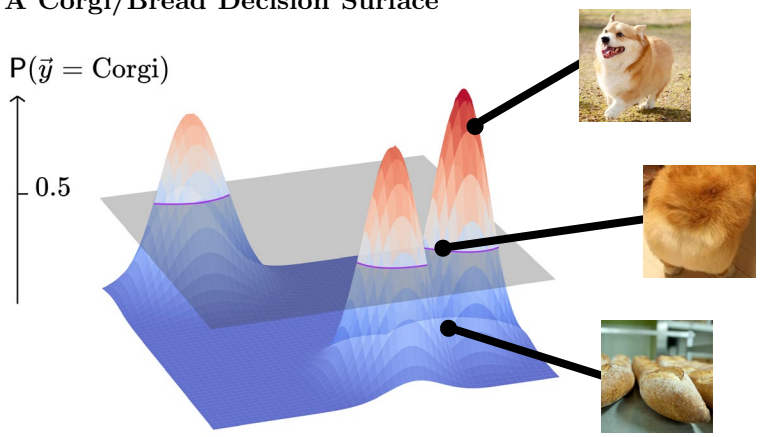
#8526: Bayes-TrEx: A Bayesian Sampling Approach To Transparency by Example

Serena Booth*, Yilun Zhou*, Ankit Shah, Julie Shah
{serenabooth, yilun, ajshah, julie_a_shah}@csail.mit.edu

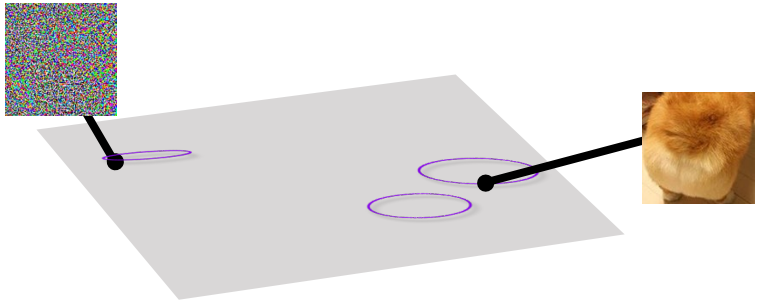
Transparency by Example

Our objective: Build a *holistic* understanding of a classifier.

A Corgi/Bread Decision Surface



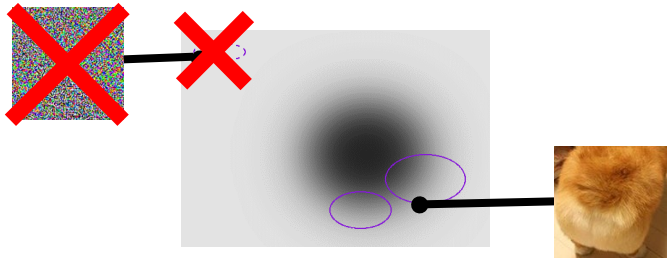
We search for **level set** examples which elicit a target prediction to help us gain insight into the classifier.



This slice corresponds to the P(Corgi) = 0.5 level set.

Method

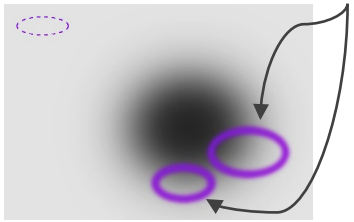
We want to find an example $\mathbf{x}$ which is *natural* and plausible under the data, and for which the classifier $f(\mathbf{x})$ has confidence $\mathbf{p}$



Want to sample from: $p(\mathbf{x}|f(\mathbf{x}) = \mathbf{p}) \propto p(\mathbf{x})\, p(f(\mathbf{x}) = \mathbf{p}|\mathbf{x})$

Two problems in applying MCMC methods:

1. $\{\mathbf{x} : f(\mathbf{x}) = \mathbf{p}\}$ has small or even zero measure.
2. $\mathbf{x}$ too high-dimensional.



To solve **problem 1**, we relax the formulation by widening the level set.

Introduce a random vector:

$$\mathbf{u}|\mathbf{x} \sim \mathcal{N}\left(f(\mathbf{x}), \sigma^2\right)$$

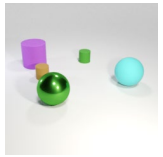
And sample from the new posterior:

$$p(\mathbf{x}|\mathbf{u} = \mathbf{u}^*) \propto p(\mathbf{x})p(\mathbf{u} = \mathbf{u}^*|\mathbf{x})$$

$$\mathbf{u}^* = \mathbf{p}$$

To solve **problem 2**, we use a generative model to represent $\mathbf{x}$ and sample from its parameter space, instead.

Applications, Evaluation, and Results



97.2%

96.0%

Classifier: "Contains 1 Cube"

Bayes-TrEx lets us find **high confidence failures**, which are more likely to be missed in assessing models.

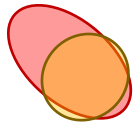
90.4%

90.5%

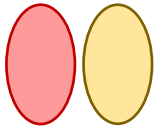
Classifier: "Contains 5 Cubes"

Bayes-TrEx lets us assess model responses to **novel classes**, like these Corgis.

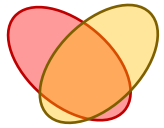
$\mathbb{P}_D \approx \mathbb{P}_C$



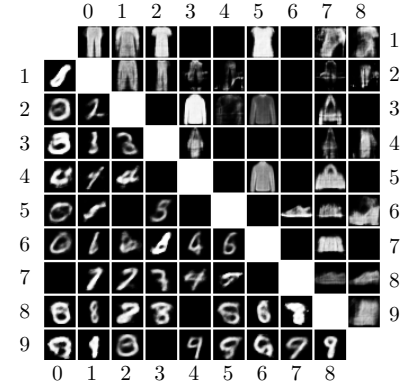
$\mathbb{P}_D \subsetneq \mathbb{P}_C$



$\mathbb{P}_D \cap \mathbb{P}_C = \varnothing$



$\mathbb{P}_D \cap \mathbb{P}_C \neq \mathbb{P}_D \neq \mathbb{P}_C$



Bayes-TrEx lets us assess **class boundaries** by finding examples where the classifier has 50/50 confidence between two classes.