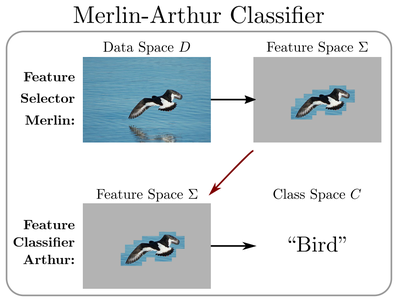
Interpretability Guarantees with Merlin-Arthur Classifiers

Stephan Wäldchen, Kartikey Sharma, Berkant Turan, Max Zimmer, and Sebastian Pokutta

Interactive Optimization and Learning
Zuse Institute Berlin & Technical University of Berlin

Formal Description



The Merlin-Arthur Classifier consists of two interactive agents, Merlin and Arthur, who communicate via a feature exchange. This feature forms the basis for classification interpretation.

Given a dataset D with a ground truth class map $c : D \to C$, where C denotes the set of classes, Merlin (M) acts as a **feature selector**, mapping each data point $\mathbf{x} \in D$ to a feature $\mathbf{z} \in \Sigma$. Arthur (A), the **feature classifier**, then assigns a class to each feature $\mathbf{z}$ from Merlin.

Feature quality is quantified using mutual information:

$$I_{x\sim\mathcal{D}}(c(\mathbf{x}); \mathbf{z} \subseteq \mathbf{x}) = H_{x\sim\mathcal{D}}(c(\mathbf{x})) - H_{x\sim\mathcal{D}}(c(\mathbf{x}) \,|\, \mathbf{z} \subseteq \mathbf{x})$$

where H denotes the class conditional entropy and $\mathcal{D}$ the data distribution.

Problem of Cheating

Merlin selects the features, but it's crucial to verify they represent true class distinctions and not just spurious correlations. Figure 1 demonstrates how "*cheating*" can artificially inflate mutual information.

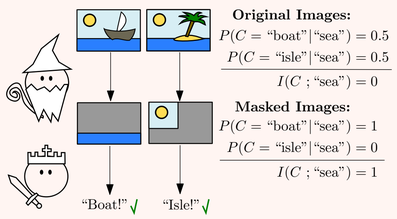


Figure 1. Illustration of "cheating" behavior: In the original dataset, "sea" and "sky" features appear equally in both "boat" and "island" classes. However, in Merlin's modified images, "sea" is only visible in "boat" images and "sky" in "island" images. This strong correlation enables Merlin to classify images using uninformative features, contrary to the idea of an interpretable classifier.

Scan Me

ArXiv Link

GitHub Repository

How Can We Identify the Most Relevant Features for Classification?

Evolution of Strategies

To prevent *cheating*, we introduce an **adversarial feature selector (Morgana, $\hat{M}$)** with the objective to convince Arthur of the wrong class. In essence, our theory shows that the only strategy that Merlin and Arthur can use is one that cannot be exploited by Morgana, see Figure 2 for an illustration.

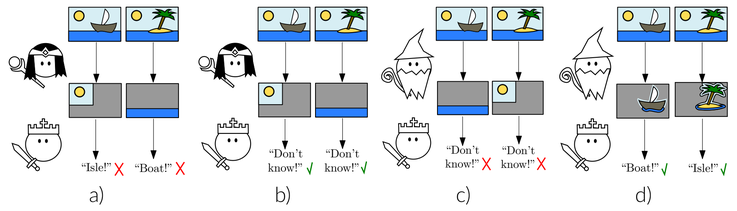


Figure 2. Evolution of feature selection for binary classification of "boats" vs. "isles." (a) Morgana exploits Arthur's expectations with "sky" and "sea" features. (b) Arthur avoids concrete classifications. (c) Arthur's uncertainty hampers cooperation with Merlin. (d) Merlin adapts with unambiguous features.

Given the adversarial influence of Morgana, we introduce two key metrics to evaluate our classifiers:

$$\begin{aligned} \textit{Completeness}: &\quad \mathbb{P}_{\mathbf{x}\sim\mathcal{D}}[A(M(\mathbf{x})) = c(\mathbf{x})], \\ \textit{Soundness}: &\quad 1 - \max_{l\in C\setminus\{c(\mathbf{x})\}} \mathbb{P}_{\mathbf{x}\sim\mathcal{D}}[A(\widehat{M}(\mathbf{x})) = l]. \end{aligned}$$

These metrics allow us to assess the accuracy and robustness of our setup involving Merlin (M) and Morgana ($\hat{M}$). More improtantly, they enable us to bound the **average precision**, which can be used to bound the *average* conditional entropy.

Quantitative Bound of Average Precision

Given a data point $\boldsymbol{x}$ with feature $\boldsymbol{z}$, precision is defined as $\mathrm{Pr}(\boldsymbol{z}) := \mathbb{P}_{\boldsymbol{y}\sim\mathcal{D}}[c(\mathbf{y}) = c(\mathbf{x}) \,|\, \mathbf{z} \subseteq \mathbf{y}]$ and has already been introduced in the context of interpretability. We extend this definition to a feature selector as follows:

Definition (Average Precision)

For a given two-class data space $\mathfrak{D}$ and a feature selector $M \in \mathcal{M}(D)$, we define the *average precision* of M with respect to $\mathcal{D}$ as

$$\mathrm{Pr}_{\mathcal{D}}(M) := \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbb{P}_{\mathbf{y}\sim\mathcal{D}}[c(\mathbf{y}) = c(\mathbf{x}) \,|\, M(\mathbf{x}) \subseteq \mathbf{y}]].$$

Theorem

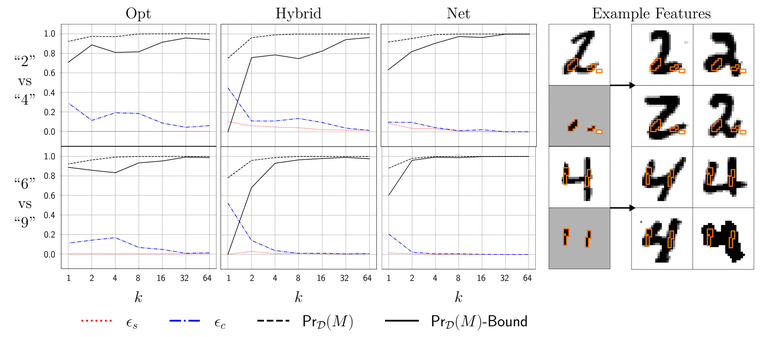
Let $(D, \mathcal{D}, c)$ have Asymmetric Feature Correlation (AFC) of κ and class imbalance B. Let $A \in \mathcal{A}$, M and $\widehat{M} \in \mathcal{M}(D)$ such that $\widehat{M}$ has a *rel. success rate* α with respect to A, M. Define

1. *Completeness*: $\min_{l\in\{-1,1\}} \mathbb{P}_{\mathbf{x}\sim\mathcal{D}_l}[A(M(\mathbf{x})) = c(\mathbf{x})] \geq 1 - \epsilon_c\,,$
2. *Soundness*: $\max_{l\in\{-1,1\}} \mathbb{P}_{\mathbf{x}\sim\mathcal{D}_l}[A((\mathbf{x})) = -c(\mathbf{x})] \leq \epsilon_s\,.$

Then it follows that

$$\mathrm{Pr}_{\mathcal{D}}(M) \geq 1 - \epsilon_c - \frac{\alpha\kappa\epsilon_s}{1 - \epsilon_c + \alpha\kappa\epsilon_s B^{-1}}\,.$$

Numerical Experiments - Evaluation of Bounds on MNIST



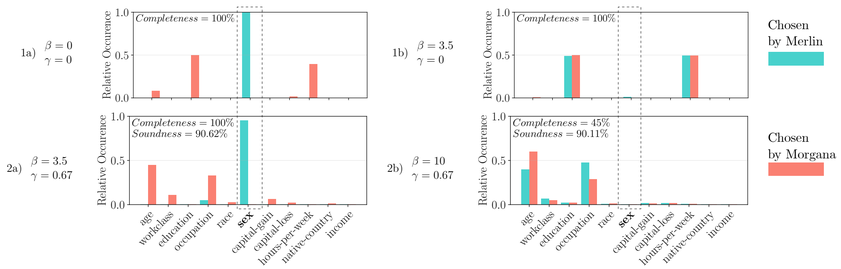
Left Panel:

- Comparison of the lower bounds on $\mathrm{Pr}_{\mathcal{D}}(M)$ with experimental results for three setups of Merlin and Morgana.
- Analysis includes different masks sizes k; bound tightness decreases with smaller k.
- Top row features labels "2" and "4", bottom row features labels "6" and "9".

Right Panel:

- Display of features selected by Merlin for individual images.
- For label "4": 13 MNIST images share the feature, all from class "4" (four shown as examples).
- For label "2": 16 images share the feature, all from class "2".

Preventing Manipulation



Scenario Context: Exploring the impact of hiding "sex" as a discriminatory factor in hiring.

1. No Soundness Required ($\gamma = 0$):

   (a) Merlin's Strategy when Unpunished ($\beta = 0$):
   - Merlin has no punishment for showing "sex". He sends Arthur the "sex" feature and they discriminate successfully (high completeness).

   (b) Merlin's Strategy when Incentivised ($\beta > 0$):
   - Merlin avoids using "sex" directly, instead conveying the "sex" to Arthur via other features such as "hours per week" and "education".
   - Morgana can exploit this strategy using the same features but switched.

2. High Soundness Required ($\gamma = 0.67$):

   (a) Merlin's Choice with Moderate Punishment ($\beta = 3.5$):
   - Despite the punishment, Merlin shows the "sex" feature, achieving high completeness.

   (b) Merlin's Choice with High Punishment ($\beta = 10$):
   - Merlin hides the "sex" feature, reducing completeness to below 50%, effectively ceasing the discrimination.