10 minutes

4 slides

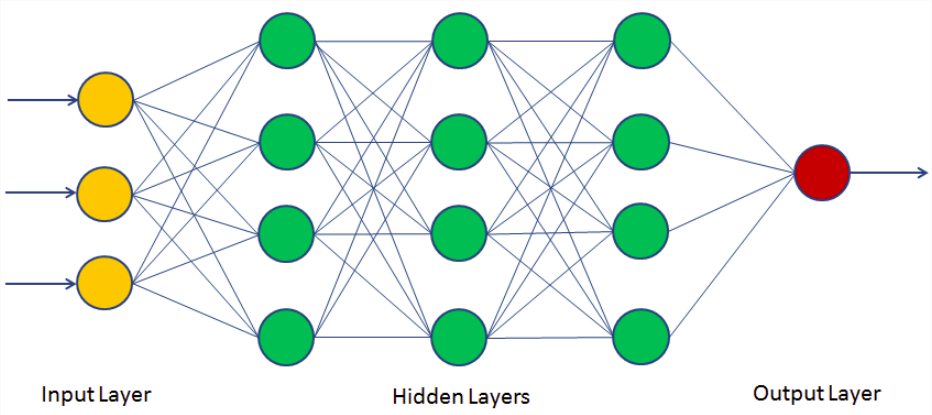
15 images

Frederik Hvilshøj — Data-Intensive Systems Group

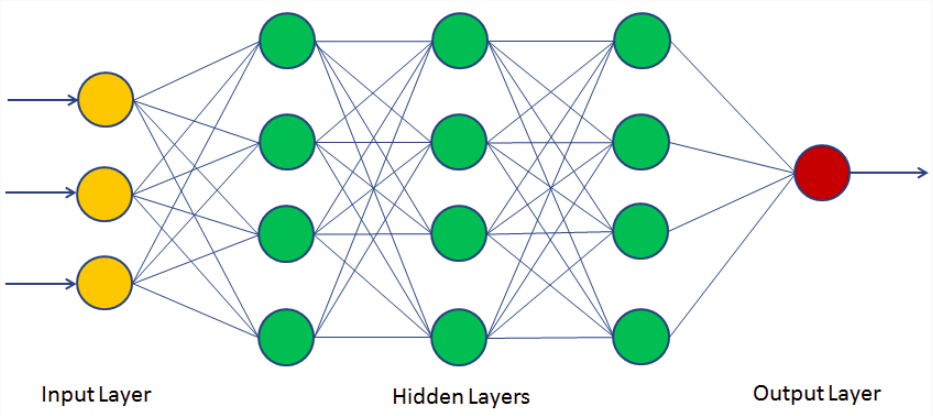
Computer Science *and* Electrical and Computer Engineering

Ira Assent *and* Alexandros Iosifidis

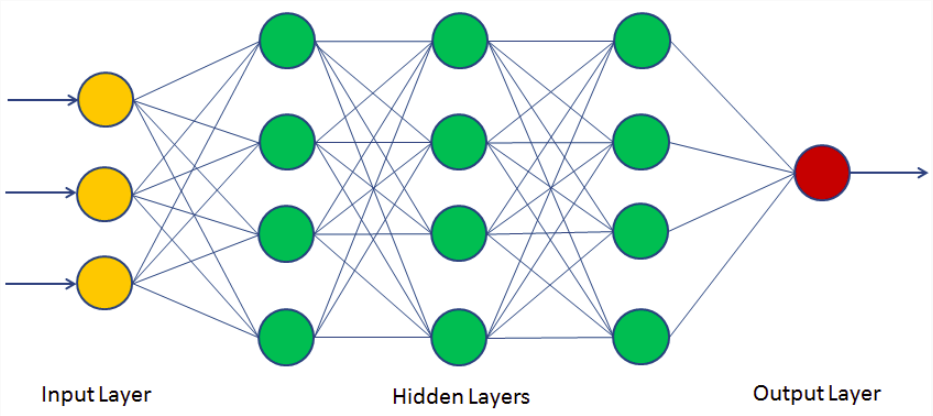
What is an Explanation?



What is an Explanation?

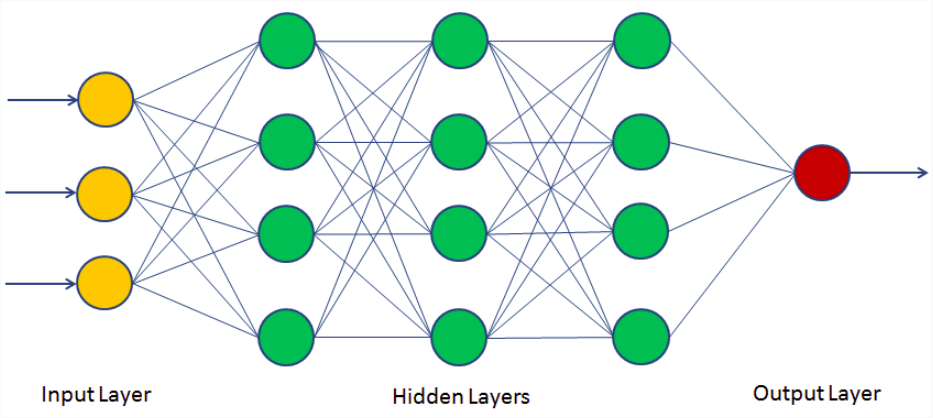


What is an Explanation?

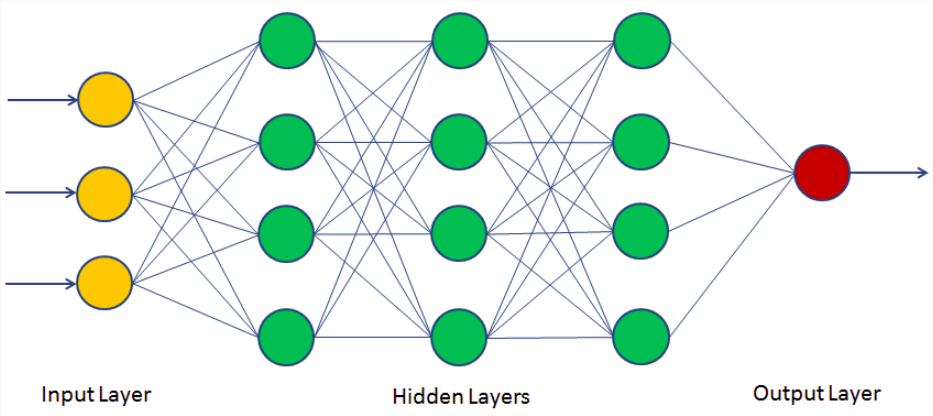


smiling

What is an Explanation?

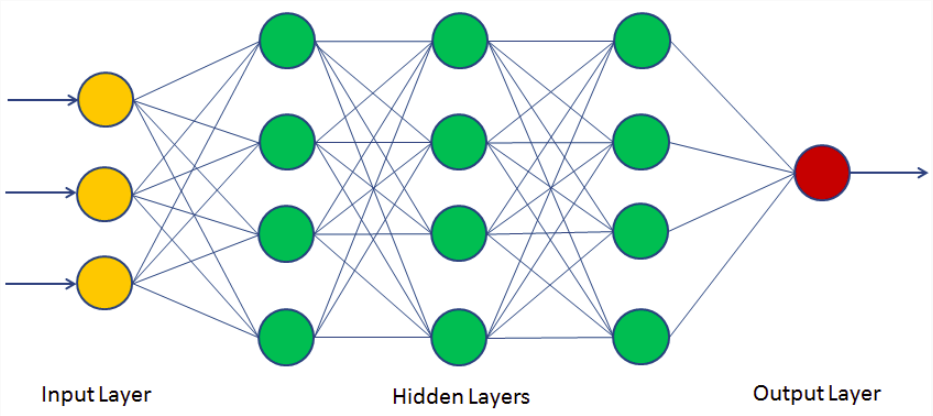


What is an Explanation?



not smiling

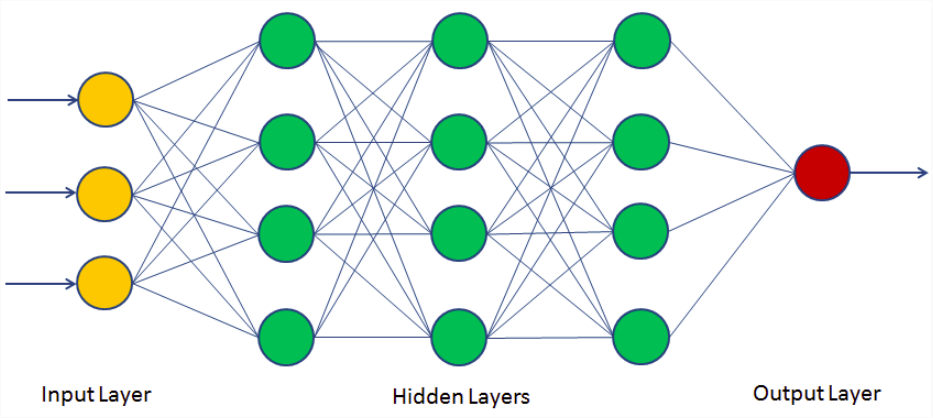
What is an Explanation?



not smiling

How do we change the prediction?

What is an Explanation?

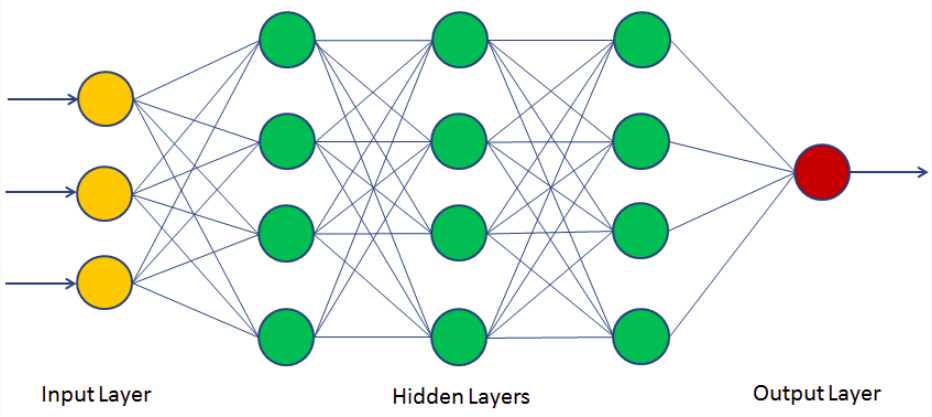


not smiling

How do we change the prediction?

→

What is an Explanation?



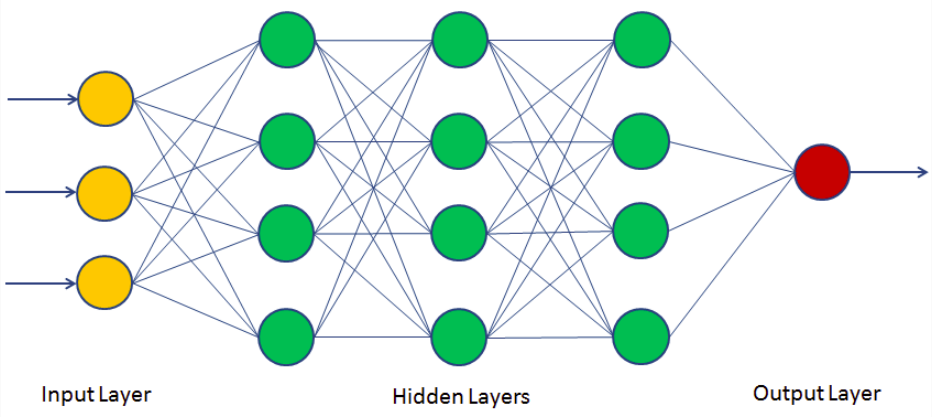
not smiling

How do we change the prediction?

→

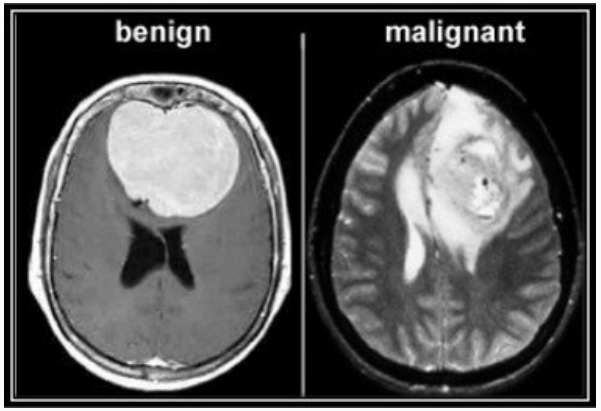
→

What is an Explanation?

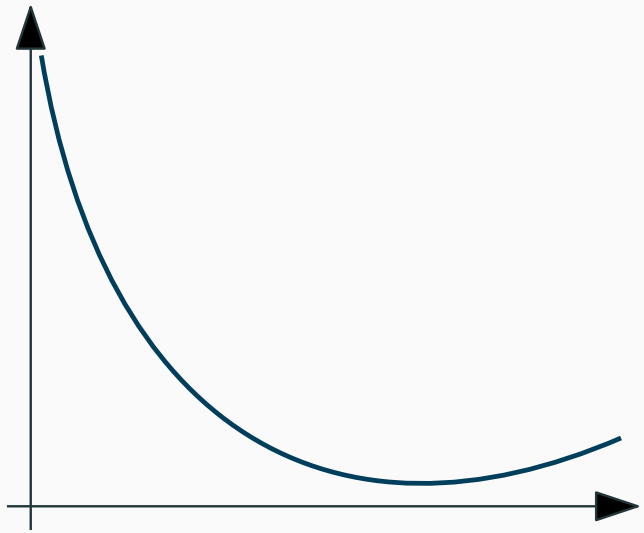


not smiling

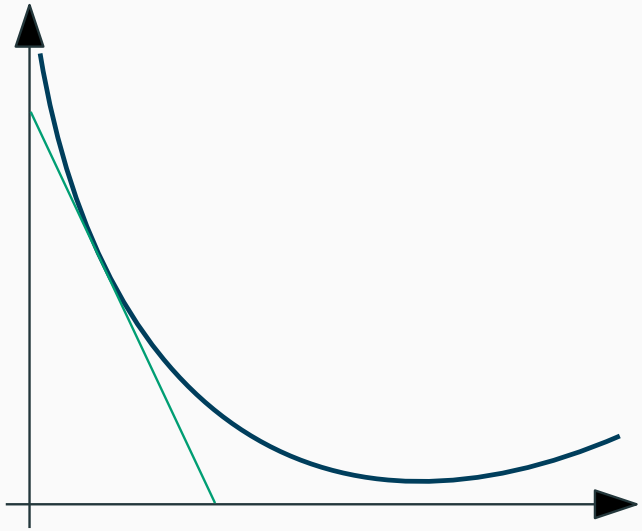
How do we change the prediction?



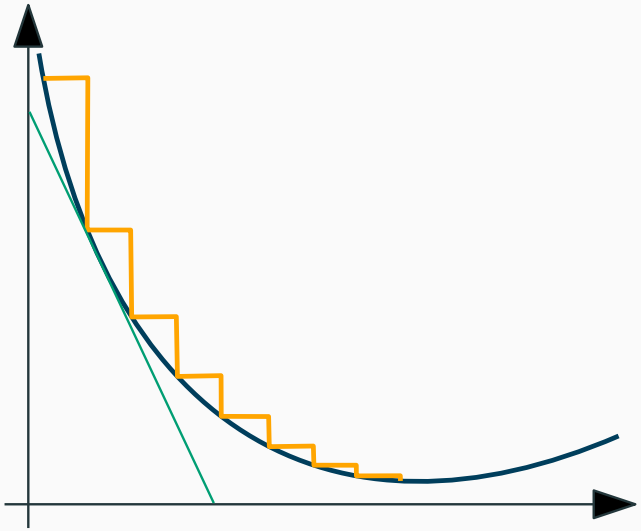
How to do it?



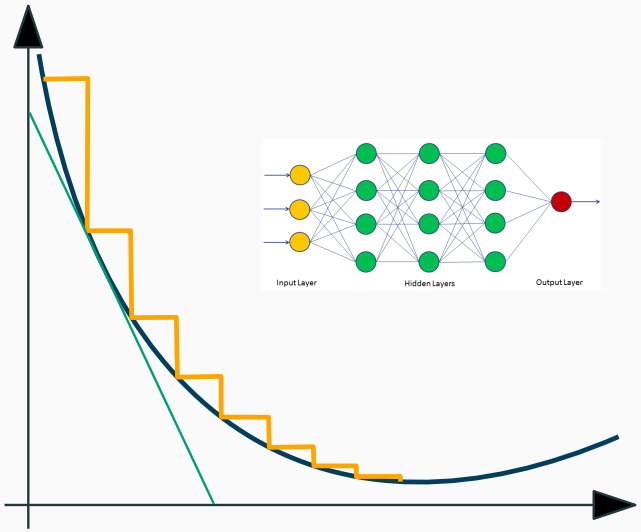
How to do it?



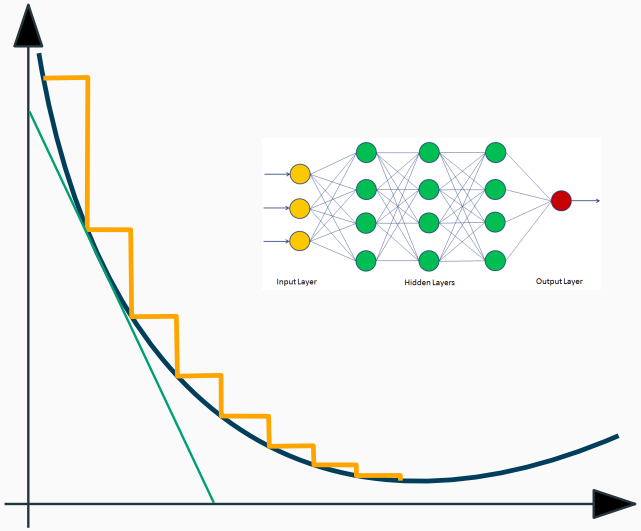
How to do it?

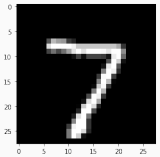


How to do it?

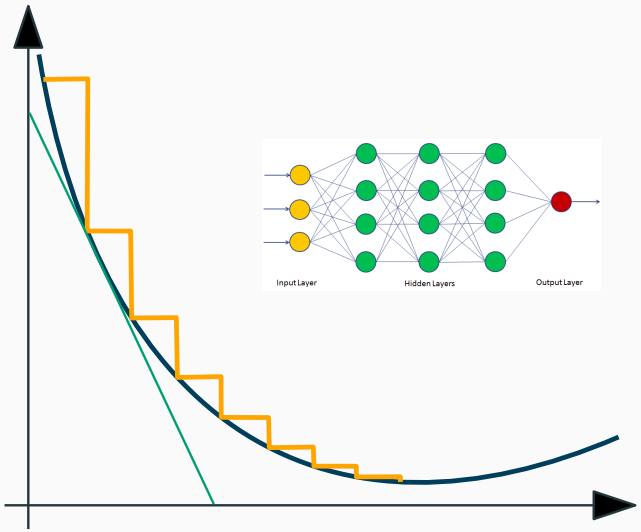


How to do it?





How to do it?



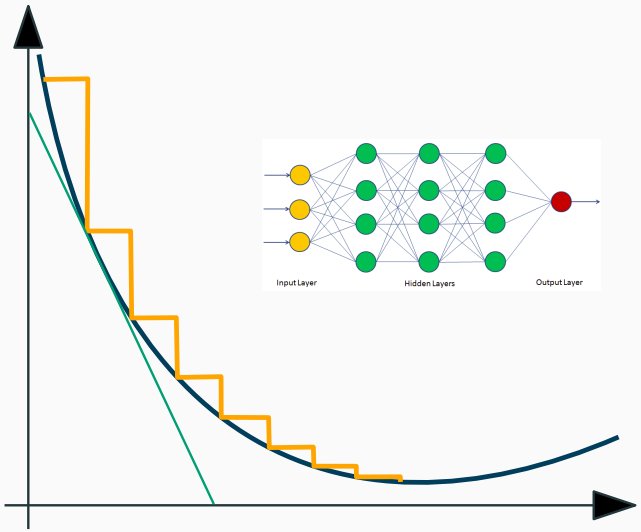
Interpretable counterfactual explanations guided by prototypes

A Van **Looveren**, J Klaise - arXiv preprint arXiv:1907.02584, **2019** - arxiv.org

We propose a fast, model agnostic method for finding interpretable counterfactual explanations of classifier predictions by using class prototypes. We show that class prototypes, obtained using either an encoder or through class specific kd trees, significantly speed up the the search for counterfactual instances and result in more interpretable explanations. We introduce two novel metrics to quantitatively evaluate local interpretability at the instance level. We use these metrics to illustrate the effectiveness of our method on an …

☆ Citeret af 61 Relaterede artikler Alle 4 versioner »

How to do it?

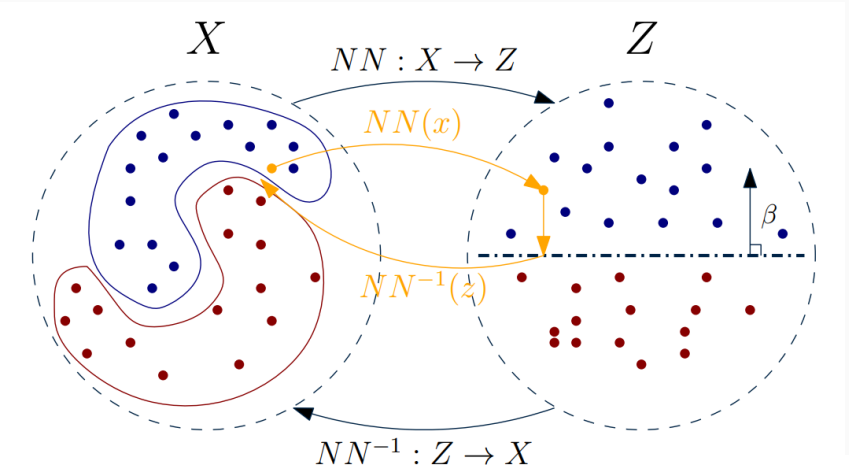


Interpretable counterfactual explanations guided by prototypes

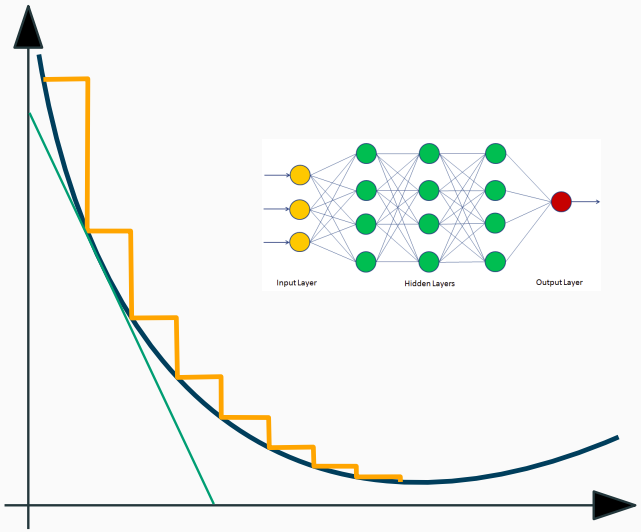
A Van **Looveren**, J Klaise - arXiv preprint arXiv:1907.02584, **2019** - arxiv.org

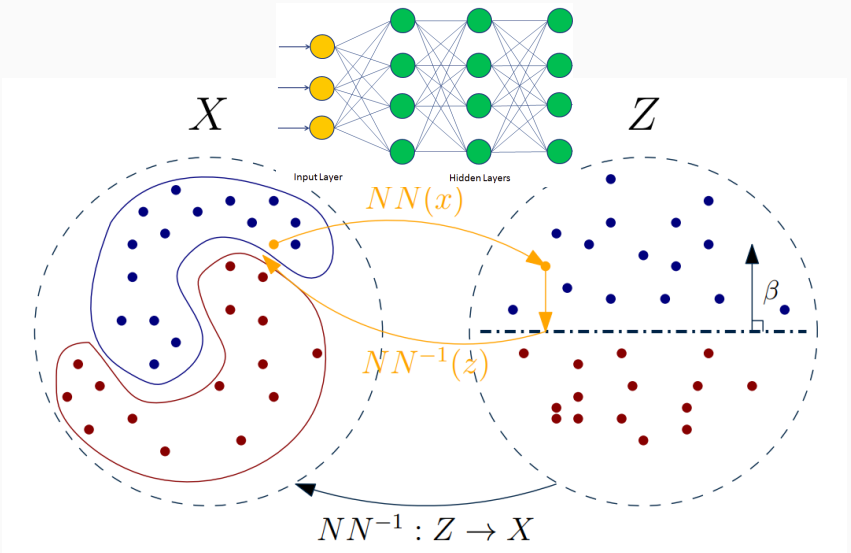
We propose a fast, model agnostic method for finding interpretable counterfactual explanations of classifier predictions by using class prototypes. We show that class prototypes, obtained using either an encoder or through class specific kd trees, significantly speed up the the search for counterfactual instances and result in more interpretable explanations. We introduce two novel metrics to quantitatively evaluate local interpretability at the instance level. We use these metrics to illustrate the effectiveness of our method on an …

☆ 99 Citeret af 61 Relaterede artikler Alle 4 versioner »



How to do it?





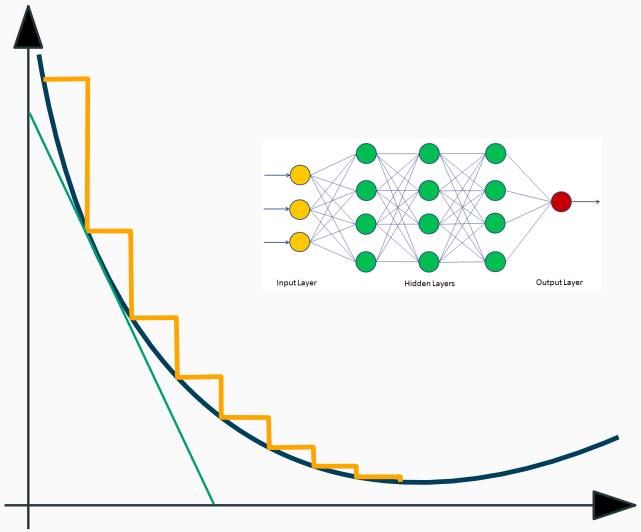
Interpretable counterfactual explanations guided by prototypes

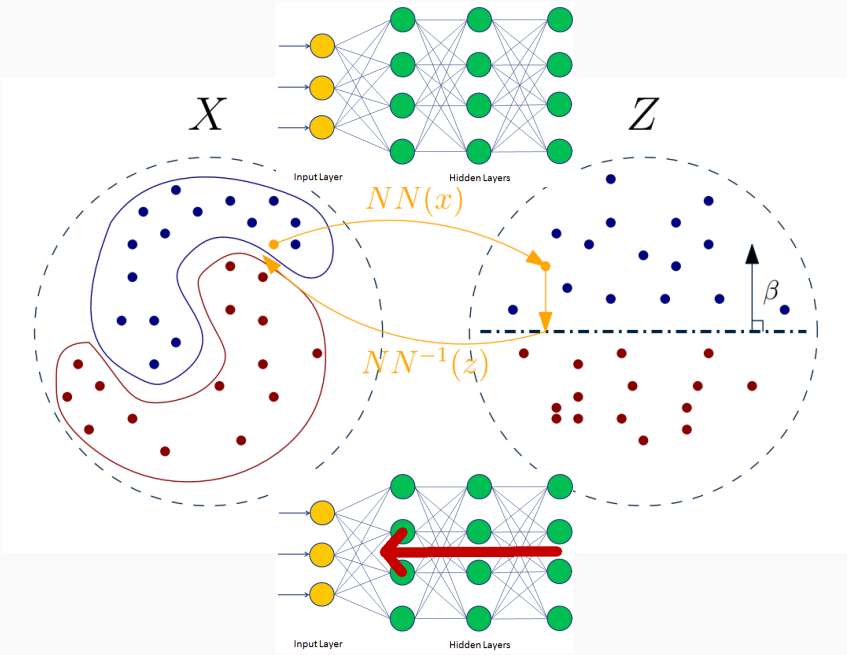
A Van **Looveren**, J Klaise - arXiv preprint arXiv:1907.02584, **2019** - arxiv.org

We propose a fast, model agnostic method for finding interpretable counterfactual explanations of classifier predictions by using class prototypes. We show that class prototypes, obtained using either an encoder or through class specific kd trees, significantly speed up the the search for counterfactual instances and result in more interpretable explanations. We introduce two novel metrics to quantitatively evaluate local interpretability at the instance level. We use these metrics to illustrate the effectiveness of our method on an …

☆ 99 Citeret af 61 Relaterede artikler Alle 4 versioner »

How to do it?





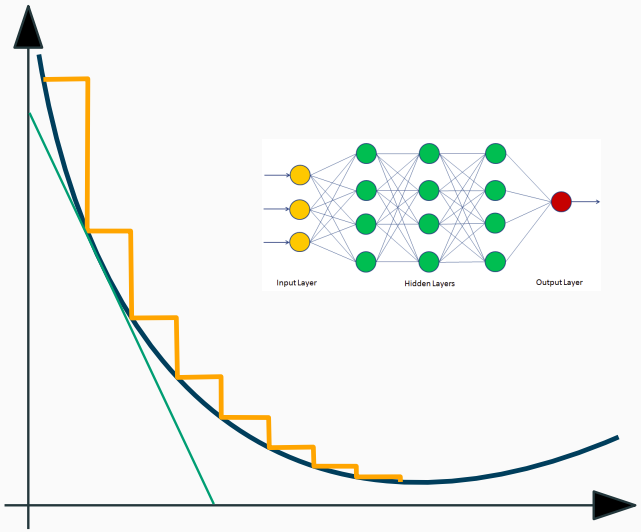
Interpretable counterfactual explanations guided by prototypes

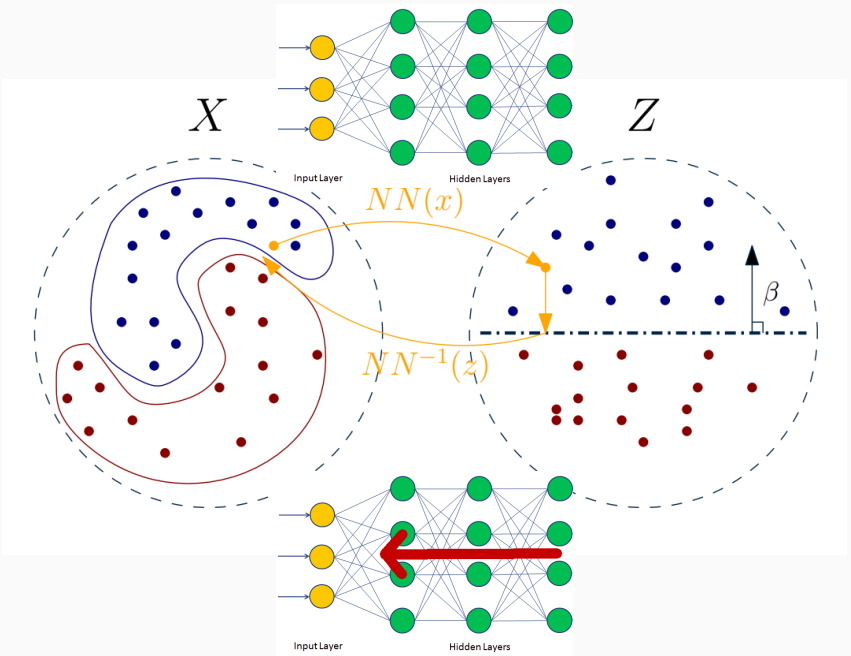
A Van **Looveren**, J Klaise - arXiv preprint arXiv:1907.02584, **2019** - arxiv.org

We propose a fast, model agnostic method for finding interpretable counterfactual explanations of classifier predictions by using class prototypes. We show that class prototypes, obtained using either an encoder or through class specific kd trees, significantly speed up the the search for counterfactual instances and result in more interpretable explanations. We introduce two novel metrics to quantitatively evaluate local interpretability at the instance level. We use these metrics to illustrate the effectiveness of our method on an …

☆ Citeret af 61 Relaterede artikler Alle 4 versioner »

How to do it?



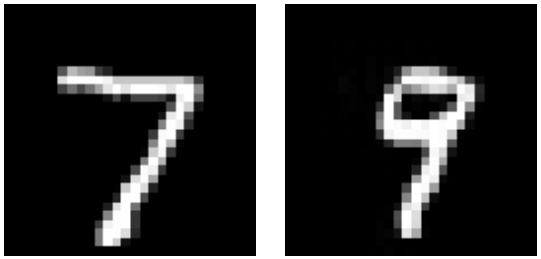


Interpretable counterfactual explanations guided by prototypes

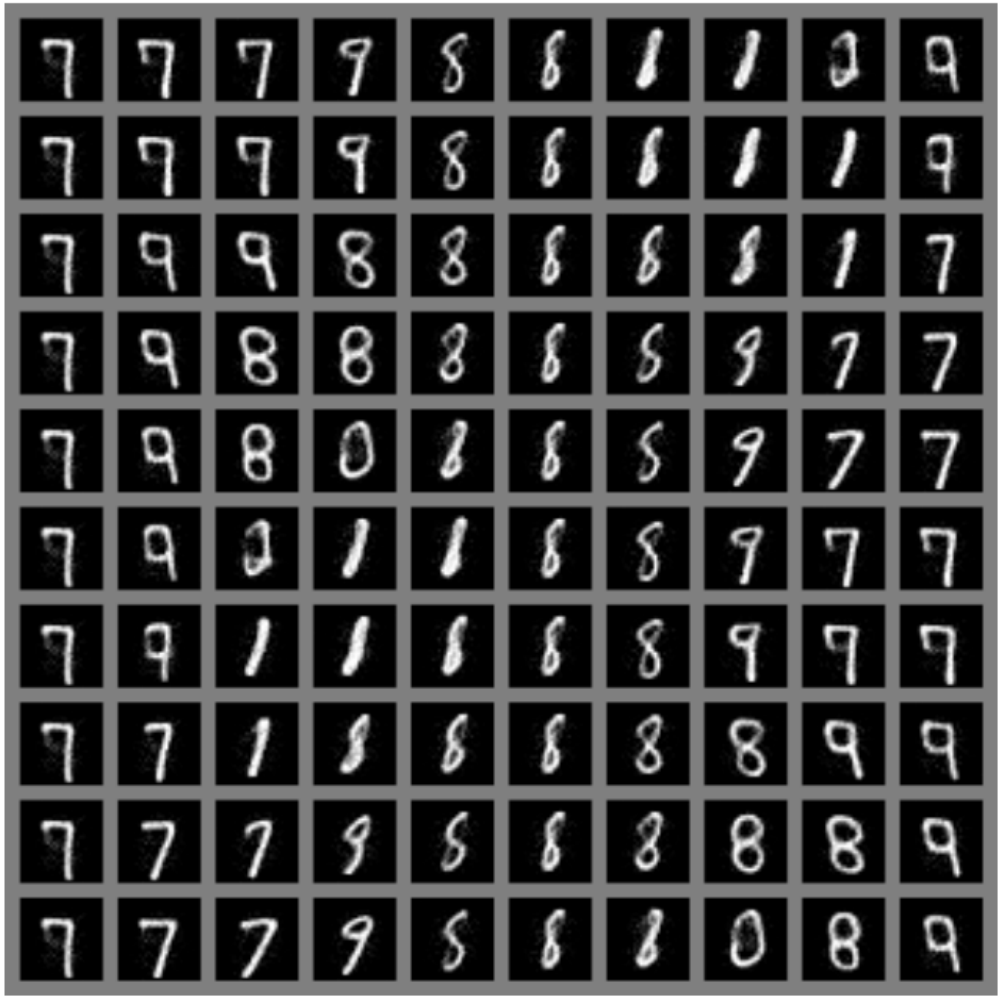
A Van **Looveren**, J Klaise - arXiv preprint arXiv:1907.02584, **2019** - arxiv.org

We propose a fast, model agnostic method for finding interpretable counterfactual explanations of classifier predictions by using class prototypes. We show that class prototypes, obtained using either an encoder or through class specific kd trees, significantly speed up the the search for counterfactual instances and result in more interpretable explanations. We introduce two novel metrics to quantitatively evaluate local interpretability at the instance level. We use these metrics to illustrate the effectiveness of our method on an …

☆ 99 Citeret af 61 Relaterede artikler Alle 4 versioner »



Why should this work?



Nice: **Non-linear independent components** estimation

L Dinh, D Krueger, Y Bengio - arXiv preprint arXiv:1410.8516, 2014 - arxiv.org

… **independent** dimensions), then we obtain the following **non-linear independent components** estimation (NICE … Directed graphical models lack the conditional **independence** structure that … is used more successfully in nonlinear **independent components analysis** (Hyvärinen and …

☆ Citeret af 798 Relaterede artikler Alle 6 versioner »

Why should this work?

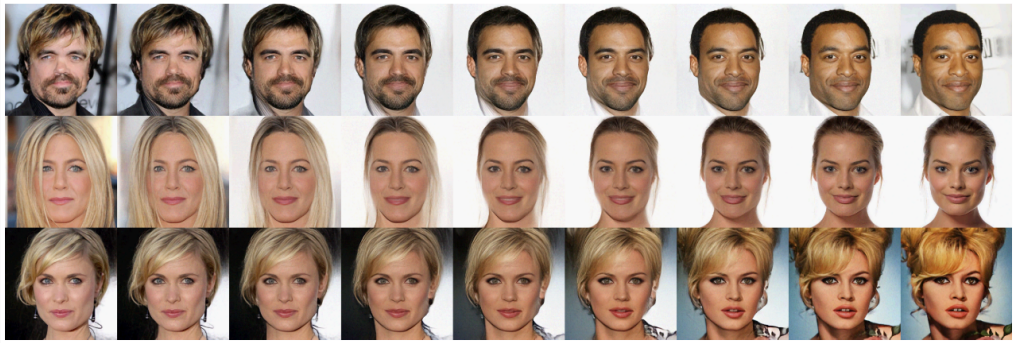
Density estimation using real nvp

L Dinh, J Sohl-Dickstein, S Bengio - arXiv preprint arXiv:1605.08803, 2016 - arxiv.org

Unsupervised learning of probabilistic models is a central yet challenging problem in machine learning. Specifically, designing models with tractable learning, sampling, inference and evaluation is crucial in solving this task. We extend the space of such models …

☆ 99 Citeret af 1185 Relaterede artikler Alle 11 versioner »

Why should this work?



Glow: **Generative flow** with invertible 1x1 convolutions

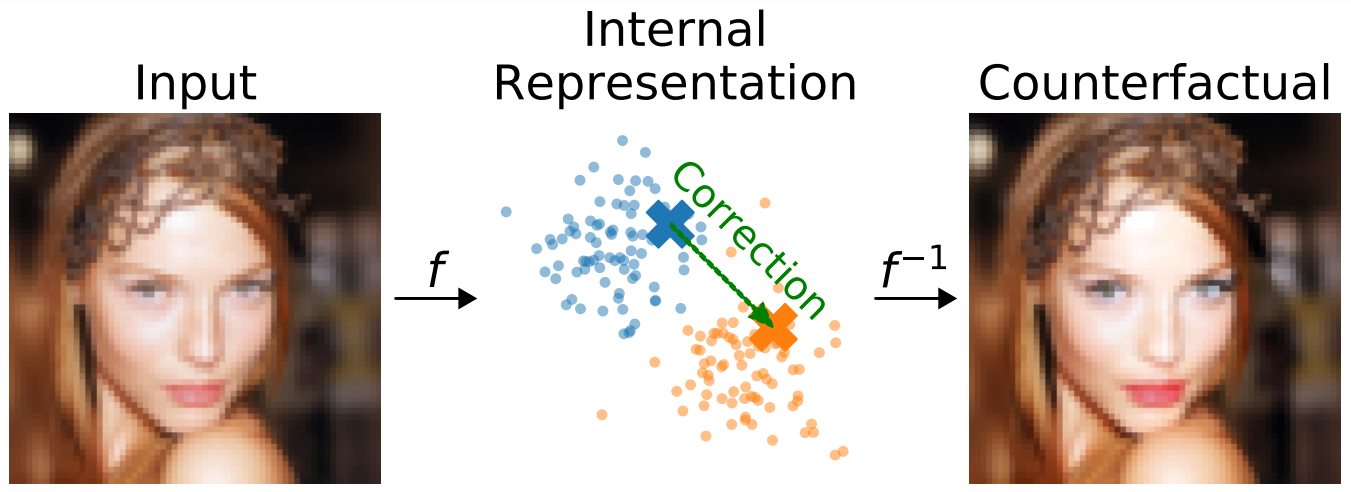
DP Kingma, P Dhariwal - arXiv preprint arXiv:1807.03039, 2018 - arxiv.org

Flow-based **generative** models (Dinh et al., 2014) are conceptually attractive due to tractability of the exact log-likelihood, tractability of exact latent-variable inference, and parallelizability of both training and synthesis. In this paper we propose **Glow**, a simple type …

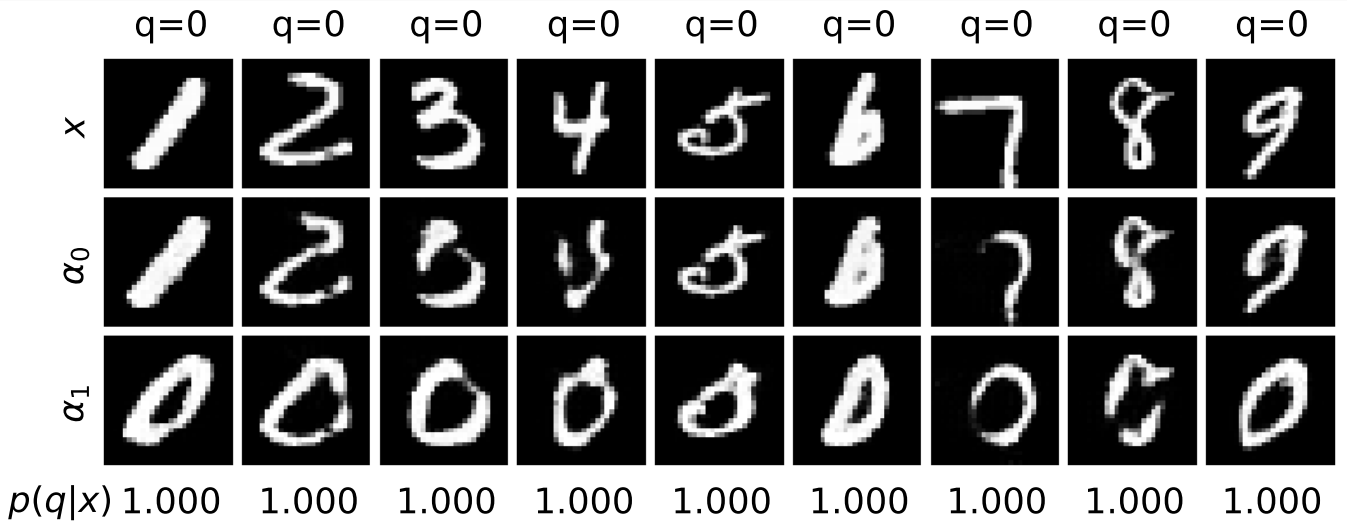
☆ 99 Citeret af 989 Relaterede artikler Alle 6 versioner »

So does it work?

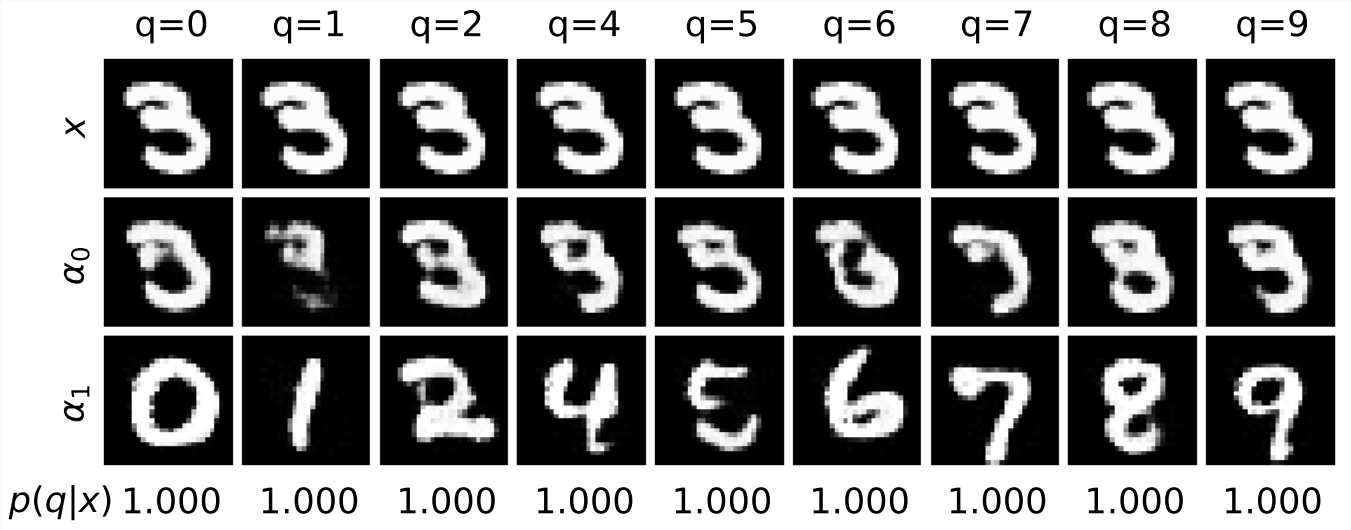
So does it work?



So does it work?



So does it work?



So does it work?