Zhang, Linli; Karakasidis, Georgios; Odnoblyudova, Arina; Dogruel, Leyla; Tian, Yu; Jung, Alex

**Explainable empirical risk minimization**

*Published in:*
Neural Computing and Applications

*DOI:*
10.1007/s00521-023-09269-3

*Published: 01/03/2024*

*Document Version:*
Publisher's PDF, also known as Version of record

*Published under the following license:*
CC BY

*Please cite the original version:*
Zhang, L., Karakasidis, G., Odnoblyudova, A., Dogruel, L., Tian, Y., & Jung, A. (2024). Explainable empirical risk minimization. *Neural Computing and Applications*, 36(8), 3983-3996. https://doi.org/10.1007/s00521-023-09269-3
Explainable empirical risk minimization

Linli Zhang1,2 · Georgios Karakasidis1 · Arina Odnoblyudova1 · Leyla Dogruel3 · Yu Tian1 · Alex Jung1

Received: 5 March 2023 / Accepted: 6 November 2023 / Published online: 8 December 2023 © The Author(s) 2023

Abstract

The successful application of machine learning (ML) methods increasingly depends on their interpretability or explainability. Designing explainable ML (XML) systems is instrumental for ensuring transparency of automated decision-making that targets humans. The explainability of ML methods is also an essential ingredient for trustworthy artificial intelligence. A key challenge in ensuring explainability is its dependence on the specific human end user of an ML system. The users of ML methods might have vastly different background knowledge about ML principles, with some having formal training in the specific field and others having none. We use information-theoretic concepts to develop a novel measure for the subjective explainability of predictions delivered by a ML method. We construct this measure via the conditional entropy of predictions, given the user signal. Our approach allows for a wide range of user signals, ranging from responses to surveys to biophysical measurements. We use this measure of subjective explainability as a regularizer for model training. The resulting explainable empirical risk minimization (EERM) principle strives to balance subjective explainability and risk. The EERM principle is flexible and can be combined with arbitrary ML models. We present several practical implementations of EERM for linear models and decision trees. Numerical experiments demonstrate the application of EERM to weather prediction and detecting inappropriate language in social media.

Keywords

Explainable machine learning · Subjective explainability · Empirical risk minimization

1 Introduction

A key requirement for trustworthy artificial intelligence (AI) and machine learning (ML) is its transparency and explainability [1]. However, there seems to be no widely accepted measure for the explainability of an ML method [2, 3]. Informally, explaining a ML method amounts to communicating some information that is relevant to the understanding of its working principles [4]. In a supervised ML setting, such information might be obtained from (local) approximations of the hypothesis map learnt by ML methods [5].

Building on our recent work [6], we interpret the act of explaining as a communication problem with the goal to effectively convey the operational principles of the ML method to the intended users. To quantify the explainability, we propose using entropy as a measure, which reflects the (lack of) uncertainty about the predictions delivered by the ML method.

A key challenge for XML is the variation in the background knowledge of human end users [6, 7]. ML methods
that are explainable for a domain expert might be opaque (“black-box”) for a lay user. For example, a deep neural network used for diagnosing skin cancer from images may be explained by quantifying the influence of individual pixels on the prediction, which could be visualized as a heat map [8, 9]. However, such an explanation may not be sufficient for a lay person without expertise in dermatology [9].

To ensure the user-specific (or personalized) explainability of ML, we introduce the concept of a user (feedback) signal. User signals are conceptually similar to the concept of labelled data in supervised ML. While labels represent quantities of interest associated with data points, user signals reflect how data points are comprehended by the human user. The main contribution of this paper is a XML method whose explainability is tailored to a specific user [1, 4].

Our approach is agnostic to the means of acquiring user signals, which could include social network behaviour, biophysical measurements, observation of facial expressions, or manually chosen interpretable features of data points. More generally, user signals might be obtained from interpretable representations of data points [5]. Section 4 considers hate speech detection in social media where data points represent short messages (“tweets”). Here, the user signal could be the presence of specific (key-) words that the user considers an indicator of hate speech [5].

This paper proposes explainable empirical risk minimization (EERM) as a novel XML method that can be applied to a wide range of models. We obtain different instances of EERM, such as explainable linear regression (see Sect. 3.1) and explainable decision tree (DT) classification (see Sect. 3.2). EERM requires a training set with known user signals which are used to estimate the subjective explainability of the hypothesis.

EERM learns a hypothesis whose predictions do not deviate too much across data points that are considered similar by a human user. In particular, if a user assigns identical user signals to data points, then predictions delivered by EERM should be close. This requirement is similar in spirit to the smoothness assumption of supervised ML, which requires similar predictions for data points with similar features [10].

EERM is an instance of structural risk minimization [11, Ch. 7] that uses the subjective explainability of a hypothesis as a regularization term [8, 12, 13]. Depending on the quality of the user signal, enforcing the subjective explainability of a learnt hypothesis might be beneficial or detrimental to the resulting prediction accuracy (see Sect. 3.1). If the user signal is correlated with the true label of a data point, our requirement for explainability helps to steer or regularize the learning task. Indeed, we might interpret the user feedback signal as a manifestation of domain expertise and, in turn, the explainability requirement as a means to incorporate domain expertise into a ML method.

1.1 Related work

Existing XML methods can be roughly divided into two main flavours: model-agnostic and intrinsically explainable (or interpretable) [2, 3, 14]. Model-agnostic methods construct explanations for any given ML method while the latter category restricts the choice of ML models to “simple” models. This paper proposes EERM as a novel XML method that bridges these two flavours. EERM shares the flexibility of model-agnostic XML in allowing for using arbitrary models. However, EERM does not construct explanations but learns a hypothesis that is intrinsically explainable to a specific user.

Model-agnostic methods can be combined with any ML method for which it is possible to efficiently compute predictions for data points. These methods do not require the details of the ML method such as the optimization algorithms used for model training. Instead, they only need to be provided the predictions for some probing points. These predictions are then used to construct a local approximation of the overall behaviour of the ML method [5, 6]. Maybe the most basic example for a XML method from this category is to locally approximate a learnt hypothesis by a linear function [5]. Instead of a local (linear) approximation of a ML method, EERM uses regularization to nudge ERM to become more explainable.

A second main category of XML methods is obtained by restricting the design choice for the ML model to a distinct set of intrinsically explainable (“simple”) models. Examples of intrinsically explainable models include linear models using few features and shallow DTs [15]. The explanation (or interpretation) of a trained linear model is typically constructed from the learned weights for the individual features. A large (in magnitude) weight indicates a high relevance of the corresponding feature for the resulting predictions. The prediction delivered by a DT might be explained by the path traversed from the root node to the decision node during the computation of the prediction [16].

In general, however, there is no widely accepted definition of which model is considered intrinsically explainable. Moreover, there is no consensus about how to measure the explainability of a “simple” model [3]. We close this gap by introducing a novel measure for the subjective explainability of a hypothesis. This measure is derived from the conditional entropy of the predictions obtained by applying the hypothesis to a random data point, given its user signal. Conditional entropy is closely related to the concept of mutual information [17], which has been used previously to quantify the effect of
explanations [6, 18]. While the XML methods of [6, 18] construct explanations for a given ML method, we train a given model such that its predictions are intrinsically explainable (without the need for additional explanations) to a specific user.

1.2 Contributions

We next enumerate the main contributions of this paper.

- We apply information-theoretic concepts to develop a novel measure for the subjective explainability of predictions delivered by a ML method.
- Our main conceptual contribution is to identify the notion of subjective explainability with predictability: A hypothesis is considered explainable to a specific user if its predictions can be anticipated by the user. The extent of anticipation is measured by the conditional entropy of the predictions, conditioned on the user signal (see Sect. 2.2).
- Our main methodological contribution is to propose the EERM principle, which is obtained from the ERM principle by adding subjective explainability as a regularizer (see Sect. 3).
- We present an overall performance assessment $E^*$ measure and illustrate the usefulness of EERM by two numeric experiments in real datasets (see Sect. 4).

1.3 Notation

Throughout the paper, the scalars are represented with normal lowercase letters while the vectors are denoted in bold. We represent different topological spaces with calligraphic font and use the symbol $\mathcal{X}$ to denote the estimation of a variable $x$.

2 Problem setup

We consider a generic ML setup that involves data points characterized by a label (quantity of interest) $y \in \mathcal{Y}$ and some features (attributes) $x = (x_1, \ldots, x_n)^T \in \mathbb{R}^n \in \mathcal{X}$ [11, 16, 19]. The XML method developed in Sect. 3 does not place any restrictions on the choice for feature space $\mathcal{X}$ and label space $\mathcal{Y}$. Each data point is also assigned a user signal $u \in \mathcal{U}$, which we will use to construct measures for the subjective explainability of ML methods.

The proposed XML method (see Sect. 3) can be combined with different sources for user signals. User signals might be obtained from facial expressions or biophysical measurements [20]. Another example for a user signal is manually chosen features that are considered interpretable [5]. Section 4.2 studies hate speech detection in social media with data points being short messages. Here, the user signal is defined via the presence of certain keywords that are considered a strong indicator of hate speech.

Our key assumption is that a hypothesis is subjectively explainable for a user if it delivers similar predictions for data points with similar user signals. To summarize, we represent a data point by a triplet $(x, y, u)$ that consists of a feature vector $x$, a label $y$, and a user signal $u$.

The goal of many important ML methods is to learn a hypothesis map [11]

$$ h(\cdot) : \mathcal{X} \rightarrow \mathcal{Y} : x \mapsto \hat{y} = h(x). \quad (1) $$

that is used to compute the predicted label $\hat{y} = h(x)$ solely from the features $x$ of a data point. Given finite computational resources can only use a subset of (computationally) feasible maps. We refer to this subset as the hypothesis space (model) $\mathcal{H}$ of a ML method. The XML method presented in Sect. 3 allows for a wide range of choices for the model, including linear models, DTs, and artificial neural networks (ANNs) [16, 19, 21]. The main requirement for the model $\mathcal{H}$ is merely that it allows for efficient training (optimization) algorithms. Examples of such models which typically include linear maps, DTs, or ANNs [21–23].

For a given data point with features $x$ and label $y$, we measure the quality of a hypothesis $h$ using some loss function $L((x, y), h)$. The quantity $L((x, y), h)$ measures the error incurred by predicting the label $y$ of a data point using the prediction $\hat{y} = h(x)$. Popular examples for loss functions are the squared error loss $L((x, y), h) = (h(x) - y)^2$ (for numeric labels $y \in \mathbb{R}$) or the logistic loss $L((x, y), h) = \log(1 + \exp(-h(x)y))$ (for label space $\mathcal{Y} = \{-1, 1\}$).

Roughly speaking, we would like to learn a hypothesis $h$ that incurs a small loss on any data point. To make this informal goal precise, we can use the notion of expected loss or risk

$$ \mathbb{E}\{L((x, y), h)\} := \mathbb{E}\{L((x, y), h)\}. \quad (2) $$

This definition tacitly uses an i.i.d. assumption where data points are interpreted as realizations of statistically independent RVs with a common pdf $p(x, y, u)$ which underlies the expectation in (2)). Ideally, we would like to learn a hypothesis $h$ with minimum risk

$$ \mathbb{E}\{L((x, y), \hat{h})\} = \min_{h \in \mathcal{H}} \mathbb{E}\{L((x, y), h)\}. \quad (3) $$

The risk minimization principle (3) is impractical as we typically do not know the probability distribution $p(x, y)$ required for evaluating the risk (2).
Section 2.1 discusses how empirical risk minimization (ERM) is obtained by approximating the risk using an average loss over some training set [11, 16, 21]. In its basic form, ERM-based methods are prone to overfitting and require some form of regularization [11, Ch. 6]. We regularize ERM by a measure for the subjective explainability of a hypothesis \( h(x) \). Section 2.2 explains how this regularization term is obtained from the conditional entropy of the predictions \( h(x) \), given the user signal \( u \). The resulting EERM principle is then discussed in Sect. 3.

### 2.1 Empirical risk minimization

ERM methods approximate the risk (2) by the average loss (or empirical risk)

\[
\tilde{L}(h|\mathcal{D}) := \frac{1}{m} \sum_{i=1}^{m} L\left( (x^{(i)}, y^{(i)}), h \right).
\]

The average loss \( \tilde{L}(h|\mathcal{D}) \) of the hypothesis \( h \) is measured on a set of labelled data points (the training set)

\[
\mathcal{D} = \{ (x^{(1)}, y^{(1)}, u^{(1)}), \ldots, (x^{(m)}, y^{(m)}, u^{(m)}) \}.
\]

The training set \( \mathcal{D} \) consists of data points characterized by features \( x^{(i)} \), label value \( y^{(i)} \), and user signal \( u^{(i)} \), for \( i = 1, \ldots, m \).

A plethora of ML methods are based on solving the ERM problem

\[
\hat{h} \in \arg \min_{h \in \mathcal{H}} \tilde{L}(h|\mathcal{D})
\]

using different choices for model \( \mathcal{H} \), training data \( \mathcal{D} \) and loss \( L \). However, a direct implementation of ERM (6) is prone to overfitting if the hypothesis space \( \mathcal{H} \) is too large compared to the size \( m \) of the training set. Two examples of such a high-dimensional regime are linear regression with a large number of features and artificial neural network (“deep learning”) using a large number of trainable parameters.

One of the most widely used techniques to avoid overfitting in this high-dimensional regime is regularization [24, 25]. There are many different implementations or regularization techniques, such as data augmentation or early stopping [11, 21]. In what follows, we regularize ERM by adding a regularization (or penalty) term \( \lambda \mathcal{R}(h) \) to the empirical risk in (6),

\[
h^{(\lambda)} \in \arg \min_{h \in \mathcal{H}} \tilde{L}(h|\mathcal{D}) + \lambda \mathcal{R}(h).
\]

The choice of the regularization parameter \( \lambda \geq 0 \) in (7) can be guided either by probabilistic models for the data [11, Ch. 7] or validation techniques [16]. For linear models \( h(x) = w^T x \), two popular choices for the regularizer are \( \mathcal{R}(h) = ||w||_2 \) (“ridge regression”) and \( \mathcal{R}(h) = ||w||_1 \) (“Lasso”) [11, 23].

A dual form of regularized ERM (7) is obtained by replacing the regularization term with a constraint,

\[
h^{(\eta)}(\mathcal{H}|\mathcal{D}) \text{ such that } \mathcal{R}(h) \leq \eta.
\]

The solutions of (8) coincide with those of (7) for an appropriate choice of \( \eta \) [26]. Solving the primal formulation (7) might be computationally more convenient as it is an unconstrained optimization problem in contrast to the dual formulation (8) [27]. On the other hand, the dual form (8) allows explicitly specifying an upper bound \( \eta \) on the value \( \mathcal{R}(h^{(\eta)}) \) for the learned hypothesis \( h^{(\eta)} \).

Regularization techniques are typically used to improve the statistical performance (risk) of the learned hypothesis. Instead, we use regularization to ensure the subjective explainability of the learned hypothesis. We use a regularization term that is not primarily meant to estimate the generalization error \( \mathbb{E}\{L((x,y), \hat{h})\} - \mathcal{L}(h|\mathcal{D}) \). but to measure the subjective explainability of the predictions \( \hat{y} = \hat{h}(x) \). The regularization parameter \( \lambda \) in (7) (or \( \eta \) in the dual formulation (8)) adjusts the level of subjective explainability of the learned hypothesis \( \hat{h} \). Larger values of \( \lambda \) (smaller values of \( \eta \)) favour a hypothesis with high explainability at the cost of incurring a higher risk. Section 3.1 studies the trade-off between subjective explainability and risk in linear regression using a simple probabilistic model for data.

### 2.2 Subjective explainability

There seems to be no widely accepted formal definition for the explainability (or interpretability) of a ML method. Some authors refer to ML methods as intrinsically interpretable if they use specific design choices for the model [4, 6, 14]. We believe that a useful concept of interpretability can only be subjective, i.e. depending on the specific human user of a ML method. Indeed, while linear regression methods might be considered interpretable for a user with formal training in statistics, the predictions obtained by applying a linear hypothesis to a huge number of features might be difficult to grasp for a layman.

The key idea of this paper is to construct a measure for subjective explainability of a hypothesis \( h \in \mathcal{H} \) via the user signal \( u \) associated with each data point. We consider this hypothesis as subjectively explainable if it delivers similar predictions \( \hat{y} = \hat{h}(x) \) for data points having similar user signals. Informally, the hypothesis is subjectively explainable to a user if

\[
h(x^{(1)}) \approx h(x^{(2)}) \text{ for data points with } u^{(1)} \approx u^{(2)}.
\]

Similar to [18], we use information-theoretic concepts to make the informal notion (9) of subjective explainability
The features realizations of i.i.d. random variables. In particular, the features \( x \), label \( y \), and user signal \( u \) associated with a data point are realizations drawn from a joint probability density function (pdf) \( p(x, y, u) \). In general, the joint pdf \( p(x,y,u) \) is unknown and needs to be estimated from data using, e.g. maximum likelihood methods [19, 23].

Note that since we model the features of a data point as the realization of a RV, the prediction \( \hat{y} = h(x) \) also becomes the realization of a RV. Figure 1 summarizes the overall probabilistic model for data points, the user signal, and the predictions delivered by (the hypothesis learned with) a ML method.

We measure the subjective explainability of the predictions \( \hat{y} \) delivered by a hypothesis \( h \) for some data points \((x, y, u)\) as,

\[
E(h|u) := C - H(h|u). \tag{10}
\]

Here, we used the conditional (differential) entropy \( H(h|u) \) (see Ch. 2 and Ch. 8 [17])

\[
H(h|u) := -\mathbb{E}_{x,y} \left\{ \log p(h(x) | u) \right\}. \tag{11}
\]

We introduce the ("calibration") constant \( C \) in (10) for notational convenience. The actual value of \( C \) is meaningless for our approach (see Sect. 3) and serves only the convention that the subjective explainability \( E(h|u) \) is non-negative.

For regression problems, the predicted label \( \hat{y} \) might be modelled as a continuous random variable. In this case, the quantity \( H(\hat{y}|u) \) is a conditional differential entropy. With slight abuse of notation we refer to \( H(\hat{y}|u) \) as a conditional entropy and do not explicitly distinguish between the case where \( \hat{y} \) is discrete, such as in classification problems studied in Sects. 3.1, 3.2 and 4.

The conditional entropy \( H(h|u) \) in (10) quantifies the uncertainty (of a user that assigns the value \( u \) to a data point) about the prediction \( \hat{y} = h(x) \) delivered by the hypothesis \( h \). Smaller values \( H(h|u) \) correspond to smaller levels of subjective uncertainty about the predictions \( \hat{y} = h(x) \) for a data point with known user signal \( u \). This, in turn, corresponds to a larger value \( E(h|u) \) of subjective explainability.

Section 4 discusses explainable methods for detecting hate speech or the use of offensive language. A data point represents a short text message (a tweet). Here, the user signal \( u \) could be the presence of specific keywords that are considered a strong indicator of hate speech or offensive language. These keywords might be provided by the user via answering a survey or they might be determined by computing word histograms on public datasets that have been manually labelled [28].

### 3 Explainable empirical risk minimization

Section 2 has introduced all the components of EERM as a novel principle for XML. EERM learns a hypothesis \( h \) by using an estimate \( \hat{H}(h|u) \) for the conditional entropy in (10) as the regularization term \( R(h) \) in (7),

\[
h^{(x)} := \arg\min_{h \in \mathcal{H}} \hat{L}(h|\mathcal{D}) + \lambda \hat{H}(h|u). \tag{12}
\]

A dual form of (12) is obtained by specializing (8),

\[
h^{(\eta)} := \arg\min_{h \in \mathcal{H}} \hat{L}(h|\mathcal{D}) \text{ such that } \hat{H}(h|u) \leq \eta. \tag{13}
\]

The empirical risk \( \hat{L}(h|\mathcal{D}) \) and the regularizer \( \hat{H}(h|u) \) are computed solely from the available training set (5). We will discuss specific choices for the estimator \( \hat{H}(\hat{y}|u) \) in Sects. 3.1 and 3.2.

The idea of EERM is that the solution of (12) (or (13)) is a hypothesis that balances the requirement of a small loss (accuracy) with a sufficient level of subjective explainability \( E(h|u) \) (\( = C - H(h|u) \)). This balance is steered by the parameter \( \lambda \) in (12) and \( \eta \) in (13), respectively. Figure 2 illustrates the parametrized solutions of (12) in the plane spanned by risk and subjective explainability. The precise location of the curve in Fig. 3 depends on the training set.

---

**Fig. 1** The features \( x \), label \( y \) and user signal \( u \) of a data point are realizations drawn from a pdf \( p(x,y,u) \). Our goal is to learn a hypothesis \( h \) such that its predictions \( \hat{y} \) have a small conditional entropy given the user signal \( u \).

**Fig. 2** The solutions of EERM, either in the primal (12) or dual (13) form, trace out a curve in the plane spanned by the risk \( \mathcal{T}(h) \) and the subjective explainability \( E(h|u) \).
We now specialize EERM in its primal form \((12)\) to linear regression.\[ \text{3.1 Explainable linear regression} \]

Choosing a large value for \(\lambda\) in \((12)\) (small value for \(\eta\) in \((13)\)) penalizes any hypothesis resulting in a large estimate \(H(h|u)\) for the conditional entropy \(H(h|u)\) of \((10)\). Assuming \(\bar{H}(h|u) \approx H(h|u)\), using a large \(\lambda\) in \((12)\) (small \(\eta\) in \((13)\)) enforces a high subjective explainability \((10)\) of the learned hypothesis \(h^{(\lambda)}\). Asymptotically (for \(\lambda \to \infty\)), the solutions \(h^{(\lambda)}\) of \((12)\) will maximize subjective explainability \(E(h|u)\) at the cost of increasing risk.

For the specific choice \(\lambda = 0\), EERM \((12)\) reduces to plain ERM that delivers a hypothesis \(h^{(\lambda=0)}\) with risk \(L_{\min}\). This special case of EERM is obtained from the dual form \((8)\) using a sufficiently large \(\eta\). The small risk of \(h^{(\lambda=0)}\) comes at the cost of a relatively small subjective explainability \(E(h^{(\lambda=0)}|u)\). We choose the constant \(C\) in \((10)\) such that \(E(h^{(\lambda=0)}|u) = 0\) for notational convenience.

We emphasize that we do not assume to have any control over the choice of user signals. Much like the process of determining the features of a data point also the construction of user signals is beyond the scope of this paper. In particular, we do not have any control over the correlation between user signals and the labels of data points. If the user signal is nearly uncorrelated with the label, requiring subjective explainability will typically result in a hypothesis with a large risk. On the other hand, if the user signal is strongly correlated with the label of a data point (e.g. when the user is a domain expert), then EERM \((12)\) might learn a hypothesis with a smaller risk compared to the hypothesis learnt by plain ERM \((6)\).

### 3.1 Explainable linear regression

We now specialize EERM in its primal form \((12)\) to linear regression \([19, 23]\). Linear regression methods learn the parameters \(w\) of a linear hypothesis \(h^{(w)}(x) = \mathbf{w}^T x\) to minimize the squared error loss of the resulting prediction error. The features \(x\) and user signal \(u\) of a data point are modelled realizations of jointly Gaussian random variables with mean zero and covariance matrix \(C\).

\[
(\mathbf{x}^T, u)^T \sim \mathcal{N}(\mathbf{0}, C) \quad .
\]

Note that \((14)\) only specifies the marginal of the joint pdf \(p(x, y, u)\) (see Fig. 1). Using the probabilistic model \((14)\), we obtain (see \([17]\))

\[
H(h|u) = (1/2) \log \sigma_{hu}^2 \quad .
\]

Here, we use the conditional variance \(\sigma_{hu}^2\) of the predicted label \(\hat{y} = h(x)\) given the user signal \(u\) for a data point.

To develop an estimator \(\bar{H}(h|u)\) for \((15)\), we use the identity \([29, \text{Sec. } 4.6]\)

\[
\sigma_{hu}^2 = \min_{x \in \mathbb{R}} \mathbb{E}\left\{ (h(x) - xu)^2 \right\} \quad .
\]

The identity \((16)\) relates the conditional variance \(\sigma_{hu}^2\) to the minimum mean squared error that can be achieved by estimating \(\hat{y}\) using a linear estimator \(xu\) with some \(x \in \mathbb{R}\). We obtain an estimator for the conditional variance \(\sigma_{hu}^2\) by replacing the expectation in \((16)\) with a sample average over the training set \(\mathcal{D}\)

\[
\bar{\sigma}_{hu}^2 = \min_{x \in \mathbb{R}} \frac{1}{m} \sum_{i=1}^{m} (\mathbf{w}^T \mathbf{x}^{(i)} - xu^{(i)})^2 \quad .
\]

It seems reasonable to estimate the conditional entropy \(\bar{H}(h^{(w)}|u)\) via the plugging in the estimated conditional variance \((17)\) into \((15)\), yielding the plug-in estimator \((1/2)\). However, in view of the duality between \((8)\) and \((13)\), any monotonic increasing function of a given entropy estimator essentially amounts to a reparametrization \(\lambda \to \lambda'\) and \(\eta \to \eta'\). Since such a reparametrization is irrelevant as we choose \(\lambda\) in a data-driven fashion, we will use the estimated conditional variance \((17)\) itself as an estimator

\[
\bar{H}(h^{(w)}|u) := \min_{x \in \mathbb{R}} \frac{1}{m} \sum_{i=1}^{m} (\mathbf{w}^T \mathbf{x}^{(i)} - xu^{(i)})^2 \quad .
\]

Note that we neither require the estimator \((18)\) to be consistent nor to be unbiased \([30]\). Our main requirement is that, with high probability, the estimator \((18)\) varies monotonically with the conditional entropy \(H(h^{(w)}|u)\).

Inserting the estimator \((18)\) into EERM \((12)\) yields Algorithm 1 as an instance of EERM for linear regression in primal form. Algorithm 1 requires as input a choice for the regularization parameter \(\lambda > 0\) and a training set

\[
\mathcal{D} = \{(x^{(1)}, y^{(1)}, u^{(1)})\ldots, (x^{(m)}, y^{(m)}, u^{(m)})\}
\]
Algorithm 1 delivers a hypothesis $h^{(2)}$ that compromises between small risk $\mathbb{E}\{L((x,y),h)\}$ and subjective explainability $E(h|u)$. This compromise is controlled by the value of $\lambda$.

Choosing a large $\lambda$ for Algorithm 1 favours a hypothesis $h^{(2)}$ with small conditional entropy $H(h^{(2)}|u)$ and, in turn, high subjective explainability $E(h^{(2)}|u)$ (see (10)). On the contrary, choosing a small $\lambda$ puts more emphasis on obtaining a small risk $\mathbb{E}\{L((x,y),h^{(2)})\}$ at the expense of increased conditional entropy $H(h^{(2)}|u)$ and, in turn, reduced subjective explainability $E(h^{(2)}|u)$.

**Algorithm 1** Explainable Linear Regression in primal form

\begin{align*}
\text{Input:} & \text{ explainability parameter } \lambda, \text{ training set } D \text{ (see (5))} \\
1: & \text{solve} \\
\hat{w} & \in \arg\min_{w \in \mathbb{R}^n} \sum_{i=1}^{m} \left( y^{(i)} - w^T x^{(i)} \right)^2 \\
& \quad \text{subject to } w^T x^{(i)} - \alpha u^{(i)} \leq \lambda \\
& \quad \text{as empirical risk} \\
& \quad \text{subjective explainability} \\
\text{Output:} & h^{(\lambda)}(x) := x^T \hat{w}
\end{align*}

**Trade-Off Between Subjective Explainability and Risk.** Let us now study the fundamental trade-off between subjective explainability $E(h|u)$ and risk of a linear hypothesis for data points characterized by a single feature $x$. We consider data points $(x,u,y)^T$, characterized by a single feature $x \in \mathbb{R}$, numeric label $y \in \mathbb{R}$ and user feedback $u \in \mathbb{R}$, as i.i.d. realizations of a Gaussian random vector

$$(x,y,u)^T \sim \mathcal{N}(\mu, C) \text{ with } \mu = \begin{pmatrix} \mu_x \\ \mu_y \\ \mu_u \end{pmatrix},$$

$$C = \begin{pmatrix} 
\sigma_x^2 & \sigma_{x,y} & \sigma_{x,u} \\
\sigma_{x,y} & \sigma_y^2 & \sigma_{y,u} \\
\sigma_{x,u} & \sigma_{y,u} & \sigma_u^2
\end{pmatrix}. \quad (20)$$

Our goal is to learn a linear hypothesis $h(x) = x^T w$ which is parametrized by a weight vector $w = (w_1 \ w_0)^T$, where $x = (x \ 1)^T$.

Let us require a minimum prescribed subjective explainability $E(h|u) \geq C - \eta$, which is equivalent to the constraint (see (10) and (15))

\begin{equation}
H(h|u) = (1/2) \log \sigma_{y,u}^2 \leq \eta. \quad (21)
\end{equation}

We minimize the risk (24) under the constraint (23), which is equivalent to enforcing subjective explainability of at least $C - \eta$,

$$\min_{w_1,w_0 \in \mathbb{R}} P_x w_1^2 + w_0^2 - 2\mu_x w_0 + P_y + 2\mu_x w_1 w_0 - 2P_{x,y} w_1$$

subject to $w_1 \leq \exp(2\eta)\sigma_{y,u}^{-2}$.

$$\text{(26)}$$

A set of necessary and sufficient conditions for a weight vector $\hat{w} = (\hat{w}_1, \hat{w}_0)^T$ to solve (26) are the Karush–Kuhn–Tucker conditions [27, Sec. 5.5.3.]

We can further develop the constraint (21) using (20) and basic calculus for Gaussian processes [31],

\begin{equation}
\sigma_{y,u}^2 = \sigma_y^2 - \sigma_{x,u}^2/\sigma_u^2
\end{equation}

and

\begin{equation}
\hat{y} = x^T \hat{w}
\end{equation}

subject to

\begin{equation}
w_1 \leq \exp(2\eta)\sigma_{y,u}^{-2} \quad (23)
\end{equation}

The goal is to find a linear hypothesis $h(x) = w^T x$, whose weight vector $w$ satisfies (23), that incurs a minimum risk

\begin{equation}
\mathbb{E}\{L((x,y),h)\} = \mathbb{E}\{(y - h(x))^2\}
\end{equation}

where

\begin{equation}
\mathbb{E}\{L((x,y),h)\} = \mathbb{E}\{(y - w_1 x - w_0)^2\}
\end{equation}

subject to $w_1 \leq \exp(2\eta)\sigma_{y,u}^{-2}$.

The goal is to find a linear hypothesis $h(x) = w^T x$, whose weight vector $w$ satisfies (23), that incurs a minimum risk

\begin{equation}
\mathbb{E}\{L((x,y),h)\} = \mathbb{E}\{(y - h(x))^2\}
\end{equation}

where

\begin{equation}
\mathbb{E}\{L((x,y),h)\} = \mathbb{E}\{(y - w_1 x - w_0)^2\}
\end{equation}

subject to $w_1 \leq \exp(2\eta)\sigma_{y,u}^{-2}$.

The goal is to find a linear hypothesis $h(x) = w^T x$, whose weight vector $w$ satisfies (23), that incurs a minimum risk

\begin{equation}
\mathbb{E}\{L((x,y),h)\} = \mathbb{E}\{(y - h(x))^2\}
\end{equation}

subject to $w_1 \leq \exp(2\eta)\sigma_{y,u}^{-2}$.
Algorithm 2: Explainable Linear Regression in dual form

Input: upper bound of conditional entropy \( \eta \), training set \( \mathcal{D} \) (see (5))

1: solve

\[
\hat{w} = (\bar{w}_1, \bar{w}_0)^T \in \arg\min_{w_1, w_0 \in \mathbb{R}} \left\{ \begin{array}{l}
\frac{\mu_{yx} - \mu_x}{\mu_x} w_1 - 2 \frac{\mu_{xy}}{\mu_y} w_0 + P_u \\
\text{subject to} \\
\bar{w}_1^2 \leq \exp(2\eta) \sigma_{y|xu}^{-2} \\
\text{upper bound of subjective explainability}
\end{array} \right. \\
\text{empirical risk}
\]

Output: \( h^{(\lambda)}(x) := x^T \hat{w} \)

By inserting (28), (29) into (24), we obtain that the minimum achievable risk \( \mathbb{E}\{L((x,y), h)\} \) of a linear hypothesis with required subjective explainability \( E(h|u) \geq C - \eta \) is

\[
\mathbb{E}\{L((x,y), h)\} = \begin{cases} 
\sigma_{y|x}^2 & \text{if } \sigma_{y|x}^2/\sigma_x^4 \leq \exp(2\eta) \sigma_{y|xu}^{-2} \\
\phi(\eta) & \text{if } \sigma_{y|x}^2/\sigma_x^4 > \exp(2\eta) \sigma_{y|xu}^{-2},
\end{cases}
\]

where \( \phi(\eta) = \sigma_x^2 \exp(2\eta) \sigma_{y|xu}^{-2} - 2|\sigma_{y|x}| \exp(\eta) / \sigma_{y|xu} + \sigma_x^2. \)

Inserting the estimator (28), (29) into EERM (13) yields Algorithm 2 as an instance of EERM for linear regression in dual form. Algorithm 2 requires as input a choice for the upper bound of conditional entropy \( \eta \) and a training set \( \mathcal{D} = \{(x^{(1)}, y^{(1)}, u^{(1)}), \ldots, (x^{(m)}, y^{(m)}, u^{(m)})\} \). Algorithm 2 delivers a hypothesis \( h^{(\lambda)} \) that achieves a minimum empirical risk \( \mathcal{L}(h) \) under the corresponding upper bound of conditional entropy \( \eta \), i.e., maximum subjective explainability \( E(h|u) \). The optimal model is controlled by the value of \( \eta \).

Choosing a large \( \eta \) for Algorithm 2 means the hypothesis \( h^{(\lambda)} \) has a large margin of conditional entropy \( H(h^{(\lambda)}|u) \) and, in turn, allows lower subjective explainability \( E(h^{(\lambda)}|u) \) and lower empirical risk \( \mathcal{L}(h) \) (see (21), (30)). On the contrary, choosing a small \( \eta \) puts more restriction on obtaining a lower conditional entropy \( H(h^{(\lambda)}|u) \) at the expense of increased empirical risk \( \mathcal{L}(h) \) and, in turn, increased subjective explainability \( E(h^{(\lambda)}|u) \).

3.2 Explainable decision trees

We now apply EERM in its dual (constraint) form (13) to DT classifiers [19, 23]. Consider data points characterized by features \( x \), a binary label \( y \in \{0,1\} \) and a binary user signal \( u \in \{0,1\} \). The restriction to binary labels and user signals is for ease of exposition. Our approach can be generalized easily to more than two label values (multiclass classification) and non-binary user signals.

The model \( \mathcal{H} \) in (13) is constituted by all DTs whose root node tests the user signal \( u \) and whose depth does not exceed a prescribed maximum depth \( d_{\text{max}} \) [16]. The depth \( d \) of a specific DT \( h \) is the maximum number of test nodes that are encountered along any possible path from the root node to a leaf node [16].
Figure 4 illustrates a hypothesis $h$ obtained from a DT with depth $d = 2$. We consider only DTs whose nodes implement a binary test, such as whether a specific feature $x_j$ exceeds some threshold. Each such binary test can maximally contribute one bit to the entropy of the resulting prediction (at some leaf node).

Thus, for a given user signal $u$, the conditional entropy of the prediction $\hat{y} = h(x)$ is upper bounded by $d - 1$ bits. Indeed, since the root node is reserved for testing the user signal $u$, the number of binary tests carried out for computing the prediction is upper bounded by $d - 1$. We then obtain Algorithm 3 from (13) by using the estimator $\hat{H}(h|u) := d - 1$.

Algorithm 3 Explainable DT Classification

Input: subjective explainability $\eta$, training set $D$ (5)

1: learn a DT classifier $\hat{g}(x)$ for user signal based on training set $D^{(u=0)}$
2: set maximum tree-depth $d_{\text{max}} := \lceil \eta \rceil$
3: partition dataset $D = D^{(u=0)} \cup D^{(u=1)}$ with

$$D^{(u=0)} := \{ (x^{(i)}, y^{(i)}, u^{(i)}): u^{(i)} = 0 \} \subseteq D$$
$$D^{(u=1)} := \{ (x^{(i)}, y^{(i)}, u^{(i)}): u^{(i)} = 1 \} \subseteq D$$

4: learn decision tree classifier $h^{(u=0)}$ with maximum depth $d_{\text{max}}$ using training set $D^{(u=0)}$
5: learn decision tree classifier $h^{(u=1)}$ with maximum depth $d_{\text{max}}$ using training set $D^{(u=1)}$
6 combine the separate DT classifiers to obtain the classifier

$$h^{(\eta)}(x, u) := \begin{cases} 
  h^{(u=1)}(x) & \text{if } u = 1 \\
  h^{(u=0)}(x) & \text{if } u = 0
\end{cases}$$

Output: hypothesis $h^{(\eta)}(x, \hat{u})$
4 Numerical experiments

This section reports the results of different numerical experiments to verify the performance of EERM. Besides, we present a new metric to evaluate the overall performance of EERM, which combines different preferences between objective accuracy and subjective explainability. It details the various choices evaluated against end users, evaluation metrics and results obtained.

A. Dataset

In order to demonstrate the usefulness of EERM (12), we conduct illustrative numeric experiments that revolve around explainable weather forecasting (see Sect. 4.2) and explainable hate speech detection in social media (see Sect. 4.1). The weather prediction dataset is obtained at the observation station “Nuuksio” in Finland from 2020-1-1 to 2021-12-31 from the Finnish Meteorological Institute (FMI). The hate speech detection dataset consists of 25.3k texts on Twitter data within hate speech, offensive language, and neither classifications from Kaggle.2

Section 4.2 discusses the application of EERM to the detection of hate speech and offensive language in social networks [32]. Hate speech is a main obstacle towards embracing the Internet’s potential for deliberation and freedom of speech [33]. Moreover, the detrimental effect of hate speech seems to have been amplified during the period of the COVID-19 pandemic [34].

Hate speech is a contested term whose meaning ranges from concrete threats to individuals to venting anger against authority [35]. Hate speech is characterized by devaluing individuals based on group-defining characteristics such as their race, ethnicity, religion, and sexual orientation [36]. Detecting hate speech requires multi-disciplinary expertise from social sciences and computer science [37, 38]. Providing subjective explainability for ML users with different backgrounds is crucial for the diagnosis and improvement of hate speech detection systems [33, 34, 39].

B. Evaluation Metrics

Since the quality of a ML model with subjective explainability is highly dependent on the domain knowledge of the end users, as well as its inherence factors which are contextual and subjective, there is a lack of consensus regarding unified metrics to effectively assess it.

To this end, we define a new comprehensive quality metric $E^*$ measure to evaluate the overall performance of a ML model with subjective explainability, i.e. the combination between the interpretability (higher subjective explainability for various levels of end users) and the fidelity (lower empirical risk). $E^*$ computes their harmonic mean and takes the effects of both empirical risk and conditional entropy into account simultaneously. It thus symmetrically represents both measures in one metric.

$$E^* = \frac{2(1 - \bar{L})(1 - \bar{H})}{2 - \bar{L} - \bar{H}},$$

(32)

where $\bar{L}$ and $\bar{H}$ denote the normalization of the empirical risk $L(h)$ and the conditional entropy $H(h|u)$ for a given human end user, respectively.

Referring to $F1$ score in statistics, $E^*$ is an indicator designed for considering two important evaluation criteria, empirical risk and conditional entropy. It combines them into a single measure to provide a balanced assessment of the model’s subjective explainability ranging between 0.0 and 1.0. The satisfying (highest) possible value of a $E^*$ is 1.0, if both empirical risk and conditional entropy are perfect, while the lowest value is 0.0, indicating poor performance, i.e. worst empirical risk or conditional entropy. For public policymakers, they should consider the different background knowledge levels of the end users. So it is a proper choice to give priority to high subjective explainability for normal people, while it might be more important to ensure that the model has a high accuracy for the research by domain-specific experts. However, high accuracy might cause low subjective explainability, which means low empirical risk but high conditional entropy. Since both measures are interactive, $E^*$ is supposed to be large enough under a certain compromise but hard to reach the ideal value of 1.0 with real datasets.

4.1 Explainable linear regression

This experiment applies EERM to learn an explainable linear predictor for the maximum daytime temperature at some observation stations of the Finnish Meteorological Institute (FMI) (https://en.ilmatieteellaislaitos.fi/download-observations). Data points that represent the daily weather recordings along with a time-stamp at the weather station “Nuuksio” in Finland. The feature $x$ of a data point is the minimum temperature during that day, while the maximum temperature of the same day is the label $y$. Each data point is also characterized by a user signal $u \in \mathbb{R}$. We compute the weights $\hat{w} = (\hat{w}_1, \hat{w}_0)^T$ of a linear hypothesis by plugging in sample estimates for means and (co-)variances in (28)–(30). Thus, we construct a linear hypothesis whose subjective explainability is approximately lower bounded by a prescribed value of $C - \eta$.

In order to simulate the variation in the background knowledge of users, we generate user signals by adding different perturbations on maximum temperatures. Specifically, the perturbations are obtained by the probability...
density function of the normal distribution $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right)$, where $m$ is the mean and $\sigma$ is the standard deviation. In Figs. 5 and 6, we set $m = 0$, $\sigma = \{0, 5, 7.5, 10\}$. With the stand deviation $\sigma$ increasing, the user signal $u$ becomes worse.

Figures 5 and 6 depict, respectively, the resulting empirical risk and weights for the varying upper bound of conditional entropy $\eta$. Overall, for increasing subjective explainability $E(h^{(l)}|u)$ (decreasing $\eta$), the empirical risk $\bar{L}(h)$ increases and the weight $\hat{w}_1$ for the feature $x$ becomes smaller (less relevant). On the other hand, for decreasing explainability $E(h^{(l)}|u)$ (increasing $\eta$), the empirical risk $\bar{L}(h)$ decreases and the weight $\hat{w}_1$ for the feature $x$ becomes larger (more relevant). Besides, it shows the larger the values of $\xi$ are, the larger the empirical risk is got to achieve the same subjective explainability, which means the worse the quality of the user signal is. Since the end user has less domain-specific background knowledge, the model needs a larger upper bound of conditional entropy $\eta$ to obtain the minimum empirical risk $\bar{L}(h)$. Moreover, all curves tend to be close to $\bar{L} = \sigma^2$ as $\eta \to -\infty$, i.e. the explainability is large enough. While $\eta \geq \frac{1}{2} \ln(\sigma_{y|x}^2/\sigma_x^2)$, the optimal model is reached that $\bar{L}_{\text{min}} = \sigma_{y|x}^2$.

Additionally, we compute the $E^*$ measure using (32) as the overall performance assessment of the model EERM in the weather prediction experiment. Different curves plotted in Figs. 7 and 8 illustrate the relationship between $E^*$ with the normalization of the conditional entropy $H$ and the empirical risk $\bar{L}$, respectively. We can see that EERM is underperforming if we just pursue a perfect subjective explainability (lower $H$) or get satisfying model accuracy (lower $\bar{L}$) solely. $E^*$ is supposed to be large enough under a certain compromise but hard to reach the ideal value of 1.0 in real datasets (see Fig. 9). Only by considering both fully can the $E^*$ increase to achieve a better performance of EERM (see Fig. 10).

4.2 Explainable hate speech detection

We now discuss a numerical experiment that uses EERM to learn an explainable hate speech detector for social networks. This experiment uses a public dataset that contains curated short messages (tweets) from a social network [28]. Each tweet has been manually rated by a varying number of users as either “hate speech”, “offensive
language”, or “neither”. For each tweet, we define its binary label as $y = 1$ (“inappropriate tweet”) if the majority of users rated the tweet either as “hate speech” or “offensive language”. If the majority of users rated the tweet as “neither”, we define its label value as $y = 0$ (“appropriate tweet”).

The feature vector $x$ of a tweet is constructed using the normalized frequencies (“tf-idf”) of individual words (stop words removed) [40]. Each tweet is also characterized by a binary user signal $u \in \{0, 1\}$. The user signal is defined to be $u = 1$ if the tweet contains at least one of the five most frequent words appearing in tweets with $y = 1$.

We use Algorithm 3 to learn an explainable decision tree classifier with its subjective explainability upper bounded by $\eta = 2$ bits. The training set $D$ used for Algorithm 3 is obtained by randomly selecting a fraction of around 90% of the entire dataset. The remaining 10% of tweets are used as a test set.

To learn the decision tree classifiers in step 3 and 4 of Algorithm 3, we used the implementations provided by the current version of the Python package scikit-learn [22]. The resulting explainable DT classifier $h^{(\eta=2)}(x)$ achieved a test-set accuracy of 0.929.

5 Conclusion

The explainability of predictions provided by ML becomes increasingly relevant for their use in automated decision-making [41, 42]. Given lay and expert users’ different levels of expertise and knowledge, providing subjective (tailored) explainability is instrumental for achieving trustworthy AI [41, 43]. Our main contribution is EERM as a new design principle for subjective XML. EERM is obtained using the conditional entropy of predictions, given
a user signal, as a regularizer. The hypothesis learned by EERM balances between small risk and subjective explainability for a specific user (explainee) of the ML method.

5.1 Future works

Though our method is flexible and agnostic to the precise means of obtaining user signals, interesting avenues for future work include user studies that allow measuring different forms of user signals, using feedback on explanations, and combining with counterfactuals. Besides, we will develop our work further in practical implementations of the EERM principle for more complex ML models like artificial neural networks (ANNs).

Author Contributions LZ performed the numerical experiments, participated in the theoretical analysis, and was a major contributor to writing the manuscript. GK contributed to the writing of the manuscript and analysis of results. AO contributed to experiment design and manuscript writing. LD contributed to the literature review and manuscript writing. YT was involved in the data and experiment results interpretation and the manuscript revision. AJ proposed the design and conception of the whole research and was a major contributor to writing the manuscript.

Funding Open Access funding provided by Aalto University. The work in this paper is supported by the following three fundings: - funding “Ensemble nowcasting of irradiance/clouds for solar energy using novel machine learning tools—can AI beat physics?” with project number 885377 granted by Austrian Research Promotion Agency (FFG) in 2021, funding “Intelligent Techniques in Condition Monitoring of Electromechanical Energy Conversion Systems” granted by Academy of Finland (decision number 331197) in 2020, and funding “XAI-based software-defined energy networks via packetized management for fossil fuel-free next-generation of industrial cyber-physical systems (X-SDEN)” granted by Academy of Finland (decision number 349966) in 2022.

Data availability Observation station data of the Finnish Meteorological Institute (FMI): https://en.ilmatieteenlaitos.fi/download-observations. Public dataset contains curated short messages (tweets) from a social network: https://dl.acm.org/doi/10.1145/2063576.2063726.

Code Availability The code used for the current study is available from the corresponding author on reasonable request.

Declarations

Conflict of interest There are no conflicts of interest or competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. High-Level Expert Group on AI (2019) Ethics guidelines for trustworthy AI, Technical report, European Commission
2. Linardatos P, Papastefanopoulos V, Kotsiantis S (2021) Explainable AI: a review of machine learning interpretability methods. Entropy 23(1):18. https://doi.org/10.3390/e23010018
3. Zhou J, Gandomi AH, Chen F, Holzinger A (2021) Evaluating the quality of machine learning explanations: a survey on methods and metrics. Electronics 10(5):593. https://doi.org/10.3390/electronics10050593
4. ISO (2020) Information technology—artificial intelligence—overview of trustworthiness in artificial intelligence, vol. ISO/IEC TR 24028:2020(E), 1st edn. ISO/IEC
5. Ribeiro MT, Singh S, Guestrin C (2016) “Why should i trust you?”: explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD, pp 1135–1144
6. Jung A, Nardelli PH (2020) An information-theoretic approach to personalized explainable machine learning. IEEE Signal Process Lett 27:825–829
7. Belle V, Papantonis I (2021) Principles and practice of explainable machine learning. Front Big Data. https://doi.org/10.3389/fdata.2021.688969
8. Bach S, Binder A, Montavon G, Klauschen F, Müller K-R, Samek W (2015) On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE 10(7):1–46
9. Ayhan MS, Kümmerle LB, Kühlewein L, Inhoffen W, Aliyeva G, Ziemssen F, Berens P (2022) Clinical validation of saliency maps for understanding deep neural networks in ophthalmology. Med Image Anal 77:102364. https://doi.org/10.1016/j.media.2022.102364
10. Chapelle O, Schölkopf B, Zien A (eds) (2006) Semi-supervised learning. The MIT Press, Cambridge
11. Jung A (2022) Machine learning: the basics. Springer, HHH, Cham
12. Montavon G, Samek W, Müller K (2018) Methods for interpreting and understanding deep neural networks. Digit Signal Process 73:1–15
13. Hagras H (2018) Toward human-understandable, explainable AI. Computer 51(9):28–36
14. Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1(5):206–215. https://doi.org/10.1038/s42256-019-0048-x
15. Molnar C (2019) Interpretable machine learning: a guide for making black box models explainable. [online] Available: https://christophm.github.io/interpretable-ml-book/
16. Hastie T, Tibshirani R, Friedman J (2001) The elements of statistical learning. Springer series in statistics. Springer, New York
17. Cover TM, Thomas JA (2006) Elements of information theory, 2nd edn. Wiley, Hoboken
18. Chen J, Song L, Wainwright MJ, Jordan MI (2018) Learning to explain: an information-theoretic perspective on model interpretation. In: Proceedings of the 35th International conference on machine learning, Stockholm, Sweden
19. Bishop CM (2006) Pattern recognition and machine learning. Springer, Cham
20. Zhang Y, Ji Q (2005) Active and dynamic information fusion for facial expression understanding from image sequences. IEEE Trans Pattern Anal Mach Intell 27(5):699–714
21. Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press, Cambridge
22. Pedregosa F (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12(85):2825–2830
23. Hastie T, Tibshirani R, Wainwright M (2015) Statistical learning with sparsity: the lasso and its generalizations. CRC Press, Boca Raton
24. Bühlmann P, van de Geer S (2011) Statistics for high-dimensional data. Springer, New York
25. Wainwright M (2019) High-dimensional statistics: a non-asymptotic viewpoint. Cambridge University Press, Cambridge
26. Bertsekas DP (1999) Nonlinear programming, 2nd edn. Athena Scientific, Belmont
27. Boyd S, Vandenberghe L (2004) Convex optimization. Cambridge University Press, Cambridge
28. Davidson T, Warmsley D, Macy M, Weber I (2017) Automated hate speech detection and the problem of offensive language. In: Proceedings of the 11th international AAAI conference on web and social media (ICWSM), vol 11, no 1, pp 512–515
29. Bertsekas DP, Tsitsiklis JN (2008) Introduction to probability, 2nd edn. Athena Scientific, Belmont
30. Lehmann EL, Casella G (1998) Theory of point estimation, 2nd edn. Springer, New York
31. Rasmussen CE, Williams CKI (2006) Gaussian processes for machine learning. MIT Press, Cambridge
32. Wang X, Wei F, Liu X, Zhou M, Zhang M (2011) Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. In: Proceedings of the 20th ACM international conference on information and knowledge management. CIKM ’11. Association for Computing Machinery, New York, NY, USA, pp 1031–1040. https://doi.org/10.1145/2063576.2063726
33. Laaksonen SM, Haapoja J, Kinnunen T, Nelimarkka M, Pöyhätäri R (2020) The datafication of hate: expectations and challenges in automated hate speech monitoring. Front Big Data 3:3
34. Hardage D, Peyman N (2020) Hate and toxic speech detection in the context of Covid-19 pandemic using XAI: ongoing applied research. In: Proceedings of the 1st workshop on NLP for COVID-19 (Part 2) at EMNLP 2020
35. Gagliardone I, Gal D, Alves T, Martínez G (2015) Countering online hate speech. UNESCO
36. Erjavec K, Kovačič MP (2012) You don’t understand, this is a new war! Mass Commun Soc 15(6):899–920
37. Papcunová J, Martončík M, Fedáková D, Kentoš M, Bozogáňová M, Srba I, Moro R, Pikuliak M, Šimko M, Adamkovič M (2021) Hate speech operationalization: a preliminary examination of hate speech indicators and their structure. Complex Intell Syst. https://doi.org/10.1007/s40747-021-00561-0
38. Liao QV, Gruen D, Miller S (2020) Questioning the AI: informing design practices for explainable AI user experiences. In: Proceedings of the 2020 CHI conference on human factors in computing systems. CHI ’20. Association for Computing Machinery, New York, NY, USA, pp 1–15. https://doi.org/10.1145/3313831.3376590
39. Bunde E (2021) AI-assisted and explainable hate speech detection for social media moderators: a design science approach. In: Proceedings of the 54th Hawaii international conference on systems sciences 2021
40. Baèza-Yates R, Ribeiro-Neto B (2011) Modern information retrieval. Addison Wesley, Boston
41. Rohlfling KJ, Cimiano P, Scharlau I, Matzner T, Buhl HM, Buschmeier H, Esposito E, Grimminger A, Hammer B, Hab Umbach R, Horwath I, Huüllermeier E, Kern F, Kopp S, Thommes K, Ngomba Ngomo A-C, Schulte C, Wachsmuth H, Wagner P, Wedre B (2021) Explanation as a social practice: toward a conceptual framework for the social design of AI systems. IEEE Trans Cogn Dev Syst 13(3):717–728. https://doi.org/10.1109/TCDS.2020.3044366
42. Larsson S, Heintz F (2020) Transparency in artificial intelligence. Internet Policy Rev. https://doi.org/10.1017/14763/2020.2.1469
43. Sokol K, Flach P (2020) One explanation does not fit all. KI-Künstliche Intell 34:235–250

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.