CONSTRUCTING EXPLAINABILITY
TRR 318

UNIVERSITÄT BIELEFELD

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Incremental Permutation Feature Importance (iPFI): Towards Online Explanations on Data Streams

Fabian Fumagalli[1,*], Maximilian Muschalik[2,*], Eyke Hüllermeier[2], and Barbara Hammer[1]

ffumagalli@techfak.uni-bielefeld.de
maximilian.muschalik@lmu.de

[1] Bielefeld University, [2] LMU Munich, [*] equal contribution

Collaboration

Fabian Fumagalli [1,*]

Maximilian Muschalik [2,*]

Eyke Hüllermeier [2]

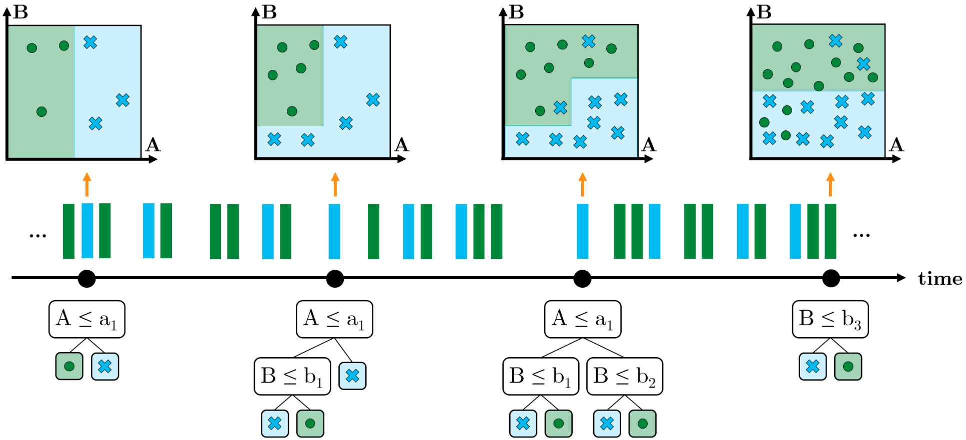
Barbara Hammer [1]

[1] UNIVERSITÄT BIELEFELD

[2] LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

* denotes equal contribution

Models in Flux: Incremental Learning from Data Streams



Various applications: Bifet and Gavaldà (2007), Gama et al. (2014), Davari et al. (2021), etc.

Examples of Models in Flux

Fraud Detection

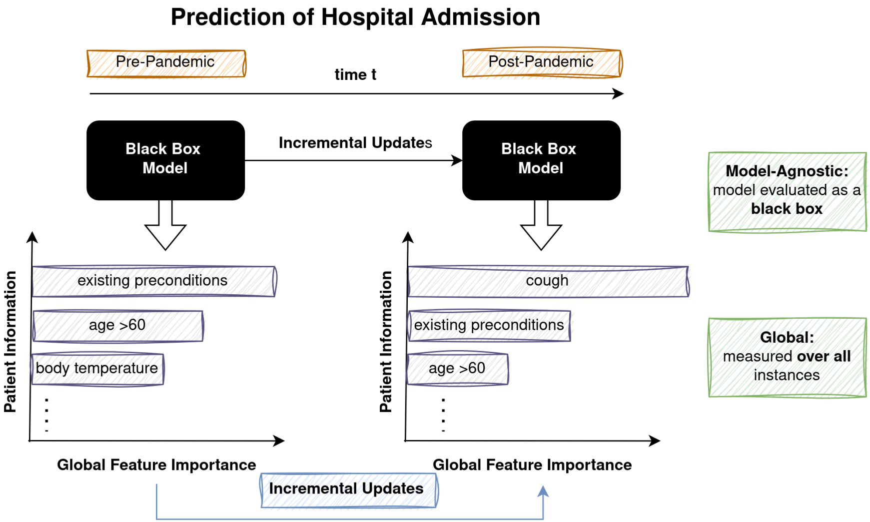
Sensor Networks

Automotive Industry

Predictive Maintenance

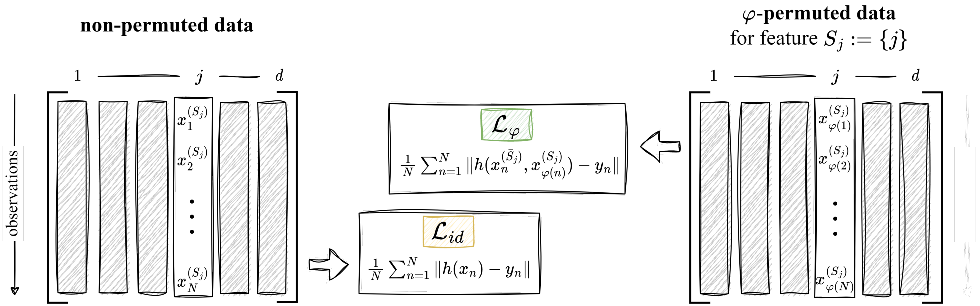
Images generated with `Leonardo.ai`.

Model-Agnostic Explanations with Global Feature Importance



Permutation Feature Importance (PFI)

Permutation Feature Importance (PFI)



Permutation Feature Importance – (Empirical) PFI

Sample permutations $\varphi_1, \ldots, \varphi_M$ uniformly and compute loss increase $\hat{\phi}_\varphi^{(S_j)} := \mathcal{L}_\varphi - \mathcal{L}_{\text{id}}$

$$\textbf{(Empirical) PFI: } \hat{\phi}^{(S_j)} := \frac{\mathbf{N}}{\mathbf{N-1}} \frac{1}{M} \sum_{m=1}^{M} \hat{\phi}_{\varphi_m}^{(S_j)}$$

Theoretical Properties of PFI

Global Feature Importance (Global FI) of a feature (set) S_j

Let $f_{S_j}(x^{(\bar{S}_j)}, y) := \mathbb{E}\left[\|h(x^{(\bar{S}_j)}, X^{(S_j)}) - y\|\right]$, then global FI is defined as

$$\phi^{(S_j)}(h) := \underbrace{\mathbb{E}_{(X,Y)}\left[f_{S_j}(X^{(\bar{S}_j)}, Y)\right]}_{\text{marginalized risk over } S_j} - \underbrace{\mathbb{E}_{(X,Y)}\left[\|h(X) - Y\|\right]}_{\text{risk}}$$

Theoretical Properties of PFI

Global Feature Importance (Global FI) of a feature (set) S_j

Let $f_{S_j}(x^{(\bar{S}_j)}, y) := \mathbb{E}\left[\|h(x^{(\bar{S}_j)}, X^{(S_j)}) - y\|\right]$, then global FI is defined as

$$\phi^{(S_j)}(h) := \underbrace{\mathbb{E}_{(X,Y)}\left[f_{S_j}(X^{(\bar{S}_j)}, Y)\right]}_{\text{marginalized risk over } S_j} - \underbrace{\mathbb{E}_{(X,Y)}\left[\|h(X) - Y\|\right]}_{\text{risk}}$$

Model Reliance Fisher, Rudin, and Dominici (2019)

$$\bar{\phi}^{(S_j)} = \underbrace{\frac{1}{N(N-1)} \sum_{n=1}^{N} \sum_{m \neq n} \|h(x_n^{(\bar{S}_j)}, x_m^{(S_j)}) - y_n\|}_{=: \hat{e}_{\text{switch}}} - \underbrace{\frac{1}{N} \sum_{n=1}^{N} \|h(x_n) - y_n\|}_{=: \hat{e}_{\text{orig}}}$$

- is a U-statistic, in particular an **unbiased estimator of global FI**
- is asymptotically Normal with finite sample boundaries

Theoretical Properties of PFI

Theorem (PFI and Model Reliance are directly linked)

Model reliance is the expectation of PFI over uniformly drawn permutations:

$$\bar{\phi}^{(S_j)} = \mathbb{E}_{\varphi \sim \text{unif}(\mathfrak{S}_N)}[\hat{\phi}^{(S_j)}] = \frac{N}{N-1} \mathbb{E}_{\varphi \sim \text{unif}(\mathfrak{S}_N)} \left[\hat{\phi}_{\varphi}^{(S_j)}\right].$$

Theoretical Properties of PFI

Theorem (PFI and Model Reliance are directly linked)

Model reliance is the expectation of PFI over uniformly drawn permutations:

$$\bar{\phi}^{(S_j)} = \mathbb{E}_{\varphi\sim\text{unif}(\mathfrak{S}_N)}[\hat{\phi}^{(S_j)}] = \frac{N}{N-1}\mathbb{E}_{\varphi\sim\text{unif}(\mathfrak{S}_N)}\left[\hat{\phi}_{\varphi}^{(S_j)}\right].$$

PFI $\hat{\phi}^{(S_j)}$ variant of Breiman (2001)

- Easy to compute in $\mathcal{O}(N)$
- Difficult to analyze theoretically due to dependence on permutations
- **Used for computation**

Expected PFI $\bar{\phi}^{(S_j)} = \mathbb{E}_{\varphi}[\hat{\phi}^{(S_j)}]$

- Hard to compute in $\mathcal{O}(N^2)$
- U-statistic with theoretical guarantees Fisher, Rudin, and Dominici (2019)
- **Used for theoretical analysis**

Incremental Permutation Feature Importance (iPFI)

Towards Online Explanations on Data Streams

Incremental PFI for Online Learning

Online Learning on Data Streams

- Unlimited data stream $(x_0, y_0), \ldots, (x_t, y_t), \ldots$
- Incrementally updated model: $h_{t+1} \leftarrow \text{incrementalUpdate}(h_t, x_t, y_t)$

Static Permutation Tests

$$\phi_\varphi^{(S_j)} = \mathcal{L}_\varphi - L_{\text{id}} = \frac{1}{N}\sum_{n=1}^{N} \|h(x_n^{(\bar{S}_j)}, x_{\varphi(n)}^{(S_j)}) - y_n\| - \|h(x_n) - y_n\|$$

At time t with (x_t, y_t) and model h_t

Stochastic Sampling Strategy

$$\varphi_t : \Omega \to \{0, \ldots, t-1\}$$

Replacement with previous Observations

$$\|h_t(x_t^{(\bar{S}_j)}, x_{\varphi_t}^{(S_j)}) - y_t\| - \|h_t(x_t) - y_t\|$$

Incremental PFI for Online Learning

Online Learning on Data Streams

- Unlimited data stream $(x_0, y_0), \ldots, (x_t, y_t), \ldots$
- Incrementally updated model: $h_{t+1} \leftarrow \text{incrementalUpdate}(h_t, x_t, y_t)$

Calculation at time t

$$\hat{\lambda}_t^{(S_j)}(x_t, x_{\varphi_t}, y_t) := \|h_t(x_t^{(\bar{S}_j)}, x_{\varphi_t}^{(S_j)}) - y_t\| - \|h_t(x_t) - y_t\|$$

Initial Computation

$$\hat{\phi}_{t_0-1}^{(S_j)} := 0 \text{ for } t \geq t_0 > 0$$

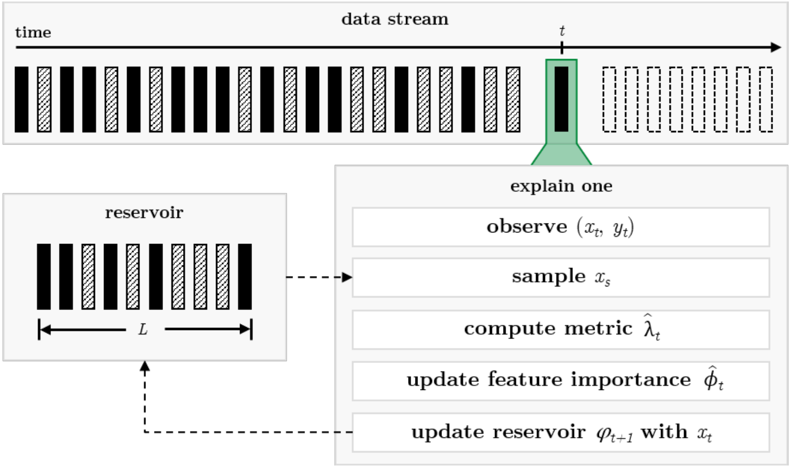
Incremental Update of iPFI

$$\hat{\phi}_t^{(S_j)} := (1-\alpha) \cdot \hat{\phi}_{t-1}^{(S_j)} + \alpha \cdot \hat{\lambda}_t^{(S_j)}(x_t, x_{\varphi_t}, y_t)$$

Smoothing Parameter

$$\alpha \in (0, 1)$$

iPFI – Algorithm Illustration

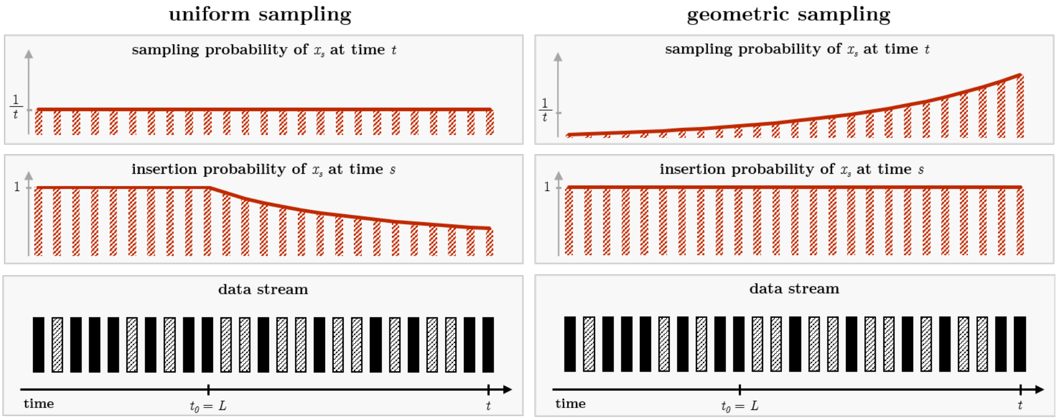


Similar Computational Complexity at time t

one static PFI score

all anytime iPFI scores

iPFI – Incremental Reservoir Sampling



iPFI – Theoretical Guarantees in Static Environments

Expected iPFI

With a (stochastic) sampling strategy $\varphi := (\varphi_s)_{s=t_0,\dots,t}$, we define

Expected iPFI: $\bar{\phi}_t^{(S_j)} := \mathbb{E}_\varphi[\hat{\phi}_t^{(S_j)}]$.

Theorem (Static Model and $(X_t, Y_t) \sim \mathbb{P}_{(X,Y)}$)

If $h \equiv h_t$ and $\mathbb{V}[\|h(X_s^{(\bar{S}_j)}, X_r^{(S_j)}) - Y_s\| - \|h(X_s) - Y_s\|] < \infty$, then

$$\phi^{(S_j)}(h) - \mathbb{E}[\bar{\phi}_t^{(S_j)}] = (1-\alpha)^{t-t_0+1}\phi^{(S_j)}(h) \qquad \text{(bias)}$$

$$\mathbb{V}\left[\lim_{t\to\infty} \bar{\phi}_t^{(S_j)}\right] = \mathcal{O}(-\alpha\log(\alpha)) \qquad \text{(uniform sampling)}$$

$$\mathbb{V}\left[\lim_{t\to\infty} \bar{\phi}_t^{(S_j)}\right] = \mathcal{O}(\alpha) + \mathcal{O}(1/L) \qquad \text{(geometric sampling)}$$

iPFI – Theoretical Guarantees in Dynamic Environments

Controlling Change in Dynamic Environments

We define a **measure of change** between two timesteps $t_0 \leq s \leq t$ as

$$f_S^{\Delta}(x^{(\bar{S}_j)}, h_s, h_t) := \mathbb{E}_{\tilde{X} \sim \mathbb{P}_S}[\|h_t(x^{(\bar{S}_j)}, \tilde{X}) - h_s(x^{(\bar{S}_j)}, \tilde{X})\|]$$

$$\Delta_S(h_s, h_t) := \mathbb{E}_X[f_S^{\Delta}(X, h_s, h_t)] \text{ and } \Delta(h_s, h_t) := \Delta_{\emptyset}(h_s, h_t).$$

Theorem (Changing Model and $(X_t, Y_t) \sim \mathbb{P}_{(X,Y)}$)

If $\Delta(h_s, h_t) \leq \delta$ and $\Delta_S(h_s, h_t) \leq \delta_S$ for $t_0 \leq s \leq t$ and finite covariances, then

$$|\mathbb{E}[\bar{\phi}_t^{(S_j)}] - \phi^{(S_j)}(h_t)| \leq \delta_S + \delta + \mathcal{O}((1-\alpha)^t) \quad \text{(bias)}$$

$$\mathbb{V}\left[\lim_{t \to \infty} \bar{\phi}_t^{(S_j)}\right] = \mathcal{O}(-\alpha \log(\alpha)) \quad \text{(uniform sampling)}$$

$$\mathbb{V}\left[\lim_{t \to \infty} \bar{\phi}_t^{(S_j)}\right] = \mathcal{O}(\alpha) + \mathcal{O}(1/L) \quad \text{(geometric sampling)}$$

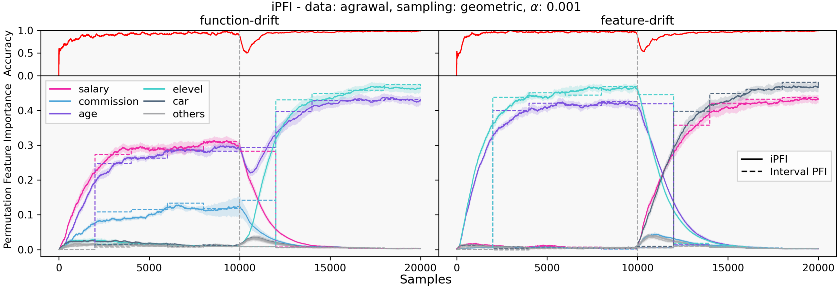
iPFI vs. Interval PFI for Concept Drifts

$h_{t_1} \rightarrow h_{t_2}$ **Function-drift**

Interval PFI
Repeated calculation over sliding window

$$(X_{t_1}^{(\{1\})}, \ldots, X_{t_1}^{(\{i\})}, \ldots, X_{t_1}^{(\{j\})}, \ldots, X_{t_1}^{(\{d\})}) \rightarrow (X_{t_2}^{(\{1\})}, \ldots, X_{t_2}^{(\{j\})}, \ldots, X_{t_2}^{(\{i\})}, \ldots, X_{t_2}^{(\{d\})})$$

Feature-drift



Conclusion & Outlook

Conclusion

Permutation Feature Importance

- (Empirical) PFI as a variant of permutation test (Breiman 2001)
- Expected PFI as model reliance (Fisher, Rudin, and Dominici 2019)
- **Expected PFI is the expectation of PFI over uniformly sampled permutations**

Incremental Permutation Feature Importance (iPFI)

- We introduce online explanations for online learning on data streams
- We propose an efficient incremental computation of PFI
- **iPFI efficiently reveals model and distribution changes over time**
- **iPFI is supported by theoretical guarantees in controlled environments**

The Road Ahead and Open Source Implementation

Towards Explaining Change

- iPFI is a **model-agnostic** XAI method to compute **global FI** for models **in flux**.
- Online XAI approaches include **iSAGE** (today at **16:30-18:30** here in room **Fucine**) and **iPDP** (xAI'23).

Workshop Friday Afternoon Slot

- Time: **14:00-18:00**
- Room: **PoliTo Room 10i**
- Title: *Explainable Artificial Intelligence: From Static to Dynamic*

iXAI

docs passing | pypi v0.1.3 | status alpha | License MIT

Installation

```
pip install ixai
```

Quickstart

```
>>> for (n, (x, y)) in enumerate(stream, start=1)
...     accuracy.update(y, model.predict_one(x))    # inference
...     incremental_pfi.explain_one(x, y)           # explaining
...     model.learn_one(x, y)                       # learning
```

References

- Bifet, Albert and Ricard Gavaldà (2007). "Learning from Time-Changing Data with Adaptive Windowing". In: *Proceedings of the Seventh SIAM International Conference on Data Mining (SIAM 2007)*, pp. 443–448. DOI: 10.1137/1.9781611972771.42.
- Breiman, Leo (2001). "Random Forests". In: *Machine Learning* 45.1, pp. 5–32.
- Davari, Narjes et al. (2021). "Predictive Maintenance Based on Anomaly Detection Using Deep Learning for Air Production Unit in the Railway Industry". In: *8th IEEE International Conference on Data Science and Advanced Analytics (DSAA 2021)*. IEEE, pp. 1–10. DOI: 10.1109/DSAA53316.2021.9564181.
- Fisher, Aaron, Cynthia Rudin, and Francesca Dominici (2019). "All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously". In: *Journal of Machine Learning Research* 20.177, pp. 1–81.
- Gama, João et al. (2014). "A Survey on Concept Drift Adaptation". In: *ACM Comput. Surv.* 46.4, 44:1–44:37. DOI: 10.1145/2523813.

Explanation Procedure

General Explanation Algorithm

Algorithm 6 Incremental explanation procedure

Require: stream $\{x_t, y_t\}_{t=1}^{\infty}$, model $f(.)$, loss function $\mathcal{L}(.)$

1: **for all** $(x_t, y_t) \in$ stream **do**
2: $\quad \hat{y}_t \leftarrow f_t(x_t)$
3: $\quad \hat{\phi}_t \leftarrow$ `explain_one`(x_t, y_t)
4: $\quad f_{t+1} \leftarrow$ `learn_one`$(\mathcal{L}(\hat{y}_t, y_t))$
5: **end for**

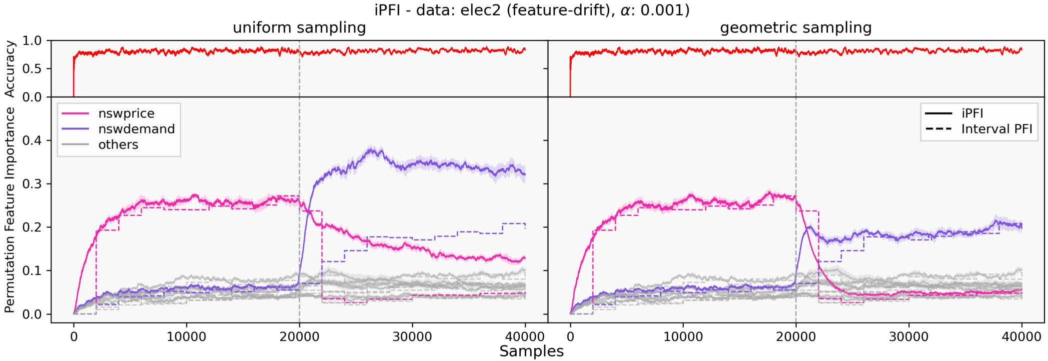
- Similarly to the **prequential** training, we explain models prequentially.
- Data points are used first for explanations (model has not seen the observation, line 3) and then the model is allowed to use it for training (line 4).

Computational Complexity

data	*stagger*	*elec2*	*agrawal*	*adult*	*bank*	*insects*	*ozone*
feature count	3	8	9	14	16	33	72
explanation time	0.734 (.017)	1.210 (.039)	1.411 (.020)	1.976 (.118)	2.386 (.048)	5.070 (.078)	7.717 (.182)
inference time	0.959 (.001)	0.989 (.002)	0.987 (.001)	0.991 (.002)	0.991 (.001)	0.990 (.021)	0.998 (.000)

Table 1: Summary of the additional time complexity of iPFI. The additional *explanation time* is given relatively to the case where the models are trained without explaining. The *inference time* denotes the portion of the explanation time in which the models are queried. All values for each dataset are derived from ten independent runs. The run time of iPFI scales *linearly* with $0.104 \cdot |D|$ over the number of features ($R^2 = 0.966$).

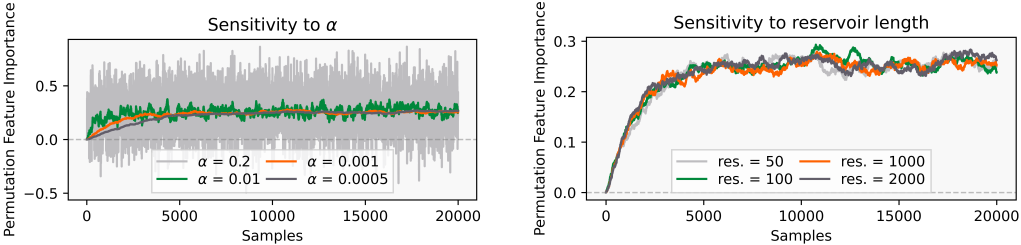
Uniform vs. Geometric Sampling



Geometric Sampling for Feature-Drift

If feature distributions change, then geometric sampling should be preferred.

Parameters



Choice of Smoothing Parameter α

The choice depends on the application. We recommend

$$\alpha = 0.001 \text{ (conservative) and } \alpha = 0.01 \text{ (reactive).}$$