Supplementary Information

1. Proofs

Before we bound $\hat{Cov}$ and $\hat{Pre}$, we would like to make the following observation for the function $\Gamma(x)$ which is defined as $\Gamma(x) = c_1\sigma(c_2x) + c_3(sgn(x)c_4 + c_5)$.

Observation 1. *When $c_4 = c_5 = 0.5$ and $c_3 = 1 - c_1$, we have:*

- *If $x > 0$, $\Gamma(x) = c_1\sigma(c_2x) + c_3$*
- *If $x < 0$, $\Gamma(x) = c_1\sigma(c_2x)$*
- *If $x = 0$, $\Gamma(x) = 0.5$*

The following Lemma bounds the term $h(l, u, x)$ for any point x and shows that it is a good enough approximation for the indicator function.

Lemma 1.1. *Let $c = \frac{c_1}{2}$ and $c_h > 1 - c$. If $c < \frac{1}{4D}$, we have $\forall x_i$:*

- *If $l_j < x_{ij} \leq u_j \ \forall j = \{1, 2, \ldots, D\}$, we have:*

$$h(l, u, x) \leq 1 \text{ and}$$
$$h(l, u, x) \geq 1 - c$$

i.e. for all points lying inside the hypercuboid, function $h(\cdot)$ is very close to 1.

- *If $\exists k, m$ with $k + m \geq 1$, such that $x_{ij} \leq l_j$ for k attributes or $x_{ij} > u_j$ for m attributes then:*

$$h(l, u, x) \leq c \text{ and}$$
$$h(l, u, x) \geq 0$$

i.e. for all points lying outside the hypercuboid, function $h(\cdot)$ is very close to 0.

Proof. The proof considers four cases depending on the number of attributes of a data point that lie between the lower bound and upper bound of the hyperrectangle.

case 1: $\forall j \in \{1, 2, \ldots, D\}, l_j < x_{ij} \leq u_j$.

$$h(l, u, x_i)$$
$$= \Gamma\left(\frac{\sum_{j=1}^{D}\Gamma\left(x_{ij} - l_j\right) + \sum_{j=1}^{D}\Gamma\left(u_j - x_{ij} + c_l\right)}{2D} - c_h\right)$$
$$= \Gamma\left(\frac{\sum_{j=1}^{D}(c_1\sigma\left(c_2\left(x_{ij} - l_j\right)\right) + c_3)}{2D}\right.$$
$$\left. + \frac{\sum_{j=1}^{D}(c_1\sigma\left(c_2\left(u_j - x_{ij} + c_l\right)\right) + c_3)}{2D} - c_h\right)$$

(From Observation 1)

Let, $t = c_1\frac{\sum_{j=1}^{D}\sigma(c_2(x_{ij}-l_j)) + \sum_{j=1}^{D}\sigma(c_2(u_j - x_{ij}+c_l))}{2D} + c_3 - c_h$, then using the fact that $\sigma(x) \geq 0.5$ if $x > 0$, we have:

$$\begin{aligned} t &\geq \frac{c_1}{2} + c_3 - c_h \\ &\geq 1 - \frac{c_1}{2} - c_h \\ &> 0 \qquad \text{(if } c_h + \tfrac{c_1}{2} < 1\text{)} \end{aligned}$$

Thus, if $c_h + \frac{c_1}{2} < 1$, we have $t > 0$. Thus, we get, $h(l, u, x_i) = \Gamma(t) = c_1\sigma(c_2t) + c_3$ from Observation 1. Since, $t > 0$, $c_2t > 0$ for any $c_2 > 0$, we have, $h(l, u, x_i) \geq \frac{c_1}{2} + c_3 \geq 1 - \frac{c_1}{2}$. Also, $h(l, u, x_i) = c_1\sigma(c_2t) + c_3 \leq c_1 + c_3 \leq 1$

Case 2: Let us assume that $\exists k$ such that $x_{ij} \leq l_j$ for k attributes i.e. point lie outside or on the lower bound of hypercuboid for $k \geq 1$ attributes and $\exists m$ such that $x_{ij} > u_j$ for $m \geq 1$ attributes. Out of k attributes, let k_1 attributes have $x_{ij} = l_j$ and $k - k_1$ attributes $x_{ij} < l_j$. Then, we have:

- For all k_1 attributes: $\Gamma(x_{ij} - l_j) = 0.5$
- For $k - k_1$ attributes:$\Gamma(x_{ij} - l_j) = c_1\sigma(c_2(x_{ij} - l_j)) \leq c_1$
- For $D - k$ attributes: $\Gamma(x_{ij} - l_j) = c_1\sigma(c_2(x_{ij} - l_j)) + c_3 \leq 1$
- For all m attributes: $\Gamma(u_j - x_{ij} + c_l) = c_1\sigma(c_2(u_j - x_{ij} + c_l)) \leq c_1$
- For $D - m$ attributes: $\Gamma(u_j - x_{ij} + c_l) = c_1\sigma(c_2(u_j - x_{ij} + c_l)) + c_3 \leq 1$

We get,

$$\begin{aligned}
&h(l,u,x_i)\\
&=\Gamma\left(\frac{0.5k_1+\sum_{j=1}^{k-k_1}c_1\sigma(c_2(x_{ij}-l_j))}{2D}\right.\\
&+\frac{\sum_{j=k+1}^{D}((c_1\sigma(c_2(x_{ij}-l_j))+c_3)}{2D}\\
&+\frac{\sum_{j=1}^{m}c_1\sigma(c_2(u_j-x_{ij}+c_l)}{2D}\\
&\left.+\frac{\sum_{j=m+1}^{D}c_1\sigma(c_2(u_j-x_{ij}+c_l)+c_3}{2D}-c_h\right)
\end{aligned}$$

Let, $h(l,u,x_i)=\Gamma(t)$ i.e. consider the entire term in Γ expression to be t then:

$$\begin{aligned}
t\leq&\frac{0.5k_1+(k-k_1)0.5c_1+(D-k)(c_1+c_3)}{2D}\\
&+\frac{0.5mc_1+(D-m)(c_1+c_3)}{2D}-c_h\\
t\leq&\frac{0.5k_1(1-c_1)+0.5kc_1+0.5mc_1+2D-k-m}{2D}-c_h\\
&(1\leq k+m\leq 2D,\ k_1\leq D,\text{ and } 0<c_1<1)\\
\leq&\frac{0.5D(1-c_1)+0.5c_1D}{2D}+\frac{2D-1}{2D}-c_h\\
\leq&\frac{1}{4D}+\frac{2D-1}{2D}-c_h\\
\leq&\frac{4D-1}{4D}-c_h
\end{aligned}$$

Thus, when $c_h>\frac{4D-1}{4D}$, then we get $t<0$. In this case, we have: $h(l,u,x_i)=\Gamma(t)=c_1\sigma(c_2t)\leq\frac{c_1}{2}$. From Case 1, we have $\frac{c_1}{2}<1-c_h$. Substituting $c_h>\frac{4D-1}{4D}$, we get, $\frac{c_1}{2}<\frac{1}{4D}$. Thus, if any of the attribute of the example lies outside the boundary, we get $h(l,u,x_i)\leq\frac{1}{4D}$ and if all the attributes lie inside the boundary, we get $h(l,u,x_i)\geq\frac{4D-1}{4D}$ □

We now prove the main theorem which bounds our approximate coverage in terms of the true coverage.

Theorem 1.2. $\left(\frac{4D-1}{4D}\right)Cov\leq\hat{Cov}\leq\frac{1}{4D}+\left(\frac{4D-1}{4D}\right)Cov$

Proof. Let the actual coverage from the hypercuboid (l,u) be $\frac{k}{N}$ i.e. $\sum_{i=1}^{N}\mathbb{I}(x_i\in S(l,u))=k$. Then:

$$\begin{aligned}
\hat{Cov}&=\frac{1}{N}\sum_{i=1}^{N}h(l,u,x_i)\\
&=\frac{1}{N}\sum_{x_i\in S(l,u)}h(l,u,x_i)+\frac{1}{N}\sum_{x_i\notin S(l,u)}h(l,u,x_i)\\
&\geq\frac{1}{N}k(1-c) && \text{(From Lemma 1.1)}\\
&\geq Cov\left(\frac{4D-1}{4D}\right) && (c<\tfrac{1}{4D})
\end{aligned}$$

Also,

$$\begin{aligned}
\hat{Cov}&=\frac{1}{N}\sum_{x_i\in S(l,u)}h(l,u,x_i)+\frac{1}{N}\sum_{x_i\notin S(l,u)}h(l,u,x_i)\\
&\leq\frac{k}{N}+\frac{N-k}{N}c && \text{From Lemma 1.1}\\
&\leq c+Cov\,(1-c)\\
&\leq\frac{1}{4D}+Cov\left(\frac{4D-1}{4D}\right) && (c<\tfrac{1}{4D})
\end{aligned}$$

□

The above result is interesting not only because it bounds the approximate coverage in terms of true coverage but it also suggests that as the features (dimension) increases, approximate coverage becomes closer to the true coverage. We also verify this from our experiments in Table 1

We also have additional result for the bounds on the approximate precision.

Theorem 1.3. $\hat{Pre}\leq Pre\left(1+\frac{1}{Cov}\left(\frac{4D}{4D-1}\right)\right)$. *Thus, When algorithm returns a hypercuboid with* $\hat{Pre}\geq P$ *then* $Pre\geq\frac{1}{\left(1+\frac{1}{Cov}\left(\frac{4D}{4D-1}\right)\right)}P$

Proof. Let, k points be inside the hyper-cuboid, out of k points, q points satisfy $f(x_i)=f(x_q)$ and m points satisfy $f(x_i)=f(x_q)$ in total.

$$\begin{aligned}
\hat{Pre}&=\frac{\sum_{x_i=x_q\atop x_i\in S(l,u)}h(l,u,x_i)+\sum_{x_i=x_q\atop x_i\notin S(l,u)}h(l,u,x_i)}{\sum_{x_i\in S(l,u)}h(l,u,x_i)+\sum_{x_i\notin S(l,u)}h(l,u,x_i)}\\
&\leq\frac{q+(m-q)c}{(1-c)k}\\
&\leq Pre+\frac{m}{k}\left(\frac{c}{1-c}\right)\\
&\leq Pre+\frac{qN}{k^2}\left(\frac{4D}{4D-1}\right) \quad \left(\tfrac{M}{N}\leq\tfrac{q}{k}\text{ and }c<\tfrac{1}{4D}\right)\\
&\leq Pre+\frac{Pre}{Cov}\left(\frac{4D}{4D-1}\right)
\end{aligned}$$

□

2. Methodology

2.1. Extension to Discrete Attributes

The MAIRE framework is directly applicable on ordered discrete attributes. The final explanation is a set of consecutive discrete values. The generated explanation is slightly modified for ordered discrete attributes by changing l_i to the smallest discrete value that is greater than or equal to l_i and changing u_i to the largest discrete value that is lesser

than or equal to u_i. This modification does not affect coverage or precision and improves readability. In the case of a categorical attribute (unordered), finding intervals is not meaningful. We instead convert all categorical attributes to their equivalent one-hot encoding. The transformed boolean representation is treated as ordered discrete attributes. If an explanation contains both the values of a boolean attribute, the corresponding attribute is dropped from the explanation. If only the value one is selected, then the value of the unordered attribute in $\mathbf{x}'_q$ is included in the explanation. Due to the enforcement of the second constraint, selection of only 0 is not possible as $\mathbf{x}'_q$ has the value 1 for the corresponding boolean attribute.

3. Additional Experiments and Results

Code for the experiments mentioned is available at https://github.com/anonymousID2242/code-submission.

3.1. Synthetic datasets

The MAIRE framework is tested on several 2D synthetic datasets. The instances for all these datasets are sampled from the interval $[0, 1]$. For these datasets, a simple shape was chosen for positive class ($f(\mathbf{x}) = 1$) region. Everywhere else, $f(\mathbf{x})$ is 0. Using simple shapes allows for easy visualization of the explanations generated by the model. Some results included in the main paper are repeated here to ease the understanding of the results.

Figure 1 illustrates the explanations generated by the MAIRE framework on various synthetic datasets. In all of these figures, the blue regions represent $f(\mathbf{x}) = 1$, the red rectangle marks the final explanation generated by the framework and the black point refers to the query point, i.e., $\mathbf{x}'_q$. The lighter colors are used for marking regions that are not included in the explanations. In Figures 1 (a)-(f), the non-blue regions represent $f(\mathbf{x}) = 0$. In Figures 1 (g) and (h), the red regions represent $f(\mathbf{x}) = 0$ and non-blue regions are not included in the instance space, i.e. the instance space is discretized. The attribute along axis 1 is discretized to take 5 values - $\{\frac{1}{6}, \frac{2}{6}, \frac{3}{6}, \frac{4}{6}, \frac{5}{6}\}$.

Figures 1(a) and (b) represents the MAIRE explanations on a rectangular decision boundary with different query points. We observe that in Figure 1(a), because $\mathbf{x}'_q$ belongs to $f(\mathbf{x}) = 1$ region, the explanation completely covers the rectangle as this is the largest region that includes $\mathbf{x}'_q$ and has a high precision. Similarly, in Figure 1(b), $\mathbf{x}'_q$ belongs to $f(\mathbf{x}) = 0$ region. So, the explanation generated is correspondingly the largest rectangle that includes $\mathbf{x}'_q$ such that most of the region has $f(\mathbf{x}) = 0$, thus having a high precision. An interesting observation here is that the framework has two potential options - horizontally cover the whole range or vertically cover the whole range. We want to point out that for the MAIRE framework, these correspond to two local maxima. Vertical cover has a larger area and so, is the global maxima. We observe that the MAIRE explanation corresponds to the vertical cover. However, it could have very well chosen the other local maxima, i.e., the horizontal cover instead. This is an artifact of any gradient-based optimization routine.

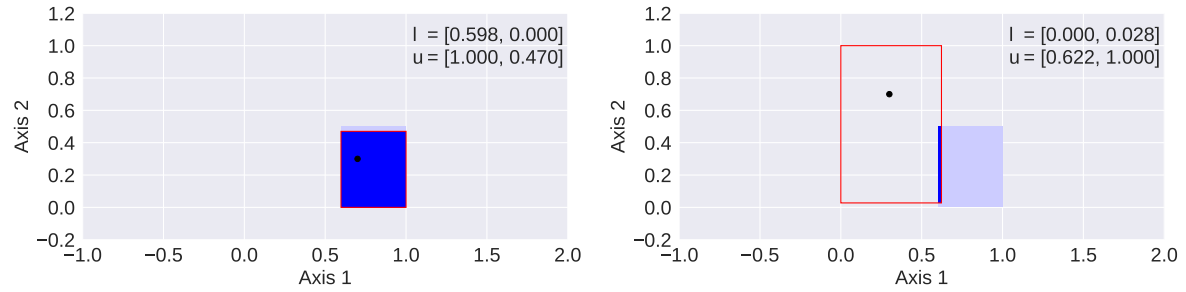
Figures 1(c) and (d) represent the MAIRE explanations on a circular decision boundary with different values of the precision threshold P. In Figure 1(c), as P was 0.80, we observe that the final explanation almost completely circumscribes the circle. On the other hand, in Figure 1(d), as P was 0.95, the explanation generated is smaller in size as this size has lesser percentage of points with $f(\mathbf{x}) = 0$.

Figures 1(e) and (f) compare the explanations generated by the MAIRE framework when the second constraint (i.e., the explanation must contain $\mathbf{x}'_q$) is active and inactive. In this set of experiments, there are two $f(\mathbf{x}) = 1$ regions. One is marked in blue (the blue rectangle). Other than that, $f(\mathbf{x}'_q)$ is also 1. In the Figure 1(e), the constraint was inactive (with $\lambda_2 = 0$). We observe that the final explanation does not contain $\mathbf{x}'_q$. This is simply because the framework maximizes the precision by minimizing the thin $f(\mathbf{x}) = 0$ strip. In the Figure 1(f), the constraint was active (with $\lambda_2 = 5$). We observe that the final explanation contains $\mathbf{x}'_q$. Here, the entire vertical range was not covered because that would have led to a precision lower than the threshold P.

Figures 1(g) and (h) represents the MAIRE explanations for a synthetic dataset where one attribute has an ordered discrete domain and the other has a continuous domain. In the Figure 1(g), $f(\mathbf{x}'_q) = 1$ and so, the corresponding blue regions from the two adjacent strips were selected. While in the Figure 1(h), $f(\mathbf{x}'_q) = 0$ and so, the corresponding red strip was selected as the explanation.

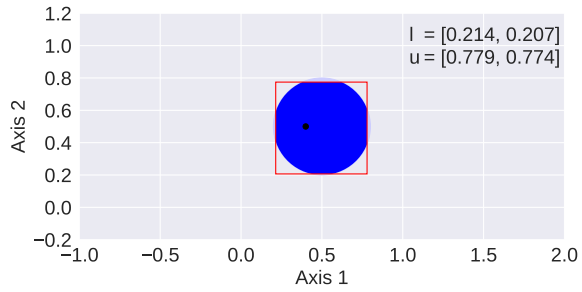
3.2. Tabular datasets

We conducted experiments to study the quality of the approximations to coverage and precision using the tabular datasets. Explanations for 100 randomly sampled data points for each of the datasets were computed. The true coverage and precision were determined for each explanation as well as the values for the corresponding approximations. The mean squared error between the true and approximate values averaged over 100 data points for the three datasets is presented in Table 1. It can be noticed that difference in the true values and the corresponding approximations is not significant. Further this difference reduces as the number of attributes increases supporting our theoretical analysis. The German credit dataset has the highest number of attributes (20), followed by Adult (14), and Abalone (8) data sets.

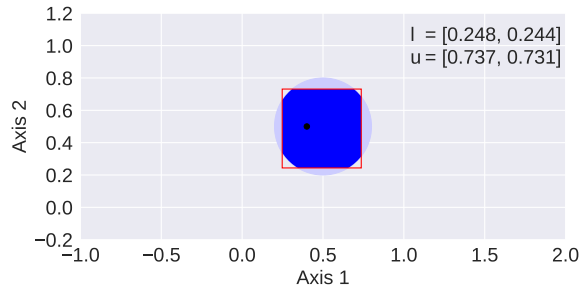


(a) $f(\mathbf{x}'_q) = 1$

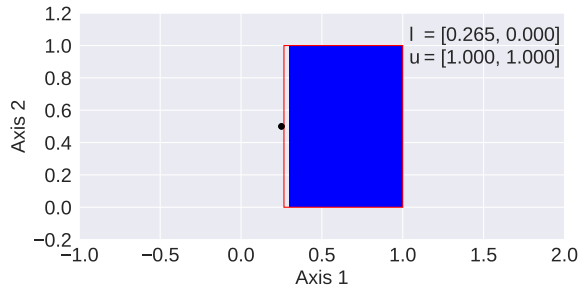
(b) $f(\mathbf{x}'_q) = 0$



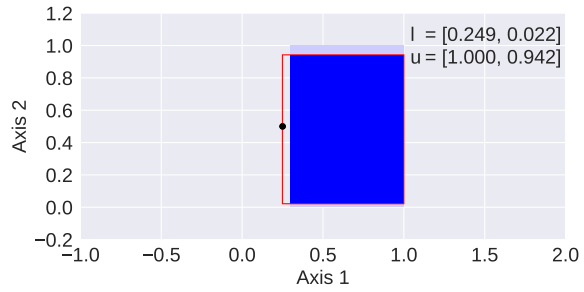
(c) P = 0.80



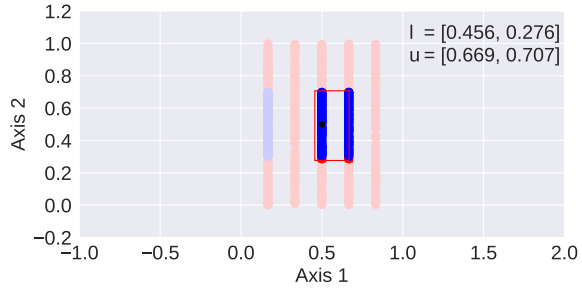
(d) P = 0.95



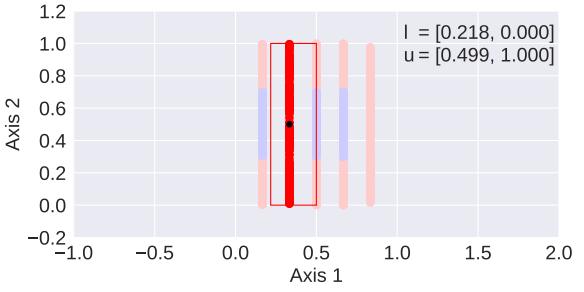
(e) Second Constraint Off ($\lambda_2 = 0$)



(f) Second Constraint On ($\lambda_2 = 5$)



(g) $f(\mathbf{x}'_q) = 1$



(h) $f(\mathbf{x}'_q) = 0$

Figure 1. [Best viewed in color]MAIRE Explanations for Synthetic Datasets (a) and (b) for Rectangular Decision Boundaries, (c) and (d) for Circular Decision Boundaries, (e) and (f) Effect of the Second Constraint (with $\mathbf{x}'_q = [0.250, 0.500]$), (g) and (h) for Synthetic Datasets with Discrete Attributes.

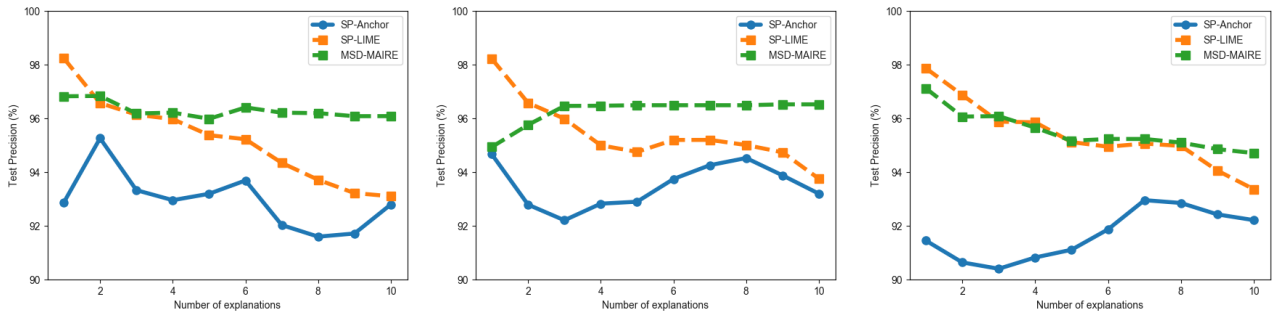


Figure 2. [Best viewed in color] Change in test precision as a function of number of local explanations included in the global explanation Test Coverage for (a) Adult (b) Abalone (c) German-Credit datasets for SP-LIME, SP-Anchors and MSD-MAIRE.

	Adult	Abalone	German credit
MSD Coverage	0.0015	0.0004	8.552e-05
MSD Precision	0.3217	0.1265	0.0985

Table 1. Mean Square difference between Cov and $\hat{Cov}$, Pre and $\hat{Pre}$ for adult, abalone and German credit datasets averaged over 100 data points.

Figures 2 (a-c) compare the change in precision as the local explanations are incrementally added to the global explanation for the three tabular datasets. It is observed that the proposed framework results in a minimal reduction in precision consistently across the three datasets. The observation is in line with the mechanism the MAIRE framework employs to create a global explanation ensuring a minimum reduction in precision. LIME shows the maximum decrease in precision.

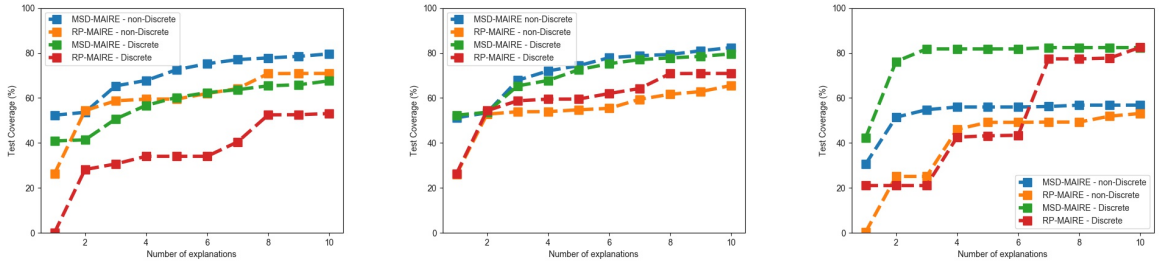
Figures 3(a-c) compares the performance of the MAIRE framework for both the discretized and non-discretized versions of the tabular datasets. We observe that the coverage of the global explanation for MSD-MAIRE for both versions of the datasets is comparable for Adult and Abalone datasets. However, we notice a significant improvement in the performance of MAIRE on the discretized version of the German-Credit dataset. Further investigation is required to understand this anomaly.

3.3. Text datasets

The datasets are divided into train and test splits in the ratio of 4:1. A bag of words representation was used to characterize the reviews and documents. In the case of IMDB movie reviews, we considered a random forest with 500 trees as our black-box model to be explained. The test accuracy of the above model is 87.3%. In the case of 20-Newsgroup dataset, we have only considered output labels 'medicine,' 'graphics,' 'Christian,' and 'atheism' as it is not feasible to present 20 labels to a human subject. We use a four-layer network consisting of two hidden layers with 512 nodes each, having ReLU activation, and dropout probability set to 0.3 among layers, and softmax activation at the output layer as the base classifier. The model is trained for 30 epochs using Adam optimizer. The test accuracy of the above black-box model is 81.17%. We use ten data points (three medicine, three atheism, two graphics, two Christian) and generated explanations for each review using 5 different approaches mentioned in the main paper. For generating the MAIRE explanation, the review was converted into a bag of words vector, and the sample points for computing $Cov, \hat{Cov}, Pre,$ and $\hat{Pre}$ were taken by randomly flipping bits in the bag of words. The words are ranked based on the effect they have on the classification using Greedy Attribute Elimination. Table 2 presents explanations generated by various explanatory models for both correct and incorrect classifications by the base classifier.

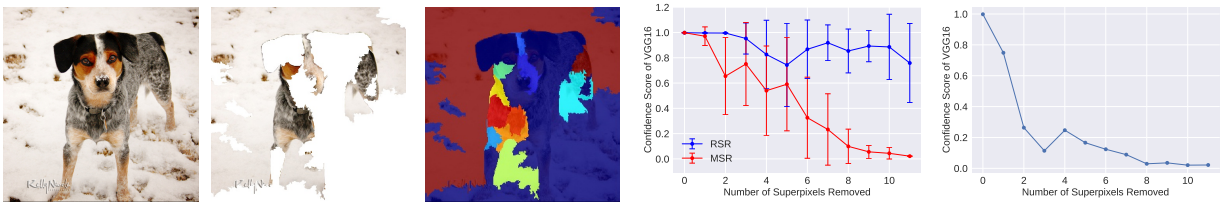
3.4. Image datasets

We use the MAIRE framework to explain the classification results of the VGG16 model for images. The procedure described in the main paper is used for generating the explanations for the test bluetick image. The Figure 4 shows the explanation generated by the MAIRE framework and heat map of the explanation (generated by ordering the superpixels chosen in the local explanation using Greedy Attribute Elimination) for the bluetick image. The VGG model has high confidence in its prediction for this image as well. The decrease in the classifier confidence (Figure 4(d)) with the removal of superpixels picked by the MAIRE framework is more significant than randomly selecting a superpixel. This also illustrates that the MAIRE framework does indeed select the superpixels having a significant impact on the classifier. We also observe that by removing the top 2 superpixels selected by the MAIRE framework, the classifier confidence drops to less than 0.5 for the bluetick image. It is interesting



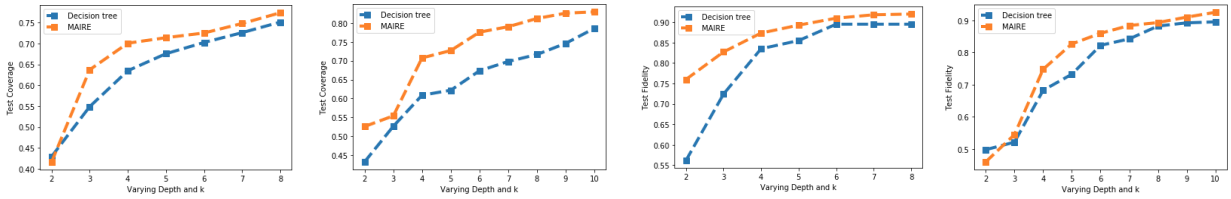
(a) (b) (c)

Figure 3. [Best viewed in color]Change in test coverage as a function of number of local explanations included in the global explanation Test Coverage for (a) Adult (b) Abalone (c) German-Credit Data sets comparing of discretized vs non-discretized version of the datasets for RP-MAIRE and MSD-MAIRE



(a) (b) (c) (d) (e)

Figure 4. [Best viewed in color] Results on the bluetick image (a) Original Image (b) Explanation Generated (c)Heat Map (d) Confidence Score as more number of Superpixels are Removed (RSR = Random Superpixels Removed, MSR = MAIRE Superpixels Removed) (e) Confidence Score as more number of Superpixels are Removed (the removal order is from most important to least as given by Greedy Attribute Elimination)



(a) (b) (c) (d)

Figure 5. [Best viewed in color] Test Coverage for MAIRE and Decision tree (a) Abalone (b) German-Credit, Test Precision for MAIRE and Decision tree (c) Abalone (d) German-Credit datasets

to note that the images in Figures 4(b) and 4(c) show that the MAIRE framework selected superpixels mostly from the background in the bluetick image. Surprisingly, the VGG16 model classified the image, containing only the superpixels selected by the MAIRE framework for the bluetick image, correctly with the confidence of 0.953. Further, when we remove the superpixel containing the background snow, the VGG16 classifier confidence drops to 0.007. This indicates that the VGG16 network is focusing on perhaps incorrect regions of the image. The MAIRE framework is effective at detecting such wrong correlations learned by the machine learning model.

3.5. Comparison against decision trees

Results of the additional experiments comparing coverage and precision of the explanations extracted from decision trees and the MAIRE framework for varying length of the explanation is presented Figure 5. These results were obtained on the Abalone and German credit datasets. Similar to the observation on the Adult data set, MAIRE has higher precision than decision trees for all values of K. It is also observed that for small K, MAIRE has higher coverage than a decision tree. This indicates that MAIRE is able to generate better explanations in terms of both precision and interpretability than a decision tree.

Review/Document	LIME	SHAP	Anchors	L2X	MAIRE
model prediction : negative, True label : negative Encouraged by the positive comments about this film on here I was looking forward to watching this film. Bad mistake. I've seen 950+ films and this is truly one of the worst of them - it's awful in almost every way: editing, pacing, storyline, 'acting,' soundtrack (the film's only song - a lame country tune - is played no less than four times). The film looks cheap and nasty and is boring in the extreme. Rarely have I been so happy to see the end credits of a film. The only thing that prevents me giving this a 1-score is Harvey Keitel - while this is far from his best performance he at least seems to be making a bit of an effort. One for Keitel obsessives only.	worst, Bad, awful, lame, boring, best, cheap, acting, thing, effort	mistake, best, lame, pacing, extreme, credits, obsessives, far, cheap, happy	bad, storyline, nasty, boring	credits, worst, comments, awful, cheap, nasty, mistake, extreme, lame, effort	Worst, awful, lame, boring, mistake, less, bad, cheap, obsessives, extreme
model prediction : graphics, True label : graphics I am looking for EISA or VESA local bus graphic cards that support at least —1024x786x24 resolution. I know Matrox has one, but it is very —expensive. All the other cards I know of, that support that —resoultion, are striaght ISA. What about the ELSA WINNER4000 (S3 928, Bt485, 4MB, EISA), or the Metheus Premier-4VL (S3 928, Bt485, 4MB, ISA/VL) ? —Also are there any X servers for a unix PC that support 24 bits? As it just happens, SGCS has a Xserver (X386 1.4) that does 1024x768x24 on those cards. Please email to info@sgcs.com for more details. - Thomas	VESA, PC, looking, 24, unix, email, resolution, graphic, info, support	cards, resoultion, support, unix, bus, details, com, bits, Metheus, servers	expensive, cards, support	support, details, expensive, cards, bits, servers, info, Premier, resolution, bus	Bus, Premier, graphic, Metheus, support, unix, expensive, details, ELSA, com
model prediction : negative, True label : positive this movie gets a 10 because there is a lot of gore in it.who cares about the plot or the acting.this is an Italian horror movie people so you know you can't expect much from the acting or the plot.everybody knows fulci took footage from other movies and added it to this one.since i never seen any of the movies that he took footage from it didn't matter to me.the Italian godfather of gore out done himself with this movie.this is one of the goriest Italian movies you will ever see.no gore hound should be without this movie in their horror movie collection.buy this movie no matter what it is a horehound's dream come true.	Plot, acting, didn, true, horror, collection, dream, footage, gets, movie	horror, cares, goriest, footage, never, expect, hound, fulci, matter, plot	horror, plot, hounds, matter	True, cares, collection, dream, never, acting, gore, fulci, matter, expect	Matter, acting, horror, dream, plot, movie, footage, collection, matter, cares
model prediction : christian, True label : atheism Pardon me if this is the wrong newsgroup. I would describe myself as an agnostic, in so far as I'm sure there is no single, universal supreme being, but if there is one and it is just, we will surely be judged on whether we lived good lives, striving to achieve that goodness that is within the power of each of us. Now, the complication is that one of my best friends has become very fundamentalist. That would normally be a non-issue with me, but he feels it is his responsibility to proselytize me (which I guess it is, according to his faith). This is a great strain to our friendship. I would have no problem if the subject didn't come up, but when it does, the discussion quickly begins to offend both of us: he is offended because I call into question his bedrock beliefs; I am offended by what I feel is a subscription to superstition, rationalized by such circular arguments as 'the Bible is God's word because He tells us in the Bible that it is so.' So my question is, how can I convince him that this is a subject better left undiscussed, so we can preserve what is (in all areas other than religious beliefs) a great friendship? How do I convince him that I am 'beyond saving' so he won't try? Thanks for any advice.	Bible, faith, beliefs, just, religious, good, word, feel, lives, Thanks	universal, lives, faith, power, religious, offend, superstition, convince, beliefs, complication	Superstition, Bible, faith	Religious, responsibility, complication, universal, subject, strain, power, circular, Subscription, convince	Strain, faith, God, superstition, saving, great, good, beliefs, religious, Bible

Table 2. Sample Explanations for correctly (first and second row) and incorrectly (third and fourth row) classified documents in the IMDB and Newsgroup Dataset.