Counterfactual Explanations for Support Vector Machine Models

Sebastian Salazar
Columbia University

Samuel Denton
Columbia University

Ansaf Salleb-Aouissi
Columbia University

Abstract

We tackle the problem of computing counterfactual explanations —minimal changes to the features that flip an undesirable model prediction. We propose a solution to this question for linear Support Vector Machine (SVMs) models. Moreover, we introduce a way to account for weighted actions that allow for more changes in certain features than others. In particular, we show how to find counterfactual explanations with the purpose of increasing model interpretability. These explanations are valid, change only actionable features, are close to the data distribution, sparse, and take into account correlations between features. We cast this as a mixed integer programming optimization problem .

Additionally, we introduce two novel scale-invariant cost functions for assessing the quality of counterfactual explanations and use them to evaluate the quality of our approach with a real medical dataset. Finally, we build a support vector machine model to predict whether law students will pass the Bar exam using protected features, and used our algorithms to uncover the inherent biases of the SVM.

1   Introduction

Support Vector Machines (SVMs) are one of the best-performing discriminative models in machine learning. Despite recent advancements in areas like Deep Learning, SVMs manage to remain competitive and come with several theoretical guarantees on stability and sample-complexity. Support vector machines are particularly good for high dimensional data since they belong to a class of learners that learn hyperplanes with low ℓ_2 norm.

In general, these concept classes have low Rademacher Complexity, which means that they tend to generalize well and have a low sample complexity. This goes hand-in-hand with the "wide margin" property of SVMs [24].

To ensure responsible use of Support Vector Machines, the aforementioned theoretical guarantees are paramount — especially in safety-critical and/or high-impact scenarios like healthcare, finance, and threat assessment. What's more, SVMs seem to perform relatively well even when data doesn't abound. This is in stark contrast to the more popular deep learning models which usually require relatively large datasets.

While there is extensive theoretical research on Support Vector machines [4, 30], there is little to no information on how to use the wide-margin property to change the labels of instances with undesirable predictions and how this can be used to enhance interpretability. As an example, consider an instance with a predicted undesirable outcome (e.g., mortgage application rejected), how do we minimally change the features to flip the prediction of the original data? Explanations of this form give the decision-maker feedback on what features are most relevant to the model in the decision-making process and are known as **counterfactual explanations**. Providing explanations of this form has become increasingly important, especially in cases where automated decision-making has the potential to drastically impact human lives [32]. Furthermore, legal regulations —like the European Union's General Data Protection Regulations— are demanding responsible deployment of machine learning models. As such, it is important to ensure that these models are used responsibly and ethically in practice.

In particular, we propose an integer programming formulation to finding counterfactual explanations for support vector machines. Taking advantage of the extra structure that comes with their wide-margin property, we show that counterfactual explanations from support vector machine models are stable. Additionally, evaluating the quality of counterfactual explanations is a well-known problem in the literature, and an active area of research. We propose two novel cost functions to evaluate the quality of counterfactual explanations that are scale invariant, and take on low values when the counterfactuals have desirable statistical properties.

Previous research (e.g., [5, 16, 21, 26]) has defined *actionability* as giving a recommendation of the *easiest* way for a user to change an undesirable model prediction, based on a set of rules. While rule-based classifiers allow for interpretability of the feature decisions, they are not the best performing models.

We illustrate our approach with a *diabetes* dataset, a serious disease characterized by eleveted levels of blood sugar. Over time, this can lead to serious damage to the heart, blood vessels, brain and kidneys. Diabetes has no known cure, however, in some cases (i.e., Type II diabetes), prevention can play an important role in reducing a patient's risk. This is done by managing risk factors like weight, BMI, physical activity, blood pressure, glucose levels in blood, etc. The way these factors interact is not fully understood, and controlling all of these factors simultaneously can be highly unrealistic for most patients.

With diabetes, as in many other health problems, one challenge that a physician might face is to make important decisions about what should be done to cure a patient or reduce the risk of a specific disease. It has been increasingly common to see Machine Learning models getting incorporated into the decision-making process. As such it is paramount to be able to understand how these models make decisions.

We define an action as a change in the value of some feature (e.g., weight of a patient, or the patient's lack of physical activity). Our approach focuses on a methodology where a datapoint is framed as a vector of features. We will then use the vector to predict an outcome based on the given SVM model. The goal will be to minimize the *weighted distance* of the action that shifts the features to a point along the margin of the desired outcome. The weighted distance will be some combination of the distance traveled in each feature multiplied by a weight that accounts for the varying scales of each feature.

Although there is recent literature on computing counterfactual explanations for linear classifiers, we are —to the best of our knowledge— the first to propose an approach that takes advantage of the wide-margin property of support vector machines.

2 Related Work

Most research on learning counterfactual explanations goes back to rule-based models, and can be classified into three main lines of research: (1) use of predefined set of simple actions that are mapped to classification rules or deviations [1, 6, 8, 19]; (2) sift through a set of association or subgroup discovery rules [7, 14, 15] to identify descriptive actionable rules of the form Condition → Class, (3) elaborate actions by comparing pairs of classification rules with contradicting classes to generate action rules

[20, 21, 22, 28]. Given an "unsatisfactory prediction", the work in [26, 27] discovers a set of action recommendations to improve that situation using classification rules. Finally, actionability is formalized as a case-based reasoning problem in [34, 33]. Given an example to reclassify, the authors propose to search for the closest cases with the desirable class label, among training examples, centroids of the clusters of the opposite class, or SVM support vectors. Choosing the closest support vector will be used as a baseline to compare our proposed approach.

Recent research on the interpretability and fairness of machine learning models also looks at how counterfactual explanations can be used to understand how a model makes decisions (e.g., [13, 12, 29, 9, 23, 11, 10]). The transformed instance that results from making changes to the feature values of a given instance in order to switch its label, is called a nearest *counterfactual* explanation or *actionable recourse*. [29] frame the problem as an *audit* task. They formulate an optimization problem expressed as an integer program solved by an IP solver, to uncover minimal cost actions in linear classifiers. Other lines of work that use IP solvers to compute counterfactual explanations include [3, 2, 3]. [32] present a comprehensive approach to counterfactual explanations without opening the black box model. Their approach is motivated by the need for the algorithmic accountability stated in the rules of the General Data Protection Regulation (GDPR). In the same context of the existence of a black box model and no information about the training data, [13] identify the closest point classified differently using a growing sphere algorithms. The idea is to find the minimal change needed to inverse the classification. Karimi et al. in [9] propose MACE, a model-agnostic counterfactual explanations approach, that uses a sequence of satisfiability (SAT) problems to find the nearest counterfactual. MACE requires to convert the model into a characteristic First Order Logic formulas. However, the applicability of their approach depends on the possibility of such a conversion (which is challenging for complex non linear models) and the efficiency of a SAT solver used as an oracle. Algorithmic recourse proposed in [10] aims to shift from counterfactual explanation to minimal interventions using a causal graph. In this context, the explanations of our model should be interpreted as causal when the SVM model only uses data from the parents of the target variable in the causal diagram for prediction. Additionally, it is worth noting that methods like inverse probability weighting could be used alongside our method to calculate the average causal effect. We leave the causal aspect of this line of research to future work.

3 Actionability with SVMs

The problem of computing robust counterfactuals for Support Vector Machines is modeled as a Mixed Inte-

ger Quadratic Program (MIQP). In this line of work, we extend the research of [29], [3], and [32] in order to: (a) leverage the wide-margin property of SVMs, and (b) handle the problem of `actionability`.

Given an instance of the training data $x \in \mathcal{X}$ with an unfavorable outcome $y \in \{\pm 1\}$, our goal is to compute the closest counterfactual $x' \in \mathcal{X}$, with label $y' = -y$ that respects the margin constraints of the Support Vector Machine. In its most general form, this is done by solving the following optimization problem

$$\underset{x' \in \mathcal{X}}{\arg\min} \quad d(x, x')$$
$$\text{s.t.} \quad y'\big(\langle w, x' \rangle + b\big) \geq 0 \quad (1)$$
$$\big|\langle w, x' \rangle + b\big| \geq 1. \quad (2)$$

Where w and b are the coefficients of the Support Vector Machine.
Now, note that the original optimization problem (1) is equivalent to:

$$\underset{x' \in \mathcal{X}}{\arg\min} \quad d(x, x')$$
$$\text{s.t.} \quad \langle w, x' \rangle + b + y \geq 0.$$

All that remains is to choose a metric to handle the problem of actionability. We choose the weighted Euclidean distance metric to reflect the fact that features may exhibit deferent length scales. Using this metric the optimization problem becomes

$$\underset{x' \in \mathcal{X}}{\arg\min} \quad \left\| W^{1/2}(x - x') \right\|_2^2 \quad (3)$$
$$\text{s.t.} \quad \langle w, x' \rangle + b + y \geq 0.$$

Where $W \in \mathbb{R}^{n \times n}$ is a diagonal matrix of positive weights to be chosen by the user. From an intuitive standpoint, changes in coordinates with a higher weight are penalized more heavily and as a result the optimization problem will tend to leave those coordinates unchanged.
In practice, datasets usually have a mix of categorical and real-valued features, meaning that $\mathcal{X} \cong \mathbb{Z}^k \times \mathbb{R}^{n-k}$. Non-ordinal categorical features are usually one-hot encoded, and any valid counterfactual explanation should maintain this property. To handle this restriction, we could introduce constraints by considering a partition $\{J_i\}_{i=1}^m$ of the set $[k]$, where the sets $\{J_i\}_{i=1}^m$ contain the entries of a one-hot encoded feature, and enforce that $\forall i : \sum_{k \in J_i} x_k = 1$. With these additional constraints, the optimization problem becomes

$$\underset{x' \in \mathbb{Z}^k \times \mathbb{R}^{n-k}}{\arg\min} \quad \left\| W^{1/2}(x - x') \right\|_2^2$$
$$\text{s.t.} \quad \langle w, x' \rangle + b + y \geq 0.$$
$$\forall i \in [m] : \sum_{s \in J_i} x'_s = 1.$$

This is just a MIQP which we express in standard form below

$$\underset{x' \in \mathbb{Z}^k \times \mathbb{R}^{n-k}}{\arg\min} \quad \frac{1}{2} x'^\mathsf{T} (2W) x' + (-2Wx)^\mathsf{T} x'$$
$$\text{s.t.} \quad \langle w, x' \rangle + b + y \geq 0 \quad (4)$$
$$\forall i \in [m] : \sum_{s \in J_i} x'_s = 1.$$

4 Desirable properties of counterfactual explanations.

The method that we have introduced so far does not have several desirable properties of counterfactual explanations. For instance, it fails to account for the relationships between input variables.
In general, evaluating the quality of counterfactual explanations is a well-known problem and an active area of research. In the following subsection we address how to deal with the issues of feature correlation, plausibility, sparsity, actionability, and stability —all of which are desirable properties of counterfactual explanations. In the context of integer programming, the problem of computing counterfactual explanations with these properties has already been explored in the literature ([3, 32, 29]. Moreover, we propose ways of adapting these ideas to support vector machines. Similarly, we introduce two novel cost functions for assessing the quality of counterfactual explanations. For a more comprehensive survey about the evaluation of counterfactual explanations, the reader is referred to [31].

4.1 Stability

We remark that an important property of the solutions of (1) is that they are stable in the sense that they are robust to perturbations. To see why this is the case, note that enforcing the wide-margin constraints of (2) implies that one can find an open neighborhood of the solution x' —call it $\mathcal{N}_{x'}$ — such that for every $v \in \mathcal{N}_{x'}$ it holds that $y'(\langle w, v \rangle + b) \geq 0$ (i.e., v has the same label as x'). This implies that, for every $v \in \mathcal{N}_{x'}$, there is a $\delta \in \mathcal{X}$ such that $v = x' + \delta$. In other words, we could perturb the counterfactual x' without changing its label! More formally,

Definition 4.1. Let $\mathcal{X}$ be a Hilbert Space over $\mathbb{R}$, $\mathcal{H} \subseteq \{f : \mathcal{X} \to \{\pm 1\}\}$ be a class of functions and $h \in \mathcal{H}$ be a classifier. Let $x \in \mathcal{X}$ be such that $h(x) = y$. We say $x' \in \mathcal{X}$ is a δ-stable counterfactual explanation for x relative to h if and only if there exists an open ball $\mathcal{B}(x', \delta)$ such that for every $v \in \mathcal{B}(x', \delta)$, we have that $h(v) = y' := -y$.

Theorem 4.2. Let $\mathcal{D} \subseteq (\mathcal{X} \times \{\pm 1\})^m$ be a dataset, $H \subseteq \{x \mapsto \text{sign}(\langle x, w \rangle + b)\}$, be the concept class of linear functions, and $f \in H$ be a linear support vector machine with $\gamma = \sup_{(x,y) \in \mathcal{D}} |\langle w, x \rangle + b|$. Let $x, x' \in \mathcal{X}$ be such that $f(x) = y$,

$f(x') = -y := y'$, and

$$x' = \arg\min_{x' \in \mathcal{X}} \quad \|x - x'\|$$
$$s.t. \quad y'(\langle w, x' \rangle + b) \geq 0 \qquad (5)$$
$$|\langle w, x' \rangle + b| \geq 1.$$

then x' is a $\gamma/\|w\|$-stable counterfactual explanation for x.

Proof. Note that, without loss of generality, we can always re-scale the parameters of the support vector machine to be such that

$$\gamma = \sup_{(x,y) \in \mathcal{D}} |\langle w, x \rangle + b|$$
$$= 1.$$

Thus, it suffices to show that the solution to 5 is a $1/\|w\|$-stable counterfactual explanation.
WLOG, take $y' = 1$ and $b = 0$. Now, let x' be a solution to 5, then it follows that $\langle w, x' \rangle \geq 1$. Let $H = \{\xi : \langle w, \xi \rangle \geq 0\}$ be the closed halfspace whose boundary ∂H is the hyperplane defined by the vector w. Since $\langle x', w \rangle \geq 1$, it follows that $x' \in H \backslash \partial H$, meaning that x' is an interior point of H. This means that there is an $\eta \in \mathbb{R}$ such that the open ball $B(x', \eta)$ is fully contained in H; that is $B(x, \eta) \subseteq H$. Now, take η to be the distance of the closest point from x' to the boundary of H (i.e., the hyperplane ∂H defined by w). More formally, take $\eta = d(x', \partial H) := \inf_{\partial h \in \partial H} \|x' - \partial h\|$. Then it is clear that $B(x', \eta) \subseteq H$. However, note that η is just the distance of the counterfactual x' to the hyperplane, which is given by $1/\|w\|$. This shows that $B(x', 1/\|w\|) \subseteq H$, meaning that every $v \in B(x', 1/\|w\|)$ is such that $\langle v, w \rangle \geq 0$ which implies that $f(v) = sign(\langle v, w \rangle) = sign(\langle x', w \rangle) = 1$, completing the proof. □

Remark 4.3. Recall that the optimization problem of the support vector machine is given by:

$$\max \frac{1}{\|w\|}$$
$$s.t. \quad \forall i : y_i(\langle w, x_i \rangle) \geq 1.$$

It is worth noting that the stability of counterfactual explanations defined by solutions to (5) scale like $1/\|w\|$ —the quantity being maximized by the SVM! Thus, it that follows that these counterfactual explanations are maximally stable.

4.2 Accounting for correlations between features

[3] first introduced a method to account for correlations between features that and can easily be integrated to SVM-ACE. In particular, given the action of a counterfactual defined by $\delta := x' - x$, we set our new counterfactual

to be $x_{cf} := x + \Sigma\delta$, where Σ is the empirical covariance matrix. This new counterfactual x_{cf} will account for linear relationships between the features. However, instead of a post-hoc application of the covariance matrix, we note that the application of the covariance matrix can be directly incorporated into the optimization problem in the following way:

$$\arg\min_{x' \in \mathbb{Z}^k \times \mathbb{R}^{n-k}} \quad \left\| W^{1/2}\Sigma^{-1/2}(x - x') \right\|_2$$
$$s.t. \quad \langle w, x' \rangle + b + y \geq 0 \qquad (6)$$
$$\forall i \in [m] : \sum_{s \in J_i} x'_s = 1.$$

Which is just a weighted version of the Manhalanobis distance. This distance measure has been widely studied in the literature and it has the desirable property of accounting for the correlations between features.

4.3 Plausibility

When computing counterfactual explanations, it is important to ensure that they are not outliers with respect to the observed data. In other words, we want the computed counterfactual explanations to be `plausible`. In the literature, plausibility is usually enforced through the use a model that is used to learn class prototypes (i.e., a vector that is a good representative for examples with a certain label) for each of the labeled examples. In this case, use linear constraints to guarantee that the counterfactual is "close" to the class prototype of the class $y' = -y$. In particular, let v_1 and v_{-1} be prototypes for the examples with labels of 1 and -1 respectively. We enforce the linear constraint $|x' - v_{y'}| \lessapprox \epsilon$, which is equivalent to $\|x' - v_{y'}\|_\infty \leq \epsilon$, for some hyperparameter $\epsilon \in \mathbb{R}^+$. With this new plausibility constraint, the new MIQP is written as:

$$\arg\min_{x' \in \mathbb{Z}^k \times \mathbb{R}^{n-k}} \quad \left\| W^{1/2}(x - x') \right\|_2^2$$
$$s.t. \quad \langle w, x' \rangle + b + y \geq 0. \qquad (7)$$
$$\forall i \in [m] : \sum_{s \in J_i} x'_s = 1$$
$$\|x' - v_{y'}\|_\infty \leq \epsilon.$$

4.4 Sparsity

In some cases, it is desirable that the computed counterfactuals are sparse so that a minimal number of features have to be changed to achieve a desirable outcome. To get sparse counterfactual explanations, a common approach is to choose a metric like the weighted ℓ_1 norm, that is used for computing sparse solutions to optimization problems. It is well-known in the literature that the best counterfactual explanations are computed using

the ℓ_1 norm. In this case, the problem can be converted into an equivalent Mixed Integer Linear Program (MILP) through the introduction of auxiliary variables. In particular, consider the following modified version of (3):

$$\underset{x' \in \mathbb{Z}^k \times \mathbb{R}^{n-k}}{\arg\min} \quad \left\| W(x - x') \right\|_1$$
$$\text{s.t.} \quad \langle w, x' \rangle + b + y \geq 0$$
$$\forall i \in [m]: \sum_{s \in J_i} x'_s = 1.$$

To turn this into an equivalent MILP, it is a standard trick to introduce auxiliary variables d_i and enforce the constraint that $\left| \sum_j W_{ij}(x_j - x'_j) \right| \leq d_i$. Doing this, it is straightforward to show that we obtain the following equivalent MILP:

$$\underset{x', d \in \mathbb{Z}^k \times \mathbb{R}^{n-k}}{\arg\min} \quad 1^\mathsf{T} d$$
$$\text{s.t.} \quad \langle w, x' \rangle + b + y \geq 0.$$
$$\forall i \in [m]: \sum_{s \in J_i} x'_s = 1$$
$$\forall i \in [n]: \sum_j W_{ij}(x_j - x'_j) \leq d_i$$
$$\forall i \in [n]: \sum_j W_{ij}(x'_j - x_j) \leq d_i$$
$$\forall i \in [n]: 0 \leq d_i.$$

4.5 Actionability

When using these methods to achieve algorithmic recourse, it is very important to account for the fact that some features may not be actionable. In this case we want to change the decision of a model by altering actionable input variables (i.e., features that are mutable or can be changed). Our method can naturally handle actionability through the (usually diagonal) weight matrix W. In this case, the entries W_{ii} represent the penalty of suggesting changes to the ith feature. Since this is a minimization problem, a higher weight means that changes in that features are penalized more heavily, and as a result aren't changed as much by the optimization problem. In the extreme case where a feature is practically inactionable (e.g., someone's family history) then the weights of those features could be set to ∞ and will thus remain unchanged by the optimization problem.

4.6 Cost functions for counterfactual explanations

We evaluate the quality of the counterfactual explanations produced by our method by using the maximum percentile shift (f_1) from [29] and two novel cost functions which we call symmetric maximum log percentile shift (f_2), and the minimax percentile shift (f_3). In closed form, these cost functions are given by:

$$f_1(x, x') = \max_{i \in [n]} \left| Q_i(x'_i) - Q_i(x_i) \right| \tag{8}$$

$$f_2(x, x') = \sum_{i \in [n]} \left| \log\left(\frac{1 - Q_i(x'_i)}{1 - Q_i(x_i)} \right) \right| + \left| \log\left(\frac{Q_i(x'_i)}{Q_i(x_i)} \right) \right| \tag{9}$$

$$f_3(x') = \min_{x \in S} \max_{i \in [n]} \left| Q_i(x'_i) - Q_i(x_i) \right|. \tag{10}$$

Where Q_i is the percentile function of the ith feature.

These loss functions all exhibit a set of desirable statistical properties that make them suitable for evaluating the quality of counterfactual explanations. In particular

- They all scale invariant, meaning that the values of these cost functions are unaffected by a change in of units in the features (i.e., using pounds vs. kilograms to measure weight).

- f_1 measures the greatest percentage change by which the features change relative to the original instance. From a qualitative standpoint a value of $f_1(x, x') = 0.1$ signifies that we would need to change all of the features of a counterfactual explanation by at most 10%.

- f_2 is a symmetrized version of the log percentile shift function introduced in [29]. It increases exponentially as $Q_j(x_j) \to 1$ and $Q_j(x_j) \to 0$, which reflects the fact that changes are harder when starting off at higher or lower percentile values (i.e., at the tails of the distribution). In other words, a change from percentiles $90 \to 95$ or $10 \to 5$ is harder than a change from percentiles $50 \to 55$. Moreover, points which are pushed towards the tails of the empirical distribution will be heavily penalized (i.e., a change from percentiles $90 \to 95$ gets penalized more heavily than a change from percentiles $90 \to 80$). Therefore, high values of this cost likely correspond to outliers and thus relatively poor counterfactual explanations. For a more detailed explanation, see [29].

- f_3 measures the greatest percentile shift of a feature relative to some instance of the dataset. In this case, a value of $f_3(x') = 0.1$ would mean that we would need to change all of the features of a counterfactual by at least 10% to look like some instance that is already in the dataset. Thus, high values of this cost function are an indication that the resulting counterfactual is an outlier relative to the original dataset. Moreover, for all of the observed datapoints note that $f_3(x) = 0$.

5 Algorithmic recourse through counterfactuals

5.1 Toy example

To demonstrate our approach we simulate a separable 2-D dataset consisting of a mixture of two Gaussians. We run three simulations to qualitatively asses the effects of enforcing the plausibility and correlation constraints discussed in Sections 3 and 4. The results of these experiments are shown in Figures 1 and **??**.

5.2 Diabetes dataset

To illustrate the usefulness of our approach, we apply SVM-ACE to the Pima Indians Diabetes dataset [25]. It has features like the number of pregnancies the patient has had, blood pressure, BMI, insulin levels, etc. that are used to predict whether a patient is at risk of diabetes. Our goal is to suggest robust perturbabtions to the features of high-risk patients to better understand the predictions of our model.

We present the results of our approach in Tables 1 and asses the quality of our counterfactual explanations as compared to other methods in Section 5.3. From a qualitative standpoint, our results are favorable. In particular, note that the changes for the three patients shown in Table 1 indicate that our model indicates that patients should decrease their glucose levels and BMI while slightly increasing the amount of insulin units. This provides a very clear and robust way of interpreting the predictions of our model. Moreover, the changes suggested to flip the predictions indicate that BMI drop to the widely accepted healthy range of between 19 and 25. Similarly, a decrease in glucose concentration to below the diabetic threshold of 140 mg/dL on most patients.

Table 1: Suggested counterfactuals (CF) on actionable features for a patient to decrease the risk of Type II diabetes

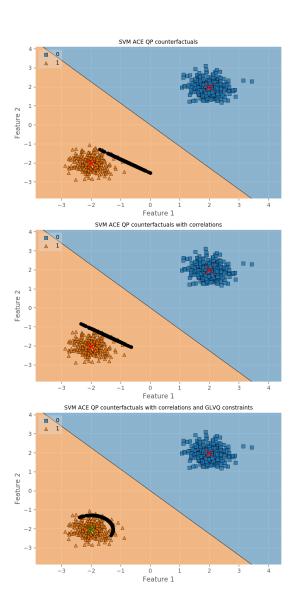| FEATURE	ORIGINAL	CF
PATIENT I		
GLUCOSE (MG/DL)	171	135
DIASTOLIC BP (MM HG)	72	75
INSULIN (MUU/ML)	135	140
BMI	33	25
PATIENT II		
GLUCOSE (MG/DL)	155	110
DIASTOLIC BP (MM HG)	62	66
INSULIN (MUU/ML)	495	501
BMI	34	24
PATIENT III		
GLUCOSE (MG/DL)	160	115
DIASTOLIC BP (MM HG)	54	58
INSULIN (MUU/ML)	175	181
BMI	30	21

Figure 1: Example of the resulting counterfactuals (shown as black dots) of all of the examples with a label of $y = 0$. (Top) Illustration of the solutions to the original optimization problem of equation 4 (Middle) Illustration of the solutions to the optimization problem using the Manhalanobbis distance to account for correlations. (Bottom) Illustration of the solutions to the optimization problem using the Manhalanobbis distance and the prototype constraints of equation 7.

5.3 Comparison to other methods

We compare the results of our approach with two methods from the literature used to compute counterfactual explanations. In particular, we compare our approach to DiCE [18] and the nearest support vector approach from [34].The results of our experiments are shown in Figures **??** and **??** and Table 2.

Table 2: Average cost of counterfactual explanations (CF) on actionable features for a diabetes patients

METHOD	SVM-ACE MEAN VALUE OF CF
MAX PERCENTILE SHIFT	
SVM-ACE	46.85
SVM-ACE WITH CORR	**46.68**
NEAREST SV	**38.88**
SVM-ACE SPARSE + CORR	65.46
DICE	78.21
SVM-ACE WITH PLAUSIBILITY	46.83
LOG PERCENTILE SHIFT	
SVM-ACE	6.65
SVM-ACE WITH CORR	6.96
NEAREST SV	4.42
SVM-ACE SPARSE + CORR	**2.95**
DICE	12.35
SVM-ACE WITH PLAUSIBILITY	7.04
MINIMAX PERCENTILE SHIFT	
SVM-ACE	24.13
SVM-ACE WITH CORR	24.26
SVM-ACE SPARSE + CORR	**23.33**
DICE	25.23
SVM-ACE WITH PLAUSIBILITY	24.35

6 Uncovering a model's bias

Counterfactual explanations have been applied in the machine learning literature as a way of enhancing interpretability (see [31]). In this section, we use our methods from Sections 3 and 4 to uncover the inherent biases of support vector machine models. As a case study, we use the reduced version of the LSAT dataset, to predict whether law students will pass the bar, based on undergraduate GPA, LSAT scores, and race as a sensitive attribute. We compute counterfactual explanations as solutions to equation (6), for all of the students that were predicted to fail the bar exam according to our SVM model. We report the mean changes of these counterfactual explanations for 884 students in Table 3. From a qualitative standpoint, we observe that the counterfactual explanations suggest higher GPAs and LSAT scores to flip the prediction of outcomes. At the same time we see that, on average, there were around 6.2% more candidates whose prediction would have been changed had the "Race (white)" attribute been set to true. These results suggest an inherent bias towards a protected feature (i.e., race) for our SVM —in agreement with previous interpretability experiments from the literature on this dataset [32].

We compare our approach with SHAP (SHapley Additive exPlanations, [17]) —a standard, widely used, approach for interpretability of machine learning models. The shap values depict the feature importance of the SVM model and are shown in Table 4. When comparing the relative importance of race as a feature in making negative predictions, we see that the features Race (Black) and Race (Hispanic) were the most negatively affected by the predictions. These results are in qualitative agreement with the results of our approach shown in Table 3.

Table 3: Mean of the suggested change in features for 884 law students predicted to fail the bar exam. The support vector machine suggested that the law students change the value of the race attribute to white 6.2% more examples relative to the other race features. The changes for categorical values are expressed as relative percentages. For continuous features, the changes are expressed in absolute terms.

COUNTERFACTUAL EXPLANATIONS	
FEATURE	SVM-ACE MEAN CHANGE
GPA	0.57
LSAT SCORE	24.9
RACE (WHITE)	6.2 %
RACE (BLACK)	(4.8)%
RACE (HISPANIC)	(1.6)%
RACE (OTHER)	(0.0)%
RACE (ASIAN)	(0.2)%

Table 4: Mean SHAP values for 884 students predicted to fail the bar exam.

SHAP VALUES	
FEATURE	SHAP VALUE
GPA	0.13
LSAT SCORE	0.28
RACE (WHITE)	(0.04)
RACE (BLACK)	(0.14)
RACE (HISPANIC)	(0.08)
RACE (OTHER)	(0.01)
RACE (ASIAN)	(0.03)

7 Conclusion and Discussion

The task of computing actionable counterfactual explanations for discriminative models has been relatively unexplored in modern machine learning research. We introduce a simple, yet effective approach for using Support

Vector Machines to compute counterfactual explanations using mixed integer programming. These methods compute counterfactuals that are valid, actionable, plausible, stable, and take into account correlations between features.

We also introduce two novel cost functions to evaluate the quality of counterfactual explanations. These cost functions are scale invariant and attain low values for counterfactual explanations that have favorable statistical properties relative to the data distribution. Moreover, we find that our approach outperforms other standard approaches for computing counterfactual explanations from the literature.

Using the counterfactual explanations, we explored the interpretability of SVM models on the LSAT dataset. In particular, we used our approach to uncover the inherent biases of the SVM model. We find that our results are both consistent with the interpretability literature, and qualitatively agree with the results produced by other interpretability methods from the literature.

To illustrate the usefulness of this approach we conducted experiments on a diabetes dataset. Our goal was to suggest actionable changes to the features of high-risk patients to decrease their risk of diabetes. When applying our approach in this context, we must note that the counterfactual recommendations are only as good as the SVM model itself. Even then, this model is limited by the fact that it fails to account for causal relationships between features. As such, this approach should always be paired with domain knowledge from experts and one should be wary about drawing causal conclusions using this method. We view the adaptation of this method for causal inference as an interesting direction for future work.

References

[1] G. Adomavicius and A. Tuzhilin. Discovery of actionable patterns in databases: The action hierarchy approach. In *KDD*, pages 111–114, 1997.

[2] André Artelt and Barbara Hammer. On the computation of counterfactual explanations - A survey. *CoRR*, abs/1911.07749, 2019.

[3] André Artelt and Barbara Hammer. Convex optimization for actionable & plausible counterfactual explanations. *CoRR*, abs/2105.07630, 2021.

[4] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[5] Y. Elovici and D. Braha. A decision-theoretic approach to data mining. *IEEE Transactions on Systems, Man, and Cybernetics, Part A*, 33(1):42–51, 2003.

[6] Y. Elovici, B. Shapira, and P. B. Kantor. Using the information structure model to compare profile-based information filtering systems. *Inf. Retr.*, 6(1):75–97, 2003.

[7] D. Gamberger and N. Lavrač. Generating actionable knowledge by expert-guided subgroup discovery. In *PKDD'02: Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery*, pages 163–174. Springer-Verlag, 2002.

[8] Y. Jiang, K. Wang, A. Tuzhilin, and A. W.-C. Fu. Mining patterns that respond to actions. In *ICDM*, pages 669–672, 2005.

[9] Amir-Hossein Karimi, Gilles Barthe, Borja Balle, and Isabel Valera. Model-agnostic counterfactual explanations for consequential decisions. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 895–905. PMLR, 26–28 Aug 2020.

[10] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: From counterfactual explanations to interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 353–362, New York, NY, USA, 2021. Association for Computing Machinery.

[11] Eoin M. Kenny and Mark T. Keane. On generating plausible counterfactual and semi-factual explanations for deep learning. In *AAAI*, 2021.

[12] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, X. Renard, and Marcin Detyniecki. Inverse classification for comparison-based interpretability in machine learning. *ArXiv*, abs/1712.08443, 2017.

[13] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. Comparison-based inverse classification for interpretability in machine learning. In Jesús Medina, Manuel Ojeda-Aciego, José Luis Verdegay, David A. Pelta, Inma P. Cabrera, Bernadette Bouchon-Meunier, and Ronald R. Yager, editors, *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Foundations*, pages 100–111, Cham, 2018. Springer International Publishing.

[14] N . Lavrač, D. Gamberger, and P. Flach. Subgroup discovery for actionable knowledge generation: deficiencies of classification rule learning and the lesson learned. In N. Lavrac, T. Fawcett, and H. Motoda, editors, *ICML 2002 workshop on Data Mining Lessons Learned*, 2002.

[15] B. Liu, W. Hsu, and Y. Ma. Identifying non-actionable association rules. In *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 329–334, New York, NY, USA, 2001. ACM Press.

[16] B. Liu, M. Hu, and W. Hsu. Multi-level organization and summarization of the discovered rules. In *Knowledge Discovery and Data Mining*, pages 208–217, 2000.

[17] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[18] Ramaravind Kommiya Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. *CoRR*, abs/1905.07697, 2019.

[19] G. Piatetsky-Shapiro and C. Matheus. The interestingness of deviations. In *AAAI Workshop on Knowledge Discovery in Databases*, pages 25–36, Menlo Park, CA, 1994. AAAI Press.

[20] Z. W. Ras and S. Gupta. Global action rules in distributed knowledge systems. *Fundam. Inf.*, 51(1):175–184, 2002.

[21] Z. W. Ras and A. A. Tzacheva. Discovering semantic inconsistencies to improve action rules mining. In *IIS*, pages 301–310, 2003.

[22] Z. W. Ras, A. A. Tzacheva, L.-S. Tsay, and O. Gurdal. Mining for Interesting Action Rules. In *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Intelligent Agent Technology (IAT'05)*, pages 187–193. IEEE, 2005.

[23] Kaivalya Rawal, Ece Kamar, and Himabindu Lakkaraju. Can i still trust you?: Understanding the impact of distribution shifts on algorithmic recourses. *ArXiv*, abs/2012.11788, 2020.

[24] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning - From Theory to Algorithms*. Cambridge University Press, 2014.

[25] Jack Smith, J. Everhart, W. Dickson, W. Knowler, and Richard Johannes. Using the adap learning algorithm to forcast the onset of diabetes mellitus. *Proceedings - Annual Symposium on Computer Applications in Medical Care*, 10, 11 1988.

[26] R. Trepos, A. Salleb, M.-O. Cordier, V. Masson, and C. Gascuel. A distance-based approach for action recommendation. In *ECML 2005, 16th European Conference on Machine Learning*, LNAI 3720, pages 425–436. Springer-Verlag Berlin Heidelberg, 2005.

[27] R. Trepos, A. Salleb-Aouissi, M-O. Cordier, V. Masson, and C. Gascuel-Odoux. Building actions from classification rules. *Knowledge and Information Systems*, 34:267–298, 2011.

[28] L.-S. Tsay and Z. W. Ras. Action rules discovery: system DEAR2, method and experiments. *Journal of Experimental & Theoretical Artificial Intelligence*, 17(1-2):119–128, 2005.

[29] Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 10–19, New York, NY, USA, 2019. Association for Computing Machinery.

[30] V.N. Vapnik. *Statistical Learning Theory*. A Wiley-Interscience publication. Wiley, 1998.

[31] Sahil Verma, John P. Dickerson, and Keegan Hines. Counterfactual explanations for machine learning: A review. *CoRR*, abs/2010.10596, 2020.

[32] Sandra Wachter, Brent D. Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *CoRR*, abs/1711.00399, 2017.

[33] Q. Yang, J. Yin, C. X. Ling, and T. Chen. Postprocessing decision trees to extract actionable knowledge. In *ICDM*, pages 685–688, 2003.

[34] Qiang Yang and Hong Cheng. Mining case bases for action recommendation. In *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, pages 522–529, 2002.

Supplementary material

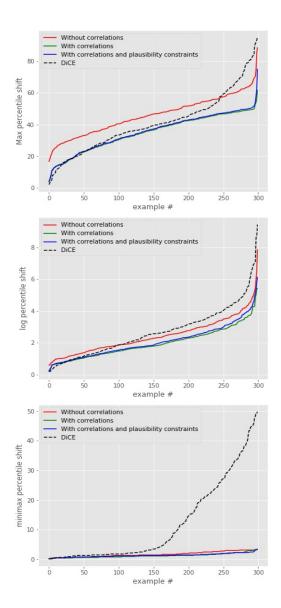## 1   Appendix: Value of counterfactuals on individual examples.

Figure 1: Example of quality of counterfactuals on the toy dataset according to the cost functions of equations 8, 9, and 10.
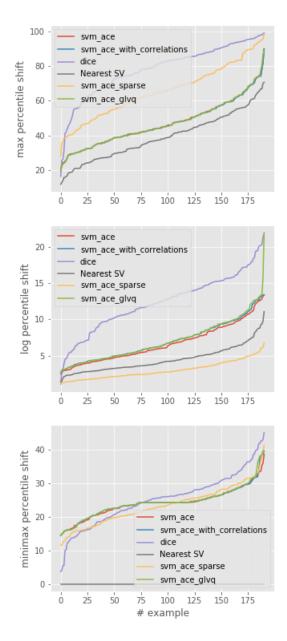
Figure 2: Example of quality of counterfactuals on the diabetes dataset according to the cost functions of equations 8, 9, and 10.