ConeOpt: An Optimization Approach for Counterfactual Explanations

June 18, 2020

Abstract

We propose an optimization approach for counterfactual explanations.

Key words. Counterfactual explanations.

1 Methodology

Given an original instance x^0, we attempt to find a counterfactual instance $x' := x^0 + d' \in \mathbb{R}^n$.

We first define the following loss function that encourages the predicted class t of the perturbed instance x' to be different than the predicted class t_0 of the original instance x^0.

$$f_p(x';\kappa) := \max\left\{ f(x';t_0) - \max_{t\neq t_0} f(x';t), \kappa \right\}, \tag{1}$$

where $f(x';t)$ is the t-th class prediction probability, and $\kappa > 0$ caps the divergence between x^0 and x'.

We also define another objective term to generate a sparse counterfactual instance that is similar to current instance:

$$f_d(x^0, x') := \beta\|x^0 - x'\|_1 + \|x^0 - x'\|_2, \tag{2}$$

where β is a positive parameter that balances the trade-off between the sparsity and similarity of the counterfactual instance x'.

To ensure the data distribution of x' lies closer to all neighboring instances in the same class, we introduce a third objective term:

$$f_d(x^0, \{x^j \mid j \in K_t\}) := -(1/|K_t|) \sum_{j\in K_t} (x^j - x^0)^T (x' - x^0). \tag{3}$$

Note that (3) encourages a counterfactual instance x' with a larger sample variance associated with samples in K_t. Geometrically, it defines a direction $x' - x^0$ in feature space along which these data vary the most.

To define K_t, we need a representative, unlabeled sample of the training dataset. First the predictive model is called to label the dataset with the classes predicted by the model. Then for each class t we encode the instances belonging to that class and order them by increasing L_2 distance to $\mathrm{ENC}(x_0)$. The K_t nearest instances in the latent space are included in K_t.

With (1)–(3), we now formally define our optimization model:

$$\begin{aligned}
\min \quad & f_p(x';\kappa) + c_d f_d(x^0, x') - c_p f_p(x', \{x^j \mid j \in K_t\}) && \text{(4a)}\\
\text{s.t.} \quad & x' = x^0 + \sum_{j=1}^{n_k} \lambda_j d^j, && \text{(4b)}\\
& d^j = x^j - x^0, j \in K_t && \text{(4c)}\\
& \sum_{j=1}^{n_k} \lambda_j \le \delta, && \text{(4d)}\\
& \lambda \ge 0. && \text{(4e)}
\end{aligned}$$

1.1 Trajectory towards counterfactuals

Now we present an algorithm that finds an trajectory towards the counterfactual.

Algorithm 1.1 *ConeOpt*

Input: An instance $x^0 \in \mathcal{X}$ to explain, and an index set K_t.

Output: A set of instances L that provides a trajectory towards the counterfactual instances.

Let $L := \{\phi\}$.
while Termination criteria not met **do**
    Solve optimization problem (4) to obtain a new instance x'.
    Update trajectory $L := L \cup \{x'\}$.
    Let $x^0 := x'$.
    Calculate the new mean of each cluster.
end while