CaDET: Interpretable Parametric Conditional Density Estimation with Decision Trees and Forests

Cyrus Cousins[1] , Matteo Riondato

Brown University, Amherst College

[1]cyrus_cousins@brown.edu

http://cs.brown.edu/~ccousins/cadet/

ECMLPKDD Würzburg 16.–20.09.2019

CaDET at a Glance

"Interpretable Parametric Conditional-Density-Estimation"

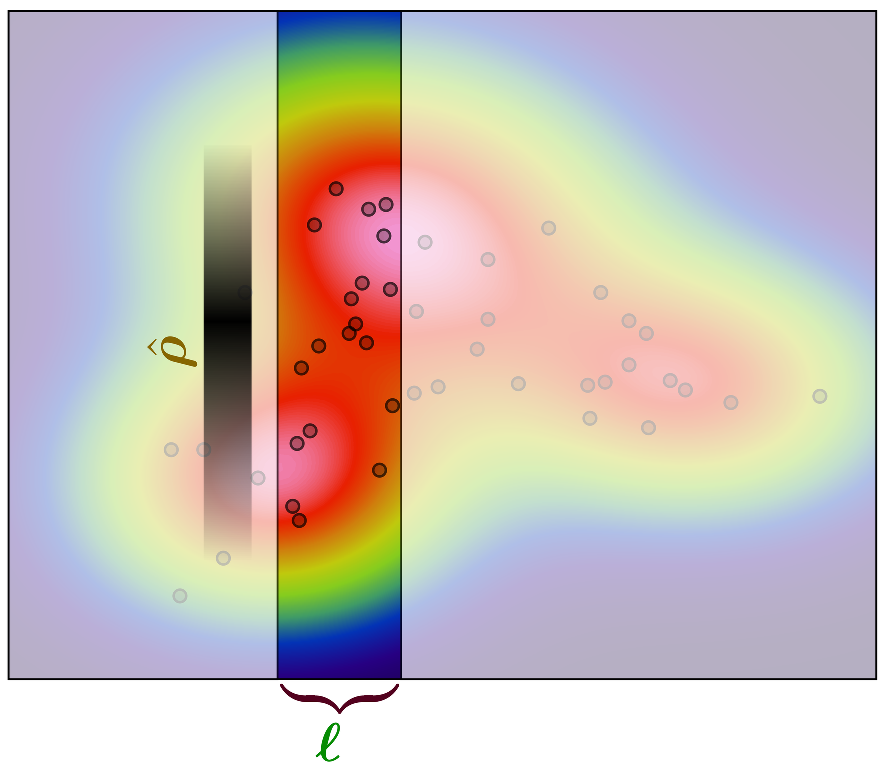
(1) CaDET trees and predictions are *interpretable* (simple to understand)
(2) CaDET predicts *parametric densities*, e.g. $\textsc{Gaussian}(1, 1)$, $\textsc{Beta}(3, 2)$
(3) *Conditional Density Estimation*: estimate *distributions*

CaDET improves upon nonparametric tree methods in:

- *Training*, *query*, and *storage* costs
- *Interpretability* of estimated conditional densities
- *Sample complexity*

The Setting: Conditional Density Estimation

- Domain $\mathcal{X}$, codomain $\mathcal{Y}$
  - PDF ρ over $\mathcal{X} \times \mathcal{Y}$
  - Sample $\{ \circ\circ\circ\circ \} \sim \rho$
- Condition on query point $q \in \mathcal{X}$
- Estimate $\hat{\rho}(\cdot \mid q) \approx \rho(\cdot \mid q)$
- Decision tree CDE:
  - Leaf ℓ contains q
  - Fit PDF $\hat{\rho}(\cdot)$ to ℓ

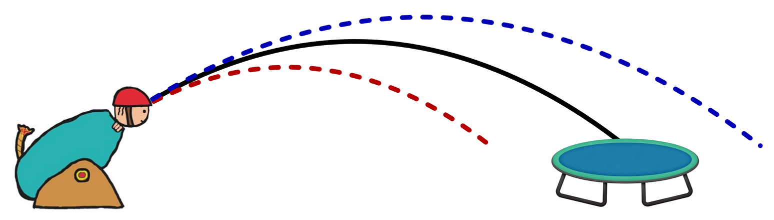


Overview of Supervised Learning

"Given query $q \in \mathcal{X}$, predict or summarize $\rho(\cdot \mid q)$"

		Prediction Type: Summary	Prediction Type: Distribution
$\mathcal{Y}$	Discrete	Hard Classification $\operatorname{argmax}_y \mathbb{P}(y \mid q)$	Soft Classification $\mathbb{P}(\cdot \mid q)$
	Continuous	Regression $\mathbb{E}_{(x,y)\sim\rho}[y \mid x = q]$	CDE $\rho(\cdot \mid q)$

- Regression is a *lossy process*
  - Only estimate *average outcome*
  - Can't reason about *other outcomes*
  - CDE tells the full story

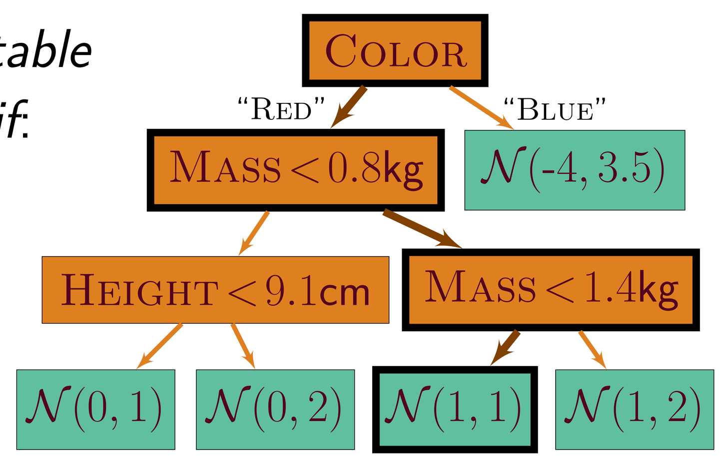


Interpretability in CDE Trees

- CDE traditionally considered *uninterpretable*
- *Parametric* CDE trees are interpretable *if*:

(1) Model structure can be visualized
(2) Predictions are intuitively understood
(3) **Decision process** is easily audited

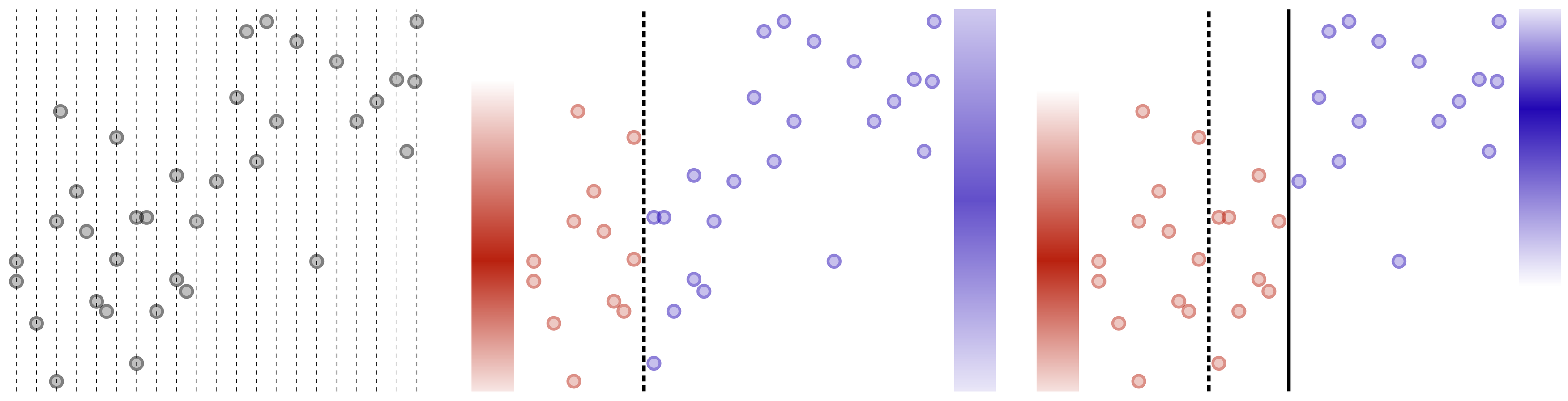
- Nonparametric CDE is uninterpretable



Constructing CaDET Trees

Hyperparameter: CaDET predicts distributions from *parametric family* $\mathcal{F}$

- Given query point q, CaDET trees
  (1) Find the leaf that contains q, with training labels y
  (2) Return $\mathrm{MLE}_{\mathcal{F}}(y)$
- CaDET trees minimize *cross entropy impurity* $\mathrm{I}_{\mathcal{F}}(\cdot)$
  - Evaluate CDE with *cross entropy loss* $\ell_{\mathrm{CE}}(y \mid \hat{\rho}) \doteq -\ln \hat{\rho}(y)$
  - *Cross entropy impurity* $\mathrm{I}_{\mathcal{F}}(y) \doteq \frac{1}{m}\sum_{i=1}^{m} \ell_{\mathrm{CE}}(y_i \mid \mathrm{MLE}(y; \mathcal{F}))$



(1) Given training points & labels
(2) Evaluate possible splits

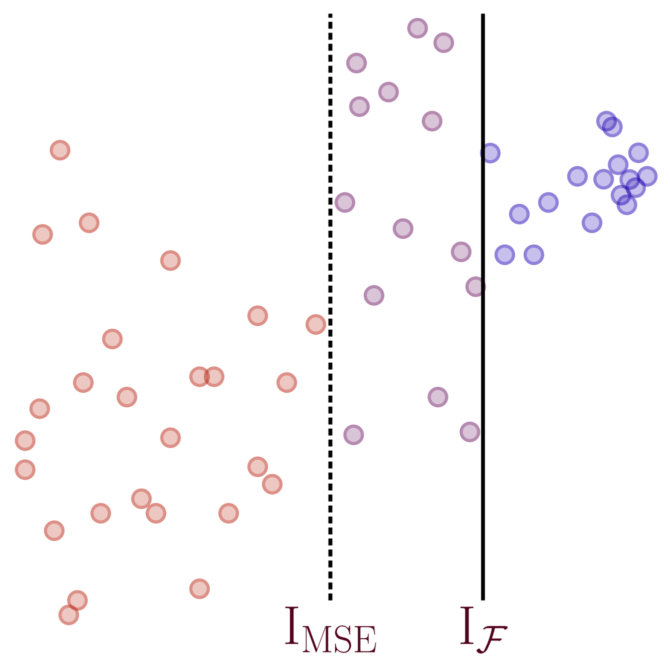
(1) Consider split ⋮ into y_L, y_R
(2) Fit $\mathrm{MLE}_{\mathcal{F}}(y_L)$, $\mathrm{MLE}_{\mathcal{F}}(y_R)$
(3) $\mathrm{I}_{\mathcal{F}}(y_L)$: low
$\mathrm{I}_{\mathcal{F}}(y_R)$: high

(1) Now consider the split |
(2) Fit $\mathrm{MLE}_{\mathcal{F}}(y_L)$, $\mathrm{MLE}_{\mathcal{F}}(y_R)$
(3) $\mathrm{I}_{\mathcal{F}}(y_L)$: same
$\mathrm{I}_{\mathcal{F}}(y_R)$: lower

Choice of Impurity Criterion

- ECE impurity criterion $\mathrm{I}_{\mathcal{F}}(\cdot)$ tailored to $\mathcal{F}$
- Handle heavy tails, multimodality, scale variation

$\mathcal{F}$	$\mathrm{I}_{\mathcal{F}}(y) \equiv$
$\textsc{Gaussian}(\cdot, 1)$	$\hat{\mathbb{V}}_{y} = \mathrm{I}_{\mathrm{MSE}}(y)$
$\textsc{Categorical}(\cdot)$	$-\hat{\mathbb{E}}_{y \in y}[\ln \hat{\mathbb{P}}_{y}(y)] = \mathrm{I}_{\mathrm{H}}(y)$
$\textsc{Uniform}(\cdot, \cdot)$	$\ln(\max(y) - \min(y))$
$\textsc{Pareto}(\cdot, 1)$	$\left(\hat{\mathbb{E}}_{y \in y}[\ln(y)]\right)^{-1}$



Organizing Computation with Sufficient Statistics

- Consider a leaf with training labels y
  - y used to select splits in training and answer queries
  - CART *summarize* y with *class frequencies* or *means*
  - "Incremental updates" make split search fast
- *Sufficient statistics* $\mathrm{w}^{(m)}(y)$ used to compute $\mathrm{MLE}_{\mathcal{F}}(y)$, $\mathrm{I}_{\mathcal{F}}(y)$

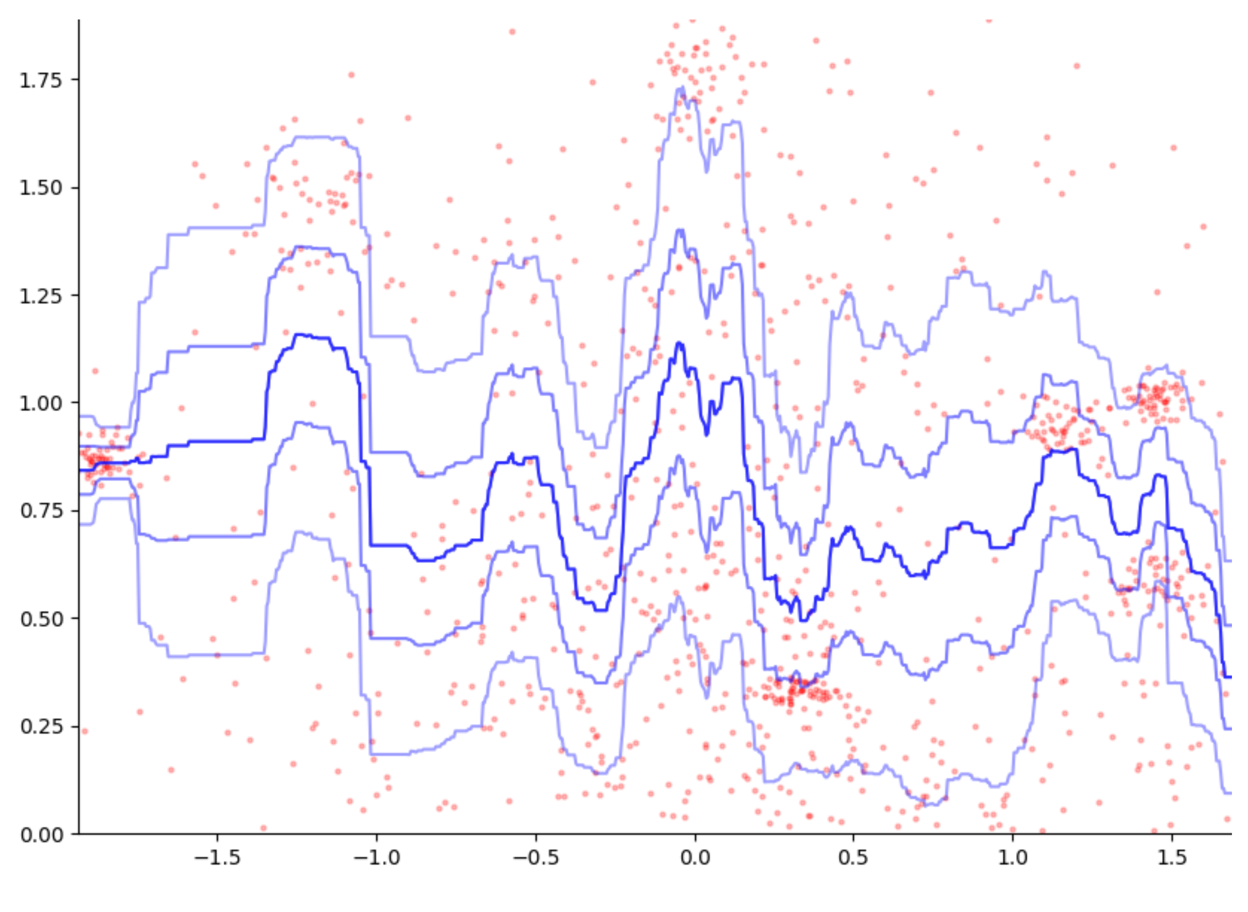
Family $\mathcal{F}$	Suff. Stat. $\mathrm{w}^{(m)}(y)$	Log Density $\ln \rho(y)$
$\textsc{Gaussian}(\mu, \sigma^2)$	$\sum_{i=1}^{m} y_i,\ \sum_{i=1}^{m} y_i^2$	$-\frac{1}{2}\ln(2\pi\sigma^2) - \frac{y^2 - 2y\mu + \mu^2}{2\sigma^2}$
$\textsc{Gamma}(\alpha, \beta)$	$\sum_{i=1}^{m} y_i,\ \sum_{i=1}^{m} \ln(y_i)$	$\alpha\ln(\beta) - \ln\Gamma(\alpha) - \beta y + (\alpha - 1)\ln(y)$

- Always exist *additive* $\mathrm{w}(\cdot)$ for $\mathcal{F}$ in the *exponential class* s.t.

$$\mathrm{w}^{(m_L + m_R)}(y_L \circ y_R) \doteq \mathrm{w}^{(m_L)}(y_L) + \mathrm{w}^{(m_R)}(y_2)$$

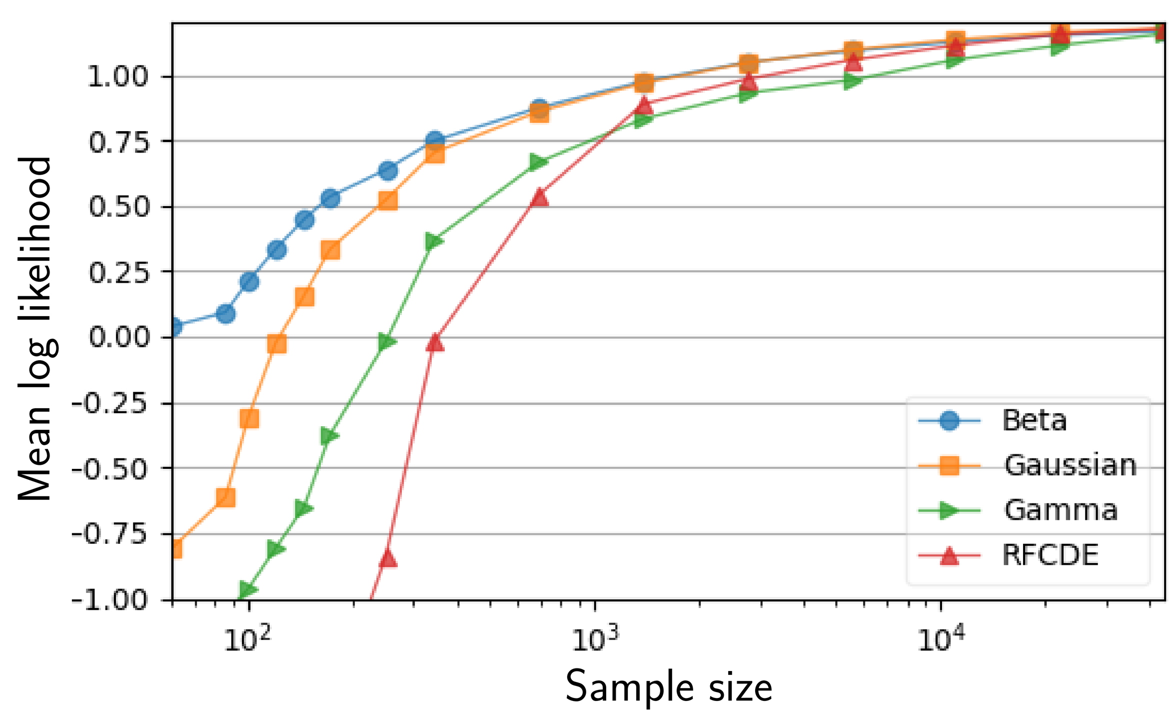
- $\Omega(m)$ speedup to compute optimal split
- Obviate storing training labels in CaDET-forests

Visualizing Gaussian-CaDET Quantiles



CaDET on Simplicial Spaces

- *Task*: Estimate baseball batting statistics
- Player attributes → (dist over) batting avgs
- Three CaDET variants vs nonparametric RFCDE



CaDET on the UCI Repository

Dataset	Gaussian MLL	Gaussian Size	Multi-$\mathcal{F}$ MLL	Multi-$\mathcal{F}$ Size	RFCDE MLL	RFCDE Size
air-quality	-12.493	2338	-12.241	1990	-15.078	25144
anneal-U	-21.718	459	-20.342	156	-29.114	4578
australian	-8.640	135	-7.630	66	-9.516	1758
auto	-28.490	100	-28.163	50	-29.073	870
balance-scale	-6.168	42	-6.168	36	-6.704	2124
breast	-5.622	238	-5.408	120	-7.070	2376
breast-cancer	-9.646	63	-9.176	30	-9.593	729
cars	-5.277	35	-5.250	29	-5.700	666
cleve	-19.040	100	-18.682	50	-19.518	1285
crx	-24.627	300	-21.487	130	-24.232	2930
diabetes	-25.802	595	-25.905	343	-25.766	4564
german	-18.434	294	-17.856	176	-17.561	3400
german-org	-12.167	207	-11.821	114	-11.832	2550
heart	-21.764	135	-21.304	60	-21.603	1374
hypothyroid	-12.765	1078	-12.759	672	-13.190	10752
iris	-3.025	14	-3.025	14	-3.650	508
winequality	-6.691	6468	-6.867	4368	-8.340	44176
Mean	-14.257	741	-13.766	494	-15.149	6458
# Optimal	3	1	14	17	2	0