_m) \to 2^{\mathcal{C}}$

(may) return multiple classes

naive assumption

"given the class, features are independent"

Probabilistic Classifiers

Class C, Features $\mathbf{F} := (F_1, \ldots, F_m)$, $f_i \in \mathcal{F}_i$
complete data $\mathcal{D} := \{(c^{(j)}, \mathbf{f}^{(j)})\}_{j=1}^{d}$

Which class label assign to instance $\mathbf{F} = \tilde{\mathbf{f}}$?

Precise classifiers learn joint $P(C, \mathbf{F})$

Assign to $\tilde{\mathbf{f}}$ most probable class label

$$\arg\max_{c' \in \mathcal{C}} P(c', \tilde{\mathbf{f}})$$

this is a classifier: $(\mathcal{F}_1 \times \ldots \times \mathcal{F}_m) \to \mathcal{C}$

Credal classifiers learn joint credal set $\mathbf{P}(C, \mathbf{F})$

Set of optimal class labels (e.g., maximality)

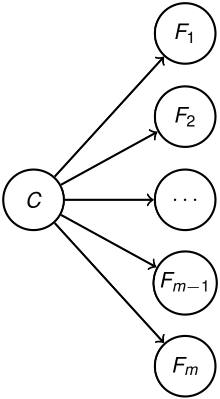
$$\{c' \in \mathcal{C} | \nexists c'' \in \mathcal{C} : P(c'', \tilde{\mathbf{f}}) > P(c', \tilde{\mathbf{f}}), \forall P \in \mathbf{P}\}$$

A credal classifier: $(\mathcal{F}_1 \times \ldots \times \mathcal{F}_m) \to 2^{\mathcal{C}}$

(may) return multiple classes

naive assumption

"given the class, features are independent"



Naive Credal Classifiers (NCC)

BAYESIAN

- IDM-based NCC (Zaffalon, 2001)
- Efficient classification algorithm based optimization under the linear constraints

FREQUENTIST

- LIK-based NCC (this paper)
- Efficient classification algorithm based on analytical derivation of the likelihood upper envelope (α-cuts identified numerically)

Feature problem zero joint counts $n(C = c', F_i = f_i) = 0$ make the classifier widely imprecisei

NCC$_\epsilon$ (Corani & Benavoli, 2010)

- Shrink the IDM set of priors by linear-vacuous contamination
- NCC$_{\epsilon=0}$ =NBC, NCC$_{\epsilon=1}$ vacuous

NCC$_\alpha$ (this paper)

- "semi-supervised" learning $\mathcal{D} := \mathcal{D} \cup (C = *, \mathbf{F} = \tilde{\mathbf{f}})$
- Assume C missing-at-random for the incomplete instance $\tilde{\mathbf{f}}$

Naive Credal Classifiers (NCC)

BAYESIAN

- IDM-based NCC (Zaffalon, 2001)
- Efficient classification algorithm based optimization under the linear constraints

FREQUENTIST

- LIK-based NCC (this paper)
- Efficient classification algorithm based on analytical derivation of the likelihood upper envelope (α-cuts identified numerically)

Feature problem zero joint counts $n(C = c', F_i = f_i) = 0$ make the classifier widely imprecisei

NCC$_\epsilon$ (Corani & Benavoli, 2010)

- Shrink the IDM set of priors by linear-vacuous contamination
- NCC$_{\epsilon=0}$ =NBC, NCC$_{\epsilon=1}$ vacuous

NCC$_\alpha$ (this paper)

- "semi-supervised" learning $\mathcal{D} := \mathcal{D} \cup (C = *, \mathbf{F} = \tilde{\mathbf{f}})$
- Assume C missing-at-random for the incomplete instance $\tilde{\mathbf{f}}$

Naive Credal Classifiers (NCC)

BAYESIAN

- IDM-based NCC (Zaffalon, 2001)
- Efficient classification algorithm based optimization under the linear constraints

FREQUENTIST

- LIK-based NCC (this paper)
- Efficient classification algorithm based on analytical derivation of the likelihood upper envelope (α-cuts identified numerically)

Feature problem zero joint counts $n(C = c', F_i = f_i) = 0$ make the classifier widely imprecisei

NCC$_\epsilon$ (Corani & Benavoli, 2010)

- Shrink the IDM set of priors by linear-vacuous contamination
- NCC$_{\epsilon=0}$ =NBC, NCC$_{\epsilon=1}$ vacuous

NCC$_\alpha$ (this paper)

- "semi-supervised" learning $\mathcal{D} := \mathcal{D} \cup (C = *, \mathbf{F} = \tilde{\mathbf{f}})$
- Assume C missing-at-random for the incomplete instance $\tilde{\mathbf{f}}$

Naive Credal Classifiers (NCC)

BAYESIAN

- IDM-based NCC (Zaffalon, 2001)
- Efficient classification algorithm based optimization under the linear constraints

FREQUENTIST

- LIK-based NCC (this paper)
- Efficient classification algorithm based on analytical derivation of the likelihood upper envelope (α-cuts identified numerically)

Feature problem zero joint counts $n(C = c', F_i = f_i) = 0$ make the classifier widely imprecisei

NCC$_\epsilon$ (Corani & Benavoli, 2010)

- Shrink the IDM set of priors by linear-vacuous contamination
- NCC$_{\epsilon=0}$ =NBC, NCC$_{\epsilon=1}$ vacuous

NCC$_\alpha$ (this paper)

- "semi-supervised" learning $\mathcal{D} := \mathcal{D} \cup (C = *, \mathbf{F} = \tilde{\mathbf{f}})$
- Assume C missing-at-random for the incomplete instance $\tilde{\mathbf{f}}$

Naive Credal Classifiers (NCC)

BAYESIAN

- IDM-based NCC (Zaffalon, 2001)
- Efficient classification algorithm based optimization under the linear constraints

FREQUENTIST

- LIK-based NCC (this paper)
- Efficient classification algorithm based on analytical derivation of the likelihood upper envelope (α-cuts identified numerically)

Feature problem zero joint counts $n(C = c', F_i = f_i) = 0$ make the classifier widely imprecisei

NCC_ϵ (Corani & Benavoli, 2010)

- Shrink the IDM set of priors by linear-vacuous contamination
- $NCC_{\epsilon=0}$ =NBC, $NCC_{\epsilon=1}$ vacuous

NCC_α (this paper)

- "semi-supervised" learning $\mathcal{D} := \mathcal{D} \cup (C = *, \mathbf{F} = \tilde{\mathbf{f}})$
- Assume C missing-at-random for the incomplete instance $\tilde{\mathbf{f}}$

Naive Credal Classifiers (NCC)

BAYESIAN

- IDM-based NCC (Zaffalon, 2001)
- Efficient classification algorithm based optimization under the linear constraints

FREQUENTIST

- LIK-based NCC (this paper)
- Efficient classification algorithm based on analytical derivation of the likelihood upper envelope (α-cuts identified numerically)

Feature problem zero joint counts $n(C = c', F_i = f_i) = 0$ make the classifier widely imprecisei

NCC$_\epsilon$ (Corani & Benavoli, 2010)

- Shrink the IDM set of priors by linear-vacuous contamination
- NCC$_{\epsilon=0}$ =NBC, NCC$_{\epsilon=1}$ vacuous

NCC$_\alpha$ (this paper)

- "semi-supervised" learning $\mathcal{D} := \mathcal{D} \cup (C = *, \mathbf{F} = \tilde{\mathbf{f}})$
- Assume C missing-at-random for the incomplete instance $\tilde{\mathbf{f}}$

Naive Credal Classifiers (NCC)

BAYESIAN

- IDM-based NCC (Zaffalon, 2001)
- Efficient classification algorithm based optimization under the linear constraints

FREQUENTIST

- LIK-based NCC (this paper)
- Efficient classification algorithm based on analytical derivation of the likelihood upper envelope (α-cuts identified numerically)

Feature problem zero joint counts $n(C = c', F_i = f_i) = 0$ make the classifier widely imprecisei

NCC$_\epsilon$ (Corani & Benavoli, 2010)

- Shrink the IDM set of priors by linear-vacuous contamination
- NCC$_{\epsilon=0}$ =NBC, NCC$_{\epsilon=1}$ vacuous

NCC$_\alpha$ (this paper)

- "semi-supervised" learning $\mathcal{D} := \mathcal{D} \cup (C = *, \mathbf{F} = \tilde{\mathbf{f}})$
- Assume C missing-at-random for the incomplete instance $\tilde{\mathbf{f}}$

Naive Credal Classifiers (NCC)

BAYESIAN

- IDM-based NCC (Zaffalon, 2001)
- Efficient classification algorithm based optimization under the linear constraints

FREQUENTIST

- LIK-based NCC (this paper)
- Efficient classification algorithm based on analytical derivation of the likelihood upper envelope (α-cuts identified numerically)

Feature problem zero joint counts $n(C = c', F_i = f_i) = 0$ make the classifier widely imprecisei

NCC$_\epsilon$ (Corani & Benavoli, 2010)

- Shrink the IDM set of priors by linear-vacuous contamination
- NCC$_{\epsilon=0}$ =NBC, NCC$_{\epsilon=1}$ vacuous

NCC$_\alpha$ (this paper)

- "semi-supervised" learning $\mathcal{D} := \mathcal{D} \cup (C = *, \mathbf{F} = \tilde{\mathbf{f}})$
- Assume C missing-at-random for the incomplete instance $\tilde{\mathbf{f}}$

EPISODE III

“The final duel”

Comparing the two classifiers
(assuming you know how to deal with it)

Extensive tests on UCI datasets

with JNCC2 (`idsia.ch/~giorgio/jncc2.html`)

(α, ϵ) tuned to reach same average output size

Very similar performances, according to the available performance descriptors

Extensive tests on UCI datasets

with JNCC2 (`idsia.ch/~giorgio/jncc2.html`)

(α, ϵ) tuned to reach same average output size

Accuracy when precise

Overall accuracy (discounted)

Very similar performances, according to the available performance descriptors

Extensive tests on UCI datasets

with JNCC2 (`idsia.ch/~giorgio/jncc2.html`)

(α, ϵ) tuned to reach same average output size

Accuracy when precise

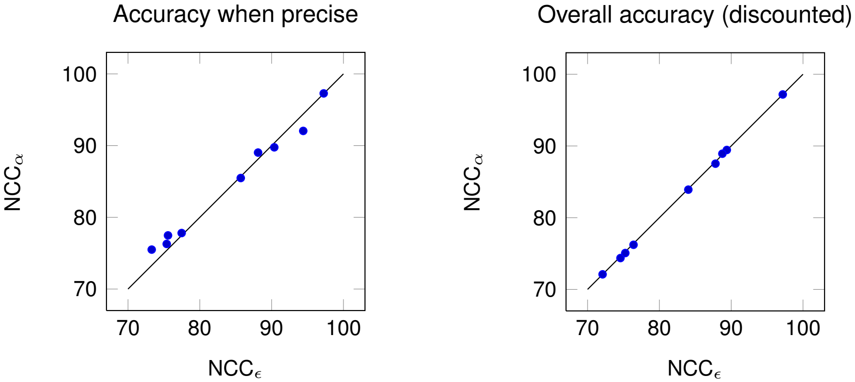
Overall accuracy (discounted)

Very similar performances, according to
the available performance descriptors

Extensive tests on UCI datasets

with JNCC2 (`idsia.ch/~giorgio/jncc2.html`)

(α, ϵ) tuned to reach same average output size



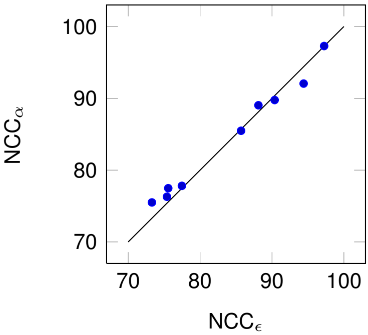
Very similar performances, according to
the available performance descriptors

Extensive tests on UCI datasets

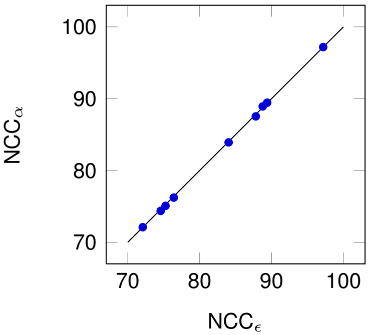
with JNCC2 (`idsia.ch/~giorgio/jncc2.html`)

(α, ϵ) tuned to reach same average output size

Accuracy when precise



Overall accuracy (discounted)



Very similar performances, according to
the available performance descriptors

THE END
(ALMOST)

What about the agnostic?

- The agnostic is still agnostic
- None of the two classifiers clearly outperforms the other (according to the actual metrics)
- “Bayesian” approach has a clear behavioural interpretation
- “Frequentist” approach promising for analytical results even with more complex independence structures
- If no classifier outperforms the other, use them sequentially! Future work: indecision in NCC_ϵ could be resolved by NCC_α minimum α resolving indecision as a confidence level

What about the agnostic?

- The agnostic is still agnostic
- None of the two classifiers clearly outperforms the other (according to the actual metrics)
- “Bayesian” approach has a clear behavioural interpretation
- “Frequentist” approach promising for analytical results even with more complex independence structures
- If no classifier outperforms the other, use them sequentially !
  Future work: indecision in NCC_ϵ could be resolved by NCC_α minimum α resolving indecision as a confidence level

What about the agnostic?

- The agnostic is still agnostic
- None of the two classifiers clearly outperforms the other (according to the actual metrics)
- "Bayesian" approach has a clear behavioural interpretation
- "Frequentist" approach promising for analytical results even with more complex independence structures
- If no classifier outperforms the other, use them sequentially!
  Future work: indecision in NCC_ϵ could be resolved by NCC_α
  minimum α resolving indecision as a confidence level

What about the agnostic?

- The agnostic is still agnostic
- None of the two classifiers clearly outperforms the other (according to the actual metrics)
- "Bayesian" approach has a clear behavioural interpretation
- "Frequentist" approach promising for analytical results even with more complex independence structures
- If no classifier outperforms the other, use them sequentially !
  Future work: indecision in NCC_ϵ could be resolved by NCC_α
  minimum α resolving indecision as a confidence level

What about the agnostic?

- The agnostic is still agnostic
- None of the two classifiers clearly outperforms the other (according to the actual metrics)
- "Bayesian" approach has a clear behavioural interpretation
- "Frequentist" approach promising for analytical results even with more complex independence structures
- If no classifier outperforms the other, use them sequentially! Future work: indecision in NCC_ϵ could be resolved by NCC_α minimum α resolving indecision as a confidence level

What about the agnostic?

- The agnostic is still agnostic
- None of the two classifiers clearly outperforms the other (according to the actual metrics)
- “Bayesian” approach has a clear behavioural interpretation
- “Frequentist” approach promising for analytical results even with more complex independence structures
- If no classifier outperforms the other, use them sequentially!
  Future work: indecision in NCC_ϵ could be resolved by NCC_α minimum α resolving indecision as a confidence level

What about the agnostic?

- The agnostic is still agnostic
- None of the two classifiers clearly outperforms the other (according to the actual metrics)
- “Bayesian” approach has a clear behavioural interpretation
- “Frequentist” approach promising for analytical results even with more complex independence structures
- If no classifier outperforms the other, use them sequentially ! Future work: indecision in NCC$_\epsilon$ could be resolved by NCC$_\alpha$ minimum α resolving indecision as a confidence level

THE END

THE END

All characters appearing in this work are fictitious. Any resemblance to real persons, living or dead, is purely coincidental.