Examples are not Enough, Learn to Criticize! Criticism for Interpretability

Been Kim*
Allen Institute for AI
beenkim@csail.mit.edu

Rajiv Khanna
UT Austin
rajivak@utexas.edu

Oluwasanmi Koyejo
UIUC
sanmi@illinois.edu

Abstract

Example-based explanations are widely used in the effort to improve the interpretability of highly complex distributions. However, prototypes alone are rarely sufficient to represent the gist of the complexity. In order for users to construct better mental models and understand complex data distributions, we also need *criticism* to explain what are *not* captured by prototypes. Motivated by the Bayesian model criticism framework, we develop `MMD-critic` which efficiently learns prototypes and criticism, designed to aid human interpretability. A human subject pilot study shows that the `MMD-critic` selects prototypes and criticism that are useful to facilitate human understanding and reasoning. We also evaluate the prototypes selected by `MMD-critic` via a nearest prototype classifier, showing competitive performance compared to baselines.

1 Introduction and Related Work

As machine learning (ML) methods have become ubiquitous in human decision making, their transparency and interpretability have grown in importance (Varshney, 2016). Interpretability is particularity important in domains where decisions can have significant consequences. For example, the pneumonia risk prediction case study in Caruana et al. (2015) showed that a more interpretable model could reveal important but surprising patterns in the data that complex models overlooked.

Studies of human reasoning have shown that the use of examples (prototypes) is fundamental to the development of effective strategies for tactical decision-making (Newell and Simon, 1972; Cohen et al., 1996). Example-based explanations are widely used in the effort to improve interpretability. A popular research program along these lines is case-based reasoning (CBR) (Aamodt and Plaza, 1994), which has been successfully applied to real-world problems (Bichindaritz and Marling, 2006). More recently, the Bayesian framework has been combined with CBR-based approaches in the unsupervised-learning setting, leading to improvements in user interpretability (Kim et al., 2014). In a supervised learning setting, example-based classifiers have been is shown to achieve comparable performance to non-interpretable methods, while offering a condensed view of a dataset (Bien and Tibshirani, 2011).

However, examples are not enough. Relying only on examples to explain the models' behavior can lead over-generalization and misunderstanding. Examples alone may be sufficient when the distribution of data points are 'clean' – in the sense that there exists a set of prototypical examples which sufficiently represent the data. However, this is rarely the case in real world data. For instance, fitting models to complex datasets often requires the use of regularization. While the regularization adds bias to the model to improve generalization performance, this same bias may conflict with the distribution of the data. Thus, to maintain interpretability, it is important, along with prototypical examples, to deliver insights signifying the parts of the input space where prototypical examples

*All authors contributed equally.

29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

do not provide good explanations. We call the data points that do not quite fit the model *criticism* samples. Together with prototypes, criticism can help humans build a better mental model of the complex data space.

Bayesian model criticism (BMC) is a framework for evaluating fitted Bayesian models, and was developed to to aid model development and selection by helping to identify where and how a particular model may fail to explain the data. It has quickly developed into an important part of model design, and Bayesian statisticians now view model criticism as an important component in the cycle of model construction, inference and criticism (Gelman et al., 2014). Lloyd and Ghahramani (2015) recently proposed an exploratory approach for statistical model criticism using the maximum mean discrepancy (MMD) two sample test, and explored the use of the *witness function* to identify the portions of the input space the model most misrepresents the data. Instead of using the MMD to compare two models as in classic two sample testing (Gretton et al., 2008), or to compare the model to input data as in the Bayesian model criticism of Lloyd and Ghahramani (2015), we consider a novel application of the MMD, and its associated witness function as a principled approach for selecting *prototype* and *criticism* samples.

We present the `MMD-critic`, a scalable framework for prototype and criticism selection to improve the interpretability of machine learning methods. To our best knowledge, ours is the first work which leverages the BMC framework to generate explanations for machine learning methods. `MMD-critic` uses the MMD statistic as a measure of similarity between points and potential prototypes, and efficiently selects prototypes that maximize the statistic. In addition to prototypes, `MMD-critic` selects criticism samples i.e. samples that are not well-explained by the prototypes using a regularized witness function score. The scalability follows from our analysis, where we show that under certain conditions, the MMD for prototype selection is a *supermodular* set function. Our supermodularity proof is general and may be of independent interest. While we are primarily concerned with prototype selection and criticism, we quantitatively evaluate the performance of `MMD-critic` as a nearest prototype classifier, and show that it achieves comparable performance to existing methods. We also present results from a human subject pilot study which shows that including the criticism together with prototypes is helpful for an end-task that requires the data-distributions to be well-explained.

2 Preliminaries

This section includes notation and a few important definitions. Vectors are denoted by lower case $\mathbf{x}$ and matrices by capital $\mathbf{X}$. The Euclidean inner product between matrices $\mathbf{A}$ and $\mathbf{B}$ is given by $\langle \mathbf{A}, \mathbf{B} \rangle = \sum a_{i,j} b_{i,j}$. Let $\det(\mathbf{X})$ denote the determinant of $\mathbf{X}$. Sets are denoted by sans serif e.g. S. The reals are denoted by $\mathbb{R}$. $[n]$ denotes the set of integers $\{1, \ldots, n\}$, and 2^{V} denotes the power set of V. The indicator function $1_{[a]}$ takes the value of 1 if its argument a is true and is 0 otherwise. We denote probability distributions by either P or Q. The notation $|\cdot|$ will denote cardinality when applied to sets, or absolute value when applied to real values.

2.1 Maximum Mean Discrepancy (MMD)

The maximum mean discrepancy (MMD) is a measure of the difference between distributions P and Q, given by the suprenum over a function space $\mathcal{F}$ of differences between the expectations with respect to two distributions. The MMD is given by:

$$\text{MMD}(\mathcal{F}, P, Q) = \sup_{f \in \mathcal{F}} \left(\mathrm{E}_{X \sim P} \left[f(X) \right] - \mathrm{E}_{Y \sim Q} \left[f(Y) \right] \right). \tag{1}$$

When $\mathcal{F}$ is a reproducing kernel Hilbert space (RKHS) with kernel function $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$, the suprenum is achieved at (Gretton et al., 2008):

$$f(x) = \mathrm{E}_{X' \sim P} \left[k(x, X') \right] - \mathrm{E}_{X' \sim Q} \left[k(x, X') \right]. \tag{2}$$

The function (2) is also known as the *witness function* as it measures the maximum discrepancy between the two expectations in $\mathcal{F}$. Observe that the witness function is positive whenever Q underfits the density of P, and negative wherever Q overfits P. We can substitute (2) into (1) and square the result, leading to:

$$\text{MMD}^2(\mathcal{F}, P, Q) = \mathrm{E}_{X, X' \sim P} \left[k(X, X') \right] - 2\mathrm{E}_{X \sim P, y \sim Q} \left[k(X, Y) \right] + \mathrm{E}_{Y, Y' \sim Q} \left[k(Y, Y') \right]. \tag{3}$$

It is clear that $\text{MMD}^2(\mathcal{F}, P, Q) \geq 0$ and $\text{MMD}^2(\mathcal{F}, P, Q) = 0$ iff. P is indistinguishable from Q on the RHKS $\mathcal{F}$. This population definition can be approximated using sample expectations. In particular, given n samples from P as $\mathsf{X} = \{x_i \sim P,\, i \in [n]\}$, and m samples from Q as $\mathsf{Z} = \{z_i \sim Q,\, i \in [m]\}$, the following is a finite sample approximation:

$$\text{MMD}_b^2(\mathcal{F}, \mathsf{X}, \mathsf{Z}) = \frac{1}{n^2} \sum_{i,j \in [n]} k(x_i, x_j) - \frac{2}{nm} \sum_{i \in [n], j \in [m]} k(x_i, z_j) + \frac{1}{m^2} \sum_{i,j \in [m]} k(z_i, z_j), \quad (4)$$

and the witness function is approximated as:

$$f(x) = \frac{1}{n} \sum_{i \in [n]} k(x, x_i) - \frac{1}{m} \sum_{j \in [m]} k(x, z_j). \quad (5)$$

3 `MMD-critic` for Prototype Selection and Criticism

Given n samples from a statistical model $\mathsf{X} = \{x_i,\, i \in [n]\}$, let $\mathsf{S} \subseteq [n]$ represent a subset of the indices, so that $\mathsf{X}_\mathsf{S} = \{x_i \; \forall i \in \mathsf{S}\}$. Given a RKHS with the kernel function $k(\cdot, \cdot)$, we can measure the maximum mean discrepancy between the samples and any selected subset using $\text{MMD}^2(\mathcal{F}, X, X_\mathsf{S})$. `MMD-critic` selects prototype indices S which minimize $\text{MMD}^2(\mathcal{F}, X, X_\mathsf{S})$. For our purposes, it will be convenient to pose the problem as a *normalized* discrete maximization. To this end, consider the following cost function, given by the negation of $\text{MMD}^2(\mathcal{F}, X, X_\mathsf{S})$ with an additive bias:

$$\begin{aligned} J_b(\mathsf{S}) =& \frac{1}{n^2} \sum_{i,j=1}^{n} k(x_i, x_j) - \text{MMD}^2(\mathcal{F}, X, X_\mathsf{S}) \\ =& \frac{2}{n|\mathsf{S}|} \sum_{i \in [n], j \in \mathsf{S}} k(x_i, y_j) - \frac{1}{|\mathsf{S}|^2} \sum_{i,j \in \mathsf{S}} k(y_i, x_j). \end{aligned} \quad (6)$$

Note that the additive bias $\text{MMD}^2(\mathcal{F}, X, \emptyset) = \frac{1}{n^2} \sum_{i,j=1}^{n} k(x_i, x_j)$ is a constant with respect to S. Further, $J_b(\mathsf{S})$ is normalized, since, when evaluated on the empty set, we have that:

$$J_b(\emptyset) = \min_{\mathsf{S} \in 2^{[n]}} J_b(\mathsf{S}) = \frac{1}{n^2} \sum_{i,j=1}^{n} k(x_i, x_j) - \frac{1}{n^2} \sum_{i,j=1}^{n} k(x_i, x_j) = 0.$$

`MMD-critic` selects m_* prototypes as the subset of indices $\mathsf{S} \subseteq [n]$ which optimize:

$$\max_{\mathsf{S} \in 2^{[n]}, |\mathsf{S}| \leq m_*} J_b(\mathsf{S}). \quad (7)$$

For the purposes of optimizing the cost function (6), it will prove useful to exploit it's linearity with respect to the kernel entries. The following Lemma is easily shown by enumeration.

Lemma 1. *Let $J_b(\cdot)$ be defined as in* (6)*, then $J_b(\cdot)$ is a linear function of $k(x_i, x_j)$. In particular, define $\mathbf{K} \in R^{n \times n}$, with $k_{i,j} = k(x_i, x_j)$, and $\mathbf{A}(\mathsf{S}) \in R^{n \times n}$ with entries $a_{i,j}(\mathsf{S}) = \frac{2}{n|\mathsf{S}|} 1_{[j \in \mathsf{S}]} - \frac{1}{|\mathsf{S}|^2} 1_{[i \in \mathsf{S}]} 1_{[j \in \mathsf{S}]}$ then: $J_b(\mathsf{S}) = \langle \mathbf{A}(\mathsf{S}), \mathbf{K} \rangle$.*

3.1 Submodularity and Efficient Prototype Selection

While the discrete optimization problem (6) may be quite complicated to optimize, we show that the cost function $J_b(\mathsf{S})$ is monotone submodular under conditions on the kernel matrix which are often satisfied in practice, and which can be easily checked given a kernel matrix. Based on this result, we describe the greedy forward selection algorithm for efficient prototype selection.

Let $F : 2^{[n]} \mapsto \mathbb{R}$ represent a set function. F is *normalized* if $F(\emptyset) = 0$. F is *monotonic*, if for all subsets $\mathsf{u} \subset \mathsf{v} \subseteq 2^{[n]}$ it holds that $F(\mathsf{U}) \leq F(\mathsf{V})$. F is *submodular*, if for all subsets $\mathsf{U}, \mathsf{V} \in 2^{[n]}$ it holds that $F(\mathsf{U} \cup \mathsf{V}) + F(\mathsf{U} \cap \mathsf{V}) \leq F(\mathsf{U}) + F(\mathsf{V})$. Submodular functions have a diminishing returns property (Nemhauser et al., 1978) i.e. the marginal gain of adding elements decreases with the size of the set. When F is submodular, $-F$ is *supermodular* (and vice versa).

We prove submodularity for a larger class of problems, then show submodularity of (6) as a special case. Our proof for the larger class may be of independent interest. In particular, the following Theorem considers general discrete optimization problems which are linear matrix functionals, and shows sufficient conditions on the matrix for the problem to be monotone and/or submodular.

Theorem 2 (Monotone Submodularity for Linear Forms)**.** *Let $\mathbf{H} \in \mathbb{R}^{n\times n}$ (not necessarily symmetric) be element-wise non-negative and bounded, with upper bound $h_* = \max_{i,j\in[n]} h_{i,j} > 0$. Further, construct the binary matrix representation of the indices that achieve the maximum as $\mathbf{E} \in [0,1]^{n\times n}$ with $e_{i,j} = 1$ if $h_{i,j} = h_*$ and $e_{i,j} = 0$ otherwise, and its complement $\mathbf{E}' = 1 - \mathbf{E}$ with the corresponding set $\mathsf{E}' = \{(i,j) \text{ s.t. } e_{i,j} = 0\}$. Given the ground set $\mathcal{S} \subseteq 2^{[n]}$ consider the linear form: $F(\mathbf{H}, \mathsf{S}) = \langle \mathbf{A}(\mathsf{S}), \mathbf{H}\rangle \; \forall \; \mathsf{S} \in \mathcal{S}$. Given $m = |\mathsf{S}|$, define the functions:*

$$\alpha(n,m) = \frac{a(\mathsf{S}\cup\{u\}) - a(\mathsf{S})}{b(\mathsf{S})}, \quad \beta(n,m) = \frac{a(\mathsf{S}\cup\{u\}) + a(\mathsf{S}\cup v\}) - a(\mathsf{S}\cup\{u,v\}) - a(\mathsf{S})}{b(\mathsf{S}\cup\{u,v\}) + d(\mathsf{S})}, \tag{8}$$

where $a(\mathsf{S}) = F(\mathbf{E}, \mathsf{S})$, $b(\mathsf{S}) = F(\mathbf{E}', \mathsf{S})$ for all $u, v \in \mathcal{S}$ (additional notation suppressed in $\alpha(\cdot)$ and $\beta(\cdot)$ for clarity). Let $m_* = \max_{\mathsf{S}\in\mathcal{S}} |\mathsf{S}|$ be the maximal cardinality of any element in the ground set.

1. *If $h_{i,j} \leq h_*\alpha(n,m) \; \forall \; 0 \leq m \leq m_*, \; \forall \; (i,j) \in \mathsf{E}'$, then $F(\mathbf{H}, \mathsf{S})$ is monotone*
2. *If $h_{i,j} \leq h_*\beta(n,m) \; \forall \; 0 \leq m \leq m_*, \; \forall \; (i,j) \in \mathsf{E}'$, then $F(\mathbf{H}, \mathsf{S})$ is submodular.*

Finally, we consider a special case of Theorem 2 for the MMD.

Corollary 3 (Monotone Submodularity for MMD)**.** *Let the kernel matrix $\mathbf{K} \in \mathbb{R}^{n\times n}$ be element-wise non-negative, with equal diagonal terms $k_{i,i} = k_* > 0 \, \forall i \in [n]$, and be diagonally dominant. If the off-diagonal terms $k_{i,j} \; \forall \; i, j \in [n], \; i \neq j$ satisfy $0 \leq k_{i,j} \leq \frac{k_*}{n^3+2n^2-2n-3}$, then $J_b(\mathsf{S})$ given by (6) is monotone submodular.*

The diagonal dominance condition expressed by Corollary 3 is easy to check given a kernel matrix. We also note that the conditions can be significantly weakened if one determines the required number of prototypes $m_* = \max |\mathsf{S}| \leq n$ a-priori. This is further simplified for the MMD since the bounds (8) are both monotonically decreasing functions of m, so the condition need only be checked for m_*. Observe that diagonal dominance is not a necessary condition, as the more general approach in Theorem 2 allows arbitrarily indexed maximal entries in the kernel. Diagonal dominance is assumed to simplify the resulting expressions.

Perhaps, more important to practice is our observation that the diagonal dominance condition expressed by Corollary 3 is satisfied by parametrized kernels with appropriately selected parameters. We provide an example for radial basis function (RBF) kernels and powers of positive standardized kernels. Further examples and more general conditions are left for future work.

Example 4 (Radial basis function Kernel)**.** *Consider the radial basis function kernel $\mathbf{K}$ with entries $k_{i,j} = k(x_i, x_j) = \exp(-\gamma\|x_i - x_j\|)$ evaluated on a sample X with non-duplicate points i.e. $x_i \neq x_j \; \forall \; x_i, x_j \in \mathsf{X}$. The off-diagonal kernel entries $k_{i,j} \; i \neq j$ monotonically decrease with respect to increasing γ. Thus, $\exists \, \gamma_*$ such that Corollary 3 is satisfied for $\gamma \geq \gamma^*$.*

Example 5 (Powers of Positive Standardized Kernels)**.** *Consider a element-wise positive kernel matrix $\mathbf{G}$ standardized to be element-wise bounded $0 \leq g_{i,j} < 1$ with unitary diagonal $g_{i,i} = 1 \; \forall \; i \in [n]$. Define the kernel power $\mathbf{K}$ with $k_{i,j} = g_{i,j}^p$. The off-diagonal kernel entries $k_{i,j} \; i \neq j$ monotonically decrease with respect to increasing p. Thus, $\exists \, p_*$ such that Corollary 3 is satisfied for $p \geq p^*$.*

Beyond the examples outlined here, similar conditions can be enumerated for a wide range of parametrized kernel functions, and are easily checked for model-based kernels e.g. the Fisher kernel (Jaakkola et al., 1999) – useful for comparing data points based on similarity with respect to a probabilistic model. Our interpretation of from these examples is that the conditions of Corollary 3 are not excessively restrictive. While constrained maximization of submodular functions is generally NP-hard, the simple greedy forward selection heuristic has been shown to perform almost as well as the optimal in practice, and is known to have strong theoretical guarantees.

Theorem 6 (Nemhauser et al. (1978))**.** *In the case of any normalized, monotonic submodular function F, the set S_* obtained by the greedy algorithm achieves at least a constant fraction $\left(1 - \frac{1}{e}\right)$ of the objective value obtained by the optimal solution i.e. $F(\mathsf{S}_*) = \left(1 - \frac{1}{e}\right) \max_{|\mathsf{S}|\leq m} F(\mathsf{s})$.*

In addition, no polynomial time algorithm can provide a better approximation guarantee unless P = NP (Feige, 1998). An additional benefit of the greedy approach is that it does not require the decision of the number of prototypes m_* to be made at training time, so assuming the kernel satisfies appropriate conditions, training can be stopped at any m_* based on computational constraints, while still returning meaningful results. The greedy algorithm is outlined in Algorithm 1.

Algorithm 1 Greedy algorithm, $\max F(\mathsf{S})$ s.t. $|\mathsf{S}| \leq m_*$

Input: $m_*, \mathsf{S} = \emptyset$
while $|\mathsf{S}| < m_*$ **do**
  foreach $i \in [n] \backslash \mathsf{S}$, $f_i = F(\mathsf{S} \cup i) - F(\mathsf{S})$
  $\mathsf{S} = \mathsf{S} \cup \{\arg\max f_i\}$
end while
Return: S.

3.2 Model Criticism

In addition to selecting prototype samples, `MMD-critic` characterizes the data points not well explained by the prototypes – which we call the model *criticism*. These data points are selected as the largest values of the witness function (5) i.e. where the similarity between the dataset and the prototypes deviate the most. Consider the cost function:

$$L(\mathsf{C}) = \sum_{l \in \mathsf{C}} \left| \frac{1}{n} \sum_{i \in [n]} k(x_i, x_l) - \frac{1}{m} \sum_{j \in \mathsf{S}} k(x_j, x_l) \right|. \tag{9}$$

The absolute value ensures that we measure both positive deviations $f(x) > 0$ where the prototypes *underfit* the density of the samples, and negative deviations $f(x) < 0$, where the prototypes *overfit* the density of the samples. Thus, we focus primarily on the magnitude of deviation, rather than its sign. The following theorem shows that (9) is a linear function of C.

Theorem 7. *The criticism function $L(\mathsf{C})$ is a linear function of C.*

We found that the addition of a regularizer which encourages a diverse selection of criticism points improved performance. Let $r : 2^{[n]} \mapsto \mathbb{R}$ represent a regularization function. We select the criticism points as the maximizers of this cost function:

$$\max_{\mathsf{C} \subseteq [n] \backslash \mathsf{S}, |\mathsf{C}| \leq c_*} L(\mathsf{C}) + r(\mathbf{K}, \mathsf{C}) \tag{10}$$

Where $[n] \backslash \mathcal{S}$ denote all indexes which not include the prototypes, and c_* is the number of criticism points desired. Fortunately, due to the linearity of (5), the optimization function (10) is submodular when the regularization function is submodular. We encourage the use of regularizers which incorporate diversity into the criticism selection. We found the best qualitative performance using the log-determinant regularizer (Krause et al., 2008). Let $\mathbf{K}_{\mathsf{C},\mathsf{C}}$ be the sub-matrix of $\mathbf{K}$ corresponding to the pair of indexes in $\mathsf{C} \times \mathsf{C}$, then the log-determinant regularizer is given by:

$$r(\mathbf{K}, \mathsf{C}) = \log\det \mathbf{K}_{\mathsf{C},\mathsf{C}} \tag{11}$$

which is known to be submodular. Further, several researchers have found, both in theory and practice (Sharma et al., 2015), that greedy optimization is an effective strategy for optimization. We apply the greedy algorithm for criticism selection with the function $F(\mathsf{C}) = L(\mathsf{C}) + r(\mathbf{K}, \mathsf{C})$.

4 Related Work

There is a large literature on techniques for selecting prototypes that summarize a dataset, and a full literature survey is beyond the scope of this manuscript. Instead, we overview a few of the most relevant references. The K-medoid clustering (Kaufman and Rousseeuw, 1987) is a classic technique for selecting a representative subset of data points, and can be solved using various iterative algorithms. K-medoid clustering is quite similar to K-means clustering, with the additional condition that the presented prototypes must be in the dataset. The ubiquity of large datasets has led to resurgence

of interest in the data summarization problem, also known as the set cover problem. Progress has included novel cost functions and algorithms for several domains including image summarization (Simon et al., 2007) and document summarizauion (Lin and Bilmes, 2011). Recent innovations also include highly scalable and distributed algorithms (Badanidiyuru et al., 2014; Mirzasoleiman et al., 2015). There is also a large literature on variations of the set cover problem tuned for classification, such as the cover digraph approach of (Priebe et al., 2003) and prototype selection for interpretable classification (Bien and Tibshirani, 2011), which involves selecting prototypes that maximize the coverage within the class, but minimize the coverage across classes.

Submodular / Supermodular functions are well studied in the combinatorial optimization literature, with several scalable algorithms that come with optimization theoretic optimality guarantees (Nemhauser et al., 1978). In the Bayesian modeling literature, submodular optimization has previously been applied for approximate inference by Koyejo et al. (2014). The technical conditions required for submodularity of (6) are due to *averaging* of the kernel similarity scores – as the average requires a division by the cardinality $|\mathsf{S}|$. In particular, the analogue of (6) which replaces all the averages by sums (i.e. removes all division by $|\mathsf{S}|$) is equivalent to the well known submodular functions previously used for scene (Simon et al., 2007) and document (Lin and Bilmes, 2011) summarization, given by: $-\frac{2}{n}\sum_{i\in[n],j\in\mathsf{S}} k(x_i, y_j) + \lambda \sum_{i,j\in\mathsf{S}} k(y_i, x_j)$, where $\lambda > 0$ is a regularization parameter. The function that results is known to be submodular when the kernel is element-wise positive i.e. without the need for additional diagonal dominance conditions. On the other hand, the averaging has a desirable built-in balancing effect. When using the sum, practitioners must tune the additional regularization parameter λ to achieve a similar balance.

5 Results

We present results for the proposed technique `MMD-critic` using USPS hand written digits (Hull, 1994) and Imagenet (Deng et al., 2009) datasets. We quantitatively evaluate the prototypes in terms of predictive quality as compared to related baselines on USPS hand written digits dataset. We also present preliminary results from a human subject pilot study. Our results suggest that the model criticism – which is unique to the proposed `MMD-critic` is especially useful to facilitate human understanding. For all datasets, we employed the radial basis function (RBF) kernel with entries $k_{i,j} = k(x_i, x_j) = \exp(-\gamma\|x_i - x_j\|)$, which satisfies the conditions of Corollary 3 for sufficiently large γ (c.f. Example 4, see Example 5 and following discussion for alternative feasible kernels).

The Nearest Prototype Classifier: While our primary interest is in interpretable prototype selection and criticism, prototypes may also be useful for speeding up memory-based machine learning techniques such as the nearest neighbor classifier by restricting the neighbor search to the prototypes, sometimes known as the *nearest prototype classifier* (Bien and Tibshirani, 2011; Kuncheva and Bezdek, 1998). This classification provides an objective (although indirect) evaluation of the quality of the selected prototypes, and is useful for setting hyperparameters. We employ a 1 nearest neighbor classifier using the Hilbert space distance induced by the kernels. Let $y_i \in [k]$ denote the label associated with each prototype $i \in \mathsf{S}$, for k classes. As we employ normalized kernels (where the diagonal is 1), it is sufficient to measure the pairwise kernel similarity. Thus, for a test point $\hat{x}$, the nearest neighbor classifier reduces to:

$$\hat{y} = y_{i*}, \text{ where } i^* = \operatorname*{argmin}_{i\in\mathsf{S}} \|\hat{x} - x_i\|^2_{\mathcal{H}_K} = \operatorname*{argmax}_{i\in\mathsf{S}} k(\hat{x}, x_i).$$

5.1 `MMD-critic` evaluated on USPS Digits Dataset

The USPS hand written digits dataset Hull (1994) consists of $n = 7291$ training (and 2007 test) greyscale images of 10 handwritten digits from 0 to 9. We consider two kinds of RBF kernels (i) *global:* where the pairwise kernel is computed between all data points, and (ii) *local:* given by $\exp(-\gamma\|x_i - x_j\|)1_{[y_i=y_j]}$, i.e. points in different classes are assigned a similarity score of zero. The local approach has the effect of pushing points in different classes further apart. The kernel hyperparameter γ was chosen based to maximize the average cross-validated classification performance, then fixed for all other experiments.

Classification: We evaluated nearest prototype classifiers using `MMD-critic`, and compared to baselines (and reported performance) from Bien and Tibshirani (2011) (abbreviated as PS) and their

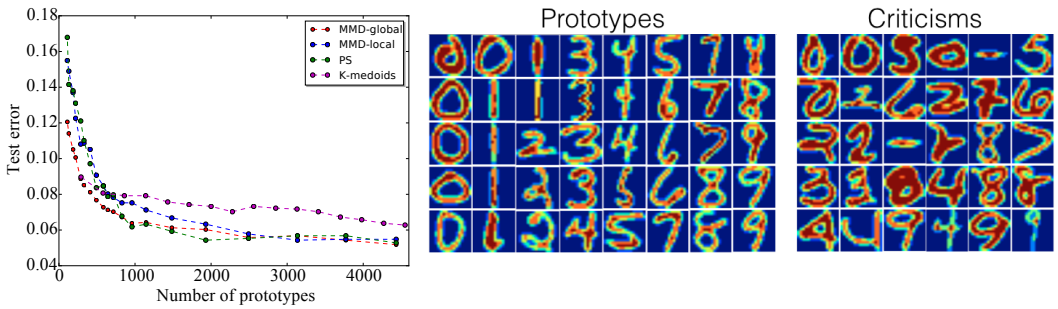

Figure 1: Classification error vs. number of prototypes $m = |S|$. `MMD-critic` shows comparable (or improved) performance as compared to other models (left). Random subset of prototypes and criticism from the USPS dataset (right).

implementation of K-medoids. Figure 1(left) compares `MMD-critic` with global and local kernels, to the baselines for different numbers of selected prototypes $m = |S|$. Our results show comparable (or improved) performance as compared to other models. In particular, we observe that the global kernels out-perform the local kernels[2] by a small margin. We note that MMD is particularly effective at selecting the first few prototypes (i.e. speed of error reduction as number of prototypes increases) suggesting its utility for rapidly summarising the dataset.

Selected Prototypes and Criticism: Fig. 1 (right) presents a randomly selected subset of the prototypes and criticism from the `MMD-critic` using the local kernel. We observe that the prototypes capture many of the common ways of writing digits, while the criticism clearly capture outliers.

5.2 Qualitative Measure: Prototypes and Criticisms of Images

In this section, we learn prototypes and criticisms from the Imagenet dataset (Russakovsky et al., 2015) using image embeddings from He et al. (2015). Each image is represented by a 2048 dimensions vector embedding, and each image belongs to one of 1000 categories. We select two breeds of one category (e.g., Blenheim spaniel) and run `MMD-critic` to learn prototypes and criticism. As shown in Figure 2, `MMD-critic` learns reasonable prototypes and criticisms for two types of dog breeds. On the left, criticisms picked out the different coloring (second criticism is in black and white picture), as well as pictures capturing movements of dogs (first and third criticisms). Similarly, on the right, criticisms capture the unusual, but potentially frequent pictures of dogs in costumes (first and second criticisms).

5.3 Quantitative measure: Prototypes and Criticisms improve interpretability

We conducted a human pilot study to collect objective and subjective measures of interpretability using `MMD-critic`. The experiment used the same dataset as Section 5.2. We define 'interpretability' in this work as the following: a method is interpretable if a user can correctly and efficiently predict the method's results. Under this definition, we designed a predictive task to quantitatively evaluate the interpretability. Given a randomly sampled data point, we measure how well a human can predict a group it belongs to (accuracy), and how fast they can perform the task (efficiency). We chose this dataset as the task of assigning a new image to a group requires groups to be well-explained but does not require specialized training.

We presented four conditions in the experiment. 1) raw images condition (Raw Condition) 2) Prototypes Only (Proto Only Condition) 3) Prototypes and criticisms (Proto and Criticism Condition) 4) Uniformly sampled data points per group (Uniform Condition). Raw Condition contained 100 images per species (e.g., if a group contains 2 species, there are 200 images) Proto Only Condition, Proto and Criticism Condition and Uniform Condition contains the same number of images.

[2] Note that the local kernel trivially achieves perfect accuracy. Thus, in order to measure generalization performance, we do not use class labels for local kernel *test* instances i.e. we use the global kernel instead of local kernel for *test* instances – regardless of training.

Figure 2: Learned prototypes and criticisms from Imagenet dataset (two types of dog breeds)

We used within-subject design to minimize the effect of inter-participant variability, with a balanced Latin square to account for a potential learning effect. The four conditions were assigned to four participants (four males) in a balanced manner. Each subject answered 21 questions, where the first three questions are practice questions and not included in the analysis. Each question showed six groups (e.g., red fox, kit fox) of a species (e.g., fox), and a randomly sampled data point that belongs to one of the groups. Subjects were encouraged to answer the questions as quickly and accurately as possible. A break was imposed after each question to mitigate the potential effect of fatigue. We measured the accuracy of answers as well as the time they took to answer each question. Participants were also asked to respond to 10 5-point Likert scale survey questions about their subjective measure of accuracy and efficiency. Each survey question compared a pair of conditions (e.g., Condition A was more helpful than condition B to correctly (or efficiently) assign the image to a group).

Subjects performed the best using Proto and Criticism Condition (M=87.5%, SD=20%). The performance with Proto Only Condition was relatively similar (M=75%, SD=41%), while that with Uniform Condition (M=55%, SD=38%, 37% decrease) and Raw Condition (M=56%, SD=33%, 36% decrease) was substantially lower. In terms of speed, subjects were most efficient using Proto Only Condition (M=1.04 mins/question, SD=0.28, 44% decrease compared to Raw Condition), followed by Uniform Condition (M=1.31 mins/question, SD=0.59) and Proto and Criticism Condition (M=1.37 mins/question, SD=0.8). Subjects spent the most time with Raw Condition (M=1.86 mins/question, SD=0.67).

Subjects indicated their preference of Proto and Criticism Condition over Raw Condition and Uniform Condition. In a survey question that asks to compare Proto and Criticism Condition and Raw Condition, a subject added that "[Proto and Criticism Condition resulted in] less confusion from trying to discover hidden patterns in a ton of images, more clues indicating what features are important". In particular, in a question that asks to compare Proto and Criticism Condition and Proto Only Condition, a subject said that "The addition of criticisms made it easier to locate the defining features of the cluster within the prototypical images". The humans' superior performance with prototypes and criticism in this preliminary study shows that providing criticisms together with prototypes is a promising direction to improve the interpretability.

6 Conclusion

We present the `MMD-critic`, a scalable framework for prototype and criticism selection to improve the interpretability of complex data distributions. To our best knowledge, ours is the first work which leverages the BMC framework to generate explanations. Further, `MMD-critic` shows competitive performance as a nearest prototype classifier compared to to existing methods. When criticism is given together with prototypes, a human pilot study suggests that humans are better able to perform a predictive task that requires the data-distributions to be well-explained. This suggests that *criticism* and prototypes are a step towards improving interpretability of complex data distributions. For future work, we hope to further explore the properties of `MMD-critic` such as the effect of the choice of kernel, and weaker conditions on the kernel matrix for submodularity. We plan to explore applications to larger datasets, aided by recent work on distributed algorithms for submodular optimization. We also intend to complete a larger scale user study on how criticism *and* prototypes presented together affect human understanding.

References

A. Aamodt and E. Plaza. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI communications*, 1994.

A. Badanidiyuru, B. Mirzasoleiman, A. Karbasi, and A. Krause. Streaming submodular maximization: Massive data summarization on the fly. In *KDD*. ACM, 2014.

I. Bichindaritz and C. Marling. Case-based reasoning in the health sciences: What's next? *AI in medicine*, 2006.

J. Bien and R. Tibshirani. Prototype selection for interpretable classification. *The Annals of Applied Statistics*, pages 2403–2424, 2011.

R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *KDD*, 2015.

M.S. Cohen, J.T. Freeman, and S. Wolf. Metarecognition in time-stressed decision making: Recognizing, critiquing, and correcting. *Human Factors*, 1996.

J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

U. Feige. A threshold of ln n for approximating set cover. *JACM*, 1998.

A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian data analysis*. Taylor & Francis, 2014.

A. Gretton, K.M. Borgwardt, M.J. Rasch, B. Schölkopf, and A. Smola. A kernel method for the two-sample problem. *JMLR*, 2008.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv:1512.03385*, 2015.

J.J. Hull. A database for handwritten text recognition research. *TPAMI*, 1994.

T.S. Jaakkola, D. Haussler, et al. Exploiting generative models in discriminative classifiers. In *NIPS*, pages 487–493, 1999.

L. Kaufman and P. Rousseeuw. *Clustering by means of medoids*. North-Holland, 1987.

B. Kim, C. Rudin, and J.A. Shah. The Bayesian Case Model: A generative approach for case-based reasoning and prototype classification. In *NIPS*, 2014.

O.O. Koyejo, R. Khanna, J. Ghosh, and R. Poldrack. On prior distributions and approximate inference for structured variables. In *NIPS*, 2014.

A. Krause, A. Singh, and C. Guestrin. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *JMLR*, 2008.

L. I. Kuncheva and J.C. Bezdek. Nearest prototype classification: clustering, genetic algorithms, or random search? *IEEE Transactions on Systems, Man, and Cybernetics*, 28(1):160–164, 1998.

H. Lin and J. Bilmes. A class of submodular functions for document summarization. In *ACL*, 2011.

J. R. Lloyd and Z. Ghahramani. Statistical model criticism using kernel two sample tests. In *NIPS*, 2015.

B. Mirzasoleiman, A. Karbasi, A. Badanidiyuru, and A. Krause. Distributed submodular cover: Succinctly summarizing massive data. In *NIPS*, 2015.

G. L Nemhauser, L.A. Wolsey, and M.L. Fisher. An analysis of approximations for maximizing submodular set functions. *Mathematical Programming*, 1978.

A. Newell and H.A. Simon. *Human problem solving*. Prentice-Hall Englewood Cliffs, 1972.

C.E. Priebe, D.J. Marchette, J.G. DeVinney, and D.A. Socolinsky. Classification using class cover catch digraphs. *Journal of classification*, 2003.

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015.

D. Sharma, A. Kapoor, and A. Deshpande. On greedy maximization of entropy. In *ICML*, 2015.

I. Simon, N. Snavely, and S.M. Seitz. Scene summarization for online image collections. In *ICCV*, 2007.

K.R. Varshney. Engineering safety in machine learning. *arXiv:1601.04126*, 2016.

Proof of Theorem 2

Observe that from the element-wise upper bound on $\mathbf{H}$, the following element-wise inequality holds $h_*\mathbf{E} \leq \mathbf{H} \leq h_*\mathbf{E} + \nu\mathbf{E}'$. Thus, from the linearity of $F(\mathbf{H}, \mathsf{S}) = \langle \mathbf{A}(\mathsf{S}), \mathbf{H} \rangle$ with respect to $\mathbf{H}$, we have that:

$$F(h_*\mathbf{E}, \mathsf{S}) \leq F(\mathbf{H}, \mathsf{S}) \leq F(h_*\mathbf{E} + \nu\mathbf{E}', \mathsf{S}),$$

where (by linearity) $F(h_*\mathbf{E} + \nu\mathbf{E}', \mathsf{S}) = h_*F(\mathbf{E}, \mathsf{S}) + \nu F(\mathbf{E}', \mathsf{S})$.

Next, employing terms: $a(\mathsf{S}) = F(\mathbf{E}, \mathsf{S}) = \langle H(\mathsf{S}), \mathbf{E} \rangle$ and $b(\mathsf{S}) = F(\mathbf{E}', \mathsf{S}) = \langle A(\mathsf{S}), \mathbf{E}' \rangle$. we may rewrite the bounds as:

$$h_*a(\mathsf{S}) \leq F(\mathbf{H}, \mathsf{S}) \leq h_*a(\mathsf{S}) + \nu b(\mathsf{S}).$$

Monotonicity:
The function $F(\mathbf{H}, \mathsf{S})$ is monotone with respect to S if: $F(\mathbf{H}, \mathsf{S} \cup \{u\}) - F(\mathbf{H}, \mathsf{S}) \geq 0$. Applying the lower and upper bounds, we have that:

$$\begin{aligned} &F(\mathbf{H}, \mathsf{S} \cup \{u\}) - F(\mathbf{H}, \mathsf{S}) \geq h_*a(\mathsf{S} \cup \{u\}) - h_*a(\mathsf{S}) - \nu b(\mathsf{S}) \geq 0 \\ &\implies \nu \leq \frac{h_*a(\mathsf{S} \cup \{u\}) - h_*a(\mathsf{S})}{b(\mathsf{S})} = h_*\alpha(n, m) \end{aligned}$$

Thus, when the off-diagonal terms satisfy $h_{i,j} \leq h_*\alpha(n, m)\ \forall\ 0 \leq m \leq m_*,\ \forall\ (i, j) \in \mathsf{E}'$, we have that $F(\mathbf{H}, \mathsf{S})$ is monotone.

Submodularity:
The function $F(\mathbf{H}, \mathsf{S})$ is submodular with respect to S if: $F(\mathbf{H}, \mathsf{S} \cup \{u\}) + F(\mathbf{H}, \mathsf{S} \cup \{v\}) \geq F(\mathbf{H}, \mathsf{S} \cup \{u, v\}) + F(\mathbf{H}, \mathsf{S})$. Again, applying the lower and upper bounds, we have that:

$$\begin{aligned} &F(\mathbf{H}, \mathsf{S} \cup \{u\}) + F(\mathbf{H}, \mathsf{S} \cup \{v\}) - F(\mathbf{H}, \mathsf{S} \cup \{u, v\}) - F(\mathbf{H}, \mathsf{S}) \\ &\geq h_*a(\mathsf{S} \cup \{u\}) + h_*a(\mathsf{S} \cup \{v\}) - h_*a(\mathsf{S} \cup \{u, v\}) - \nu b(\mathsf{S} \cup \{u, v\}) - h_*a(\mathsf{S}) - \nu b(\mathsf{S}) \geq 0 \\ &\implies \nu \leq h_* \frac{a(\mathsf{S} \cup \{u\}) + a(\mathsf{S} \cup \{v\}) - a(\mathsf{S} \cup \{u, v\}) - a(\mathsf{S})}{b(\mathsf{S} \cup \{u, v\}) + b(\mathsf{S})} = h_*\beta(n, m) \end{aligned}$$

Thus, when the off-diagonal terms satisfy $h_{i,j} \leq h_*\beta(n, m)\ \forall\ 0 \leq m \leq m_*,\ \forall\ (i, j) \in \mathsf{E}'$, we have that $F(\mathbf{H}, \mathsf{S})$ is submodular.

Proof of Corollary 3

Based on the diagonal dominance assumption on $\mathbf{K}$, it is clear that $\mathsf{E}' = \{i, j \in [n]\, i \neq j\}$ indexes the off diagonal terms, and $\mathbf{E} = 1 - \mathbf{E}' = \mathbf{I}$. Given $\mathbf{A}(\mathsf{S})$ with entries $a_{i,j}(\mathsf{S}) = \frac{2}{n|\mathsf{S}|}\mathbf{1}_{[j \in \mathsf{S}]} - \frac{1}{|\mathsf{S}|^2}\mathbf{1}_{[i \in \mathsf{S}]}\mathbf{1}_{[j \in \mathsf{S}]}$, we can compute the bounds (8) simply by enumerating sums as:

$$\begin{aligned} a(\mathsf{S}) &= \langle A(\mathsf{S}), \mathbf{I} \rangle = \frac{2m}{nm} - \frac{m}{m^2} = \frac{2}{n} - \frac{1}{m} \\ b(\mathsf{S}) &= \langle A(\mathsf{S}), 1 - \mathbf{I} \rangle = \frac{2(nm - m)}{nm} - \frac{m^2 - m}{m^2} = \frac{2(n-1)}{n} - \frac{m-1}{m} \end{aligned}$$

Monotonicity: $J_p(\cdot)$ is monotone when the upper bound of the off-diagonal terms is given by $\alpha(n, m) = \frac{a(\mathsf{S} \cup \{u\}) - a(\mathsf{S})}{b(\mathsf{S})}$ by Theorem 2. We have that:

$$a(\mathsf{S} \cup \{u\}) - a(\mathsf{S}) = \frac{-1}{m+1} + \frac{1}{m}, \quad b(\mathsf{S}) = \frac{2(n-1)}{n} - \frac{m-1}{m}.$$

Thus:

$$\alpha(n, m) = \frac{n}{(m+1)(m(n-2)+n)}.$$

This is a decreasing function wrt m. Further, for the ground set $2^{[n]}$, we have that $m_* = n$, and $\alpha(n, n) = \frac{1}{n^2-1}$

Submodularity: $J_p(\cdot)$ is submodular when the upper bound of the off-diagonal terms is given by $\beta(n, m) = \frac{a(\mathsf{S} \cup \{u\}) + a(\mathsf{S} \cup v\}) - a(\mathsf{S} \cup \{u, v\}) - a(\mathsf{S})}{b(\mathsf{S} \cup \{u, v\}) + b(\mathsf{S})}$ by Theorem 2. We have that:

$$\begin{aligned} a(\mathsf{S} \cup \{u\}) + a(\mathsf{S} \cup v\}) - a(\mathsf{S} \cup \{u, v\}) - a(\mathsf{S}) &= \frac{-2}{m+1} + \frac{1}{m+1} + + \frac{1}{m} \\ b(\mathsf{S} \cup \{u, v\}) + b(\mathsf{S}) &= \frac{4(n-1)}{n} - \frac{m+1}{m+2} - \frac{m-1}{m} \end{aligned}$$

Thus:

$$\beta(n,m) = \frac{n}{(m+1)(n(m^2+3m+1)-2(m^2+2m))}$$

This is a decreasing function wrt m. Further, for the ground set $2^{[n]}$, we have that $m_* = n$, and $\beta(n,n) = \frac{1}{n^3+2n^2-2n-3}$.

Combined Bound: Finally, we show that $\beta(n,n) \leq \alpha(n,n)$, so that the bound $k_{i,j} \leq k_*\beta(n,n)$ is sufficient to guarantee both monotonicity and submodularity.

$$\begin{aligned}
&\beta(n,n) \leq \alpha(n,n) \\
\implies &\frac{1}{n^3+2n^2-2n-3} \leq \frac{1}{n^2-1} \\
\implies &n^2-1 \leq n^3+2n^2-2n-3 \\
\implies &0 \leq n^3+n^2-n-3 \\
\implies &0 \leq (n-1)(n^2-2)
\end{aligned}$$

which holds when $n > -1$ and $n \geq \sqrt{2}$. Thus $\beta(n,n) \leq \alpha(n,n)$. The proof is complete.

Proof of Theorem 7

A discrete function is linear if it can be written in the form $F(\mathsf{C}) = \sum_{i\in[n]} w_i \mathbf{1}_{[i\in\mathsf{C}]}$. Consider (9) and observe that:

$$\begin{aligned}
L(\mathsf{C}) &= \sum_{l\in\mathsf{C}} \left| \frac{1}{n}\sum_{i\in[n]} k(x_i,x_l) - \frac{1}{m}\sum_{j\in\mathsf{S}} k(x_j,x_l) \right| \\
&= \sum_{l\in[n]} \left(\left| \frac{1}{n}\sum_{i\in[n]} k(x_i,x_l) - \frac{1}{m}\sum_{j\in\mathsf{S}} k(x_j,x_l) \right| \right) \mathbf{1}_{[l\in\mathsf{C}]} \\
&= \sum_{l\in[n]} w_l \mathbf{1}_{[l\in\mathsf{C}]},
\end{aligned}$$

where:

$$w_l = \left| \frac{1}{n}\sum_{i\in[n]} k(x_i,x_l) - \frac{1}{m}\sum_{j\in\mathsf{S}} k(x_j,x_l) \right|.$$