A Framework for Explanation of Machine Learning Decisions

Chris Brinton
Mosaic ATM; Leesburg, Virginia; USA
brinton@mosaicatm.com

Abstract

This paper presents two novel techniques to generate explanations of machine learning model results for use in advanced automation-human interaction. The first technique is "Explainable Principal Components Analysis," which creates a framework within a multi-dimensional problem space to support the explainability of model outputs. The second technique is the "Gray-Box Decision Characterization" approach, which probes the output of the machine learning model along the dimensions of the explainable framework. These two techniques are independent of the type of machine learning algorithm. Rather, the intent of these algorithms is to be applicable generally across any type of machine learning algorithm and any application domain of machine learning. The concept and computational steps of each technique are presented in the paper, along with results of experimental implementation and analysis.

1 Introduction

The average person now experiences the results of machine learning (ML) models on a frequent basis through online interaction with news sites that learn reader interests, retail websites that provide automated offers that are customized to individual consumer's buying habits, and credit card fraud detection that warns the card holder when a transaction occurs that is outside of their normal purchasing pattern. Many such applications of machine learning can be performed using an automated approach where human interpretation of the recommendations is not necessary. However, in many other applications, there would be great benefit if the machine learning model could provide an explanation of its output recommendation or other result.

In the realm of many decision support tools for military and other safety- or life-critical applications, it is necessary and appropriate for humans to be involved in decisions using the recommendations and guidance of computer automation and information systems. For example, a surgeon that receives a recommendation from an ML-based medical decision support system is not likely to perform a surgery on a patient based on the recommendation of a computer system alone. Lee and See [2004] provide an extensive description of the need for and methods to achieve user trust in a computer automation system, including the following guidance: "*Show the process and algorithms of the automation by revealing intermediate results in a way that is comprehensible to the operators.*"

Although some ML models can provide limited insight into and explanation of their intermediate results and model outputs, most machine learning model output is opaque. Such opacity can lead users of the technology to doubt the reliability of the information or recommendation that is provided. This lack of understanding of the technology can result in distrust, and to eventual failure of the technology to receive acceptance and use. Even if the technology does receive acceptance and operational use, a machine learning-based information system that can explain itself may allow more efficient and effective use of the technology.

Much previous research has been conducted in explaining ML model output, including [Andrews, et al., 1995; Fung, et al., 2005; Letham, et al., 2012; and Baehrens, et al., 2005]. The innovation that we describe herein is motivated by the explanation approach presented in the Baehrens, et al., [2005] work, but over more complex problem spaces.

2 Explainable Principal Components Analysis

Principal Components Analysis (PCA) [Hotelling, 1933] is a technique that is used extensively within machine learning model development for dimensionality reduction. From an information-theoretic perspective, regular PCA aggregates information contained in a high-dimensional space into a form that can represent an arbitrarily large portion of the information in the data through a lower-dimension vector representation. While this aggregation process identifies the orthogonal dimensions in the data over which the greatest explanation of the variance can be achieved, the "explanation" of the variance in PCA is maximized from a statistical perspective, but not from the perspective of understandability by a human. In fact, the basis vectors created by PCA are one of the primary sources of opacity in many practical machine learning applications. We present herein a formulation of a variant of PCA - which we refer to as Explainable Principal Components Analysis (EPCA) - that computes basis vectors of the problem space with human understandability as a primary objective.

The EPCA technique uses minimal human interaction to identify modes of variation in the input data that can be identified and labeled for use by subsequent explanation algorithms, such as the Gray Box Decision Characterization (GBDC) approach. The EPCA process is performed iteratively, creating one explainable basis vector for the

problem space at a time. After each basis vector is created, the input training samples are projected into a subspace that excludes any contribution from the previously fixed explainable basis vectors. Regular PCA is then run on the training samples in the subspace, and the human uses the results of the regular PCA to inform the design of the next explainable basis vector. The result of this process is a set of basis vectors that are labeled with their meanings in a manner that is intended to be understandable by a human observer of a model decision.

While we have not found this concept in the literature, our EPCA algorithm is motivated by the Kernel Near Principal Components Analysis work [Martin, 2002] and the LASSO Principal Components Analysis concept [Jolliffe, et al., 2003].

Suppose we are given n input training samples, $x_1, \dots, x_n \in \mathbb{R}^d$, for either a classification or regression problem. With any set of orthonormal basis vectors, $u_1, \dots, u_n \in \mathbb{R}^d$ that spans the space of the input training samples, we can represent the input sample data as a linear combination of the basis vectors.

$$x_i = p_{i,1}u_1 + p_{i,2}u_2 + \cdots + p_{i,n}u_n \quad (1)$$

In matrix form for all samples of the training data:

$$X = PU \quad (2)$$

where the *U* matrix is the orthonormal basis, the *X* matrix is the set of all input samples as row vectors, and *P* is the matrix of basis vector coefficients to reconstruct the input, *X*.

To initiate the EPCA procedure, we set the *U* matrix to be the orthonormal basis made up of the eigenvectors as output by a regular PCA process, sorted by the eigenvalues. The model designer then interprets the individual eigenvectors as weighting coefficients on each of the original features in the *x* vectors, and manually identifies the human-understandable concept or combination of features that is generally represented by one of the eigenvectors, favoring those that are associated with a larger eigenvalue. This may indicate that a single feature is the primary contributor to the eigenvector, or perhaps that a combination of a few related features are highlighted in the eigenvector. However, the PCA process likely also results in small, non-zero coefficients on many features in the original feature space, which can obscure the understandability of the eigenvector.

This step in the process requires the model designer to use domain knowledge and other techniques dependent on the unique nature of the problem space to create an interpretable basis vector. Figure 1 shows the eigenvectors generated by the first step in this process for a sample text-analytics problem. For text-analytics problems the Singular Value Decomposition (SVD) is used, rather than PCA, for performance reasons, but the EPCA techniques can be applied analogously to achieve explainable basis vectors.

To generate an explainable vector for this example, words that indicate a distinct concept would be excluded. For example, to distinguish medical doctors and facilities from dentists, the word 'dental' would be excluded.

ei	words
0.295841	patient
0.226496	care
0.156011	health
0.155184	therapy
0.152811	treatment
0.151564	medical
0.145705	dental
0.129402	hospital
0.123112	nurse
0.120978	therapist

ei	words
-0.18745	design
-0.1376	project
-0.12352	designer
-0.11958	graphic
-0.11022	web
-0.10993	business
-0.09552	account
-0.0937	sale
-0.09217	marketing
-0.08637	adobe

Figure 1. Sample words and coefficient values from the first eigenvector for a resume classification problem.

An additional technique that we have used is to identify input data samples that differ from each other in a single, understandable way. A simple linear model can be generated from at least two such input samples that exhibit a single mode of variation in the input feature space to create an explainable basis vector.

The first explainable basis vector, ϕ_1, is created by using these, or other, techniques to identify the input feature vector elements that are to be included as contributors to the first explainable basis vector. All other coefficients in ϕ_1 are set to zero, and ϕ_1 is then normalized to unit length.

To form the remaining explainable basis vectors, the EPCA procedure removes the contribution of the explainable basis vector from each of the input data samples. In linear algebraic terms, the input data samples are projected into the null space of the explainable basis vectors, or orthogonal complement, in the case of a single vector. This modified set of input data samples, X', is then analyzed using regular PCA and the same manual techniques described above to generate a second explainable basis vector.

Although the projection into the null space of the explainable basis vector could be done more directly, the first step that we use in this process is to create a new orthonormal basis that uses ϕ_1 as the first basis vector. This is done by combining ϕ_1 with the original *u* vectors, excluding the *u* vector that was used to form ϕ_1 (or an arbitrary *u* vector if ϕ_1 is not related to any *u* vectors), and then applying the Gram-Schmidt procedure [Wikipedia, 2017] to orthonormalize the basis.

The EPCA process terminates when the model designer is satisfied with the set of explainable basis vectors, or no additional explainable basis vectors can be identified. At this point, the basis matrix is relabeled as Φ, as is declared to be the final orthonormal explainable basis for the problem space.

A very important aspect of using the EPCA approach to establish the basis for explanation of machine learning model decisions is that in addition to obtaining orthogonal dimensions along which the ML decision can be parameterized for explainability, we also obtain explicit measures of the mean and variance of the input data samples along those dimensions. Thus, in the use of the EPCA basis for sensitivity analysis, we can compare the mean value of the entire set of input data along that dimension to the value

of the specific input vector applied to the machine learning model that generated the decision to be explained.

Inverting Equation (2) and replacing the U matrix with the Φ matrix:

$$P = X\Phi^T \quad (3)$$

Since the Φ matrix is orthonormal, its inverse is the same as its transpose. The mean and standard deviation are then computed along the columns of the P matrix to obtain the mean and standard deviation across the entire input data set.

If we are given a new input vector for prediction by the ML model, $z \in \mathbb{R}^d$, the coefficients for each of the explainable basis vectors for this new input vector are computed as:

$$p = z\Phi^T \quad (4)$$

Comparing the p vector from Equation (4) to the column means and standard deviations of the P matrix from Equation (3) provides an immediate, understandable characterization of the input data vector (not the decision, but the input vector). If labels are assigned to the dimensions of the EPCA output basis, textual descriptions could be assigned such as 'the eyes of this face are particularly close to each other,' or 'this flight path uses a particularly long final approach segment.' Although such characterizations of the input data are completely separate from the output of the model, they can provide value to the human model user who desires an explanation of the model decision. The next section, however, describes the method for explanation of the ML model decision itself.

2 Gray Box Decision Characterization

We refer to the second component of our innovation as Gray-Box Decision Characterization. As implied, this approach uses some knowledge of the inner-workings of the machine-learning approach, but does not make changes to the machine learning algorithm itself. Thus, the approach lies in between a black-box and a white-box approach. It is important to note that we refer to this approach as 'decision' characterization, not 'model' characterization. The objective of this technique is to provide an explanation for a single specific output of the machine learning model (at a time), not to provide an explainable characterization of the entire machine learning model's behavior.

The GBDC approach utilizes the results of the EPCA algorithm (or regular PCA if sufficiently explainable) as an orthogonal basis for sensitivity analysis of the output of the machine learning model around the input data vector for a single decision output. This technique simply performs a sensitivity analysis of the behavior of the model in the region of the space around the specific input data feature vector that generated the decision from the machine learning model. The GBDC approach searches for changes along explainable basis vectors that result in a change in the output of the machine learning model. Although this technique is simple in principle, the large number of dimensions (even after dimensionality reduction), and the need to search along each dimension, require additional enhancements beyond the simple concept explanation provided so far. We return to the EPCA approach to describe the additions to the GBDC algorithm.

In addition to using the mean value and standard deviation along each explainable basis vector of the Φ matrix to characterize the input, we also use the standard deviation along each dimension to determine the appropriate step size to use in the sensitivity analysis. For example, if the standard deviation, σ_j, of the coefficients of the input data samples along a particular explainable basis vector, j, is 10, then testing the sensitivity along that dimension by evaluating a change of 30 along that dimension would move the sensitivity test position by 3σ, which would likely be outside of the range of nearly all input samples.

For characterization of a classification machine learning model, the GBDC technique conducts a search to find a change in the output classification of the model. A binary search is used to find the first occurrence of change along that dimension, recognizing that no change may occur at all within the realm of reasonable change values.

Mathematically, this is represented in Equation (5), where $p_{\phi,0}$ represents the mean value of the p coefficient from the input sample data for the ϕ basis vector, y_0 being the label output by the ML model for the test input vector, z, and y being the output of the ML model for the senstivity analysis over coefficients, p:

$$\phi^* = \arg\min_p \frac{|p_\phi - p_{\phi,0}|}{\sigma_\phi}, y \neq y_0, \quad (5)$$

We represent the value of the coefficient that satisfies Equation (5) as p^*_ϕ. Then, the ML model decision can be explained as having the most significant contribution in the ϕ^* explainable basis vector, and that the decision, y_0, was output by the model because the value along the ϕ^* explainable basis vector is less than (or greater than) p^*_ϕ.

For a regression model, the rate of change of the output given a change in the input is calculated.

$$\phi^* = \arg\max_p \frac{\left|\partial y / \partial p_\phi\right|}{\sigma_\phi} \quad (6)$$

In both Equations (5) and (6), additional basis vectors can be selected for use in the explanation if the change in z-score or the partial derivative of the output is below or above a threshold value.

Thus, the GBDC approach provides an explanation of the machine learning model decision by selecting the dimensions that generate the most significant change in the machine learning model output for a regression problem, or that create a change in the classification decision with the smallest change (according to z-score) in the input vector.

3 Results

Experiments have been conducted to study the usefulness and applicability of these two concepts for generating explanations of machine learning model decisions. The research team has implemented the EPCA and GBDC algorithms in prototype software. Initial testing and evaluation of the algorithm has been performed using multiple domains.

In Figure *2*, which shows aircraft arrival tracks for flights entering the Atlanta terminal area from the northwest and landing on Runway 8L at Atlanta Hartsfield International

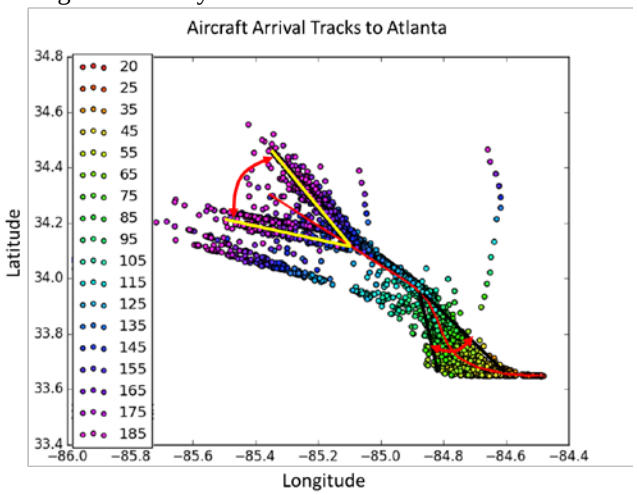

Figure 2. Human-Annotated Explainable Modes of Variation

Airport, some modes of variation are clearly evident through visual analysis of the scatter plot. The points of the scatter plot show individual surveillance positions for 246 flights on a single day. The legend indicates the altitude (in 100s of feet) associated with each color. The figure also shows two primary modes of variation of the data (one in yellow lines and one in black lines) that are clear and understandable to human observation.

In Figure 3 we show the results of both regular PCA and EPCA on this dataset. The top two plots show the first two eigenvectors of the cross-correlation matrix (i.e., regular PCA). Note that the two modes of variation determined by

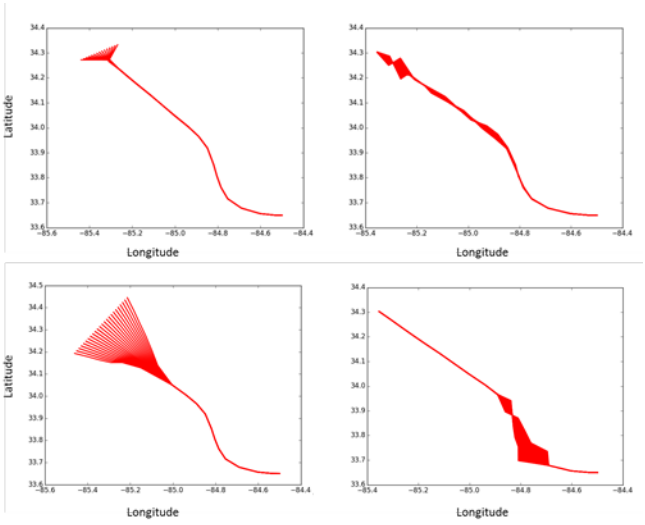

Figure 3. Two Primary Modes of Variation from Regular PCA (top) and EPCA (bottom)

PCA do not closely resemble the two primary modes that can be identified through visual inspection and understood by a human. Using EPCA, two modes of variation were generated that closely resemble the human-annoted modes of variation from Figure *2*.

Finally, we have conducted experiments with the GBDC technique as applied to the Iris dataset. In this case, the input feature vector representation is only 4-dimensional, so we are able to use the four features directly, without applying the EPCA approach first. Table 1 provides sample results that use GBDC to explain the reasons for the model output of one particular data point input to different models. All models gave the same classification, but GBDC found different explanations for each of those classifications.

Table 1. GBDC Results on the Iris Dataset

Model Type	Explanation
Decision Tree	petal width (cm) is greater than 0.79
SVM with Linear Kernel	petal length (cm) is greater than 2.1
SVM with RBF Kernel	sepal length (cm) is greater than 4.3 and petal length (cm) is greater than 2.4
SVM with Polynomial Kernel	petal length (cm) is greater than 1.7

Each of the models learns decision boundaries using different mathematical formulations and parameters. Although the thresholds chosen by GBDC to explain the decision are different in each case, they are still consistent with each other. It is also important to note that this example was selected because of the more significant differences in explanation than many other cases that were tested.

4 Conclusions

The Explainable Principal Components Analysis and Gray-Box Decision Characterization techniques provide a useful framework for analysis and explanation of machine learning model decisions. Human involvement in establishing the framework is required, but the required additional work is performed during design of the model. We have demonstrated the use of the techniques on multiple problem domains. However, additional research is needed to address more complicated problem domains, and to evaluate the feasibility and effectiveness of identifying explainable basis vectors in such problem domains.

Explainability cannot be fully evaluated without including human participants to evaluate the usefulness of the explanations generated. Our future work will include such considerations, as well as the question of whether or not a single set of basis vectors is appropriate for all users of the model, or if different people or types of users would actually need a different set of basis vectors to achieve adequate explanations.

References

[Lee and See, 2004] Lee, J.D., and See, K.A., "Trust in Automation: Designing for Appropriate Reliance," HUMAN FACTORS, Vol. 46, No. 1, Spring 2004, pp. 50–80.

[Andrews, et al., 1995] Andrews, R., Diederich, J., and Tickle A., "A survey and critique of techniques for extracting rules from trained artificial neural networks." Knowledge-Based Systems, 8:373–389, 1995.

[Letham, et al., 2012] Letham, B., Rudin, C., McCormick, T. H., and Madigan, D., "Building Interpretable Classifiers with Rules using Bayesian Analysis," Technical Report no. 609, Department of Statistics, University of Washington, December, 2012.

[Baehrens, et al., 2010] Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., and Muller, K., "How to Explain Individual Classification Decisions," Journal of Machine Learning Research, 2010.

[Fung, et al., 2005] Fung. G., Sandilya, S., and Rao, R. B., "Rule Extraction from Linear Support Vector Machines," Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, 2005.

[Hotelling, 1933] H. Hotelling, "Analysis of a complex of statistical variables into principal components," Journal of Educational Psychology, 24:417-441 and 498-520, 1933.

[Martin, 2002] Martin, Shawn, "Kernel Near Principal Components Analysis," Sandia National Laboratory Report SAND2001 -3769, July 2002.

[Jolliffe, et al., 2003] Jolliffe, I.T.; Trendafilov, N.T. and Uddin, M. A modified principal component technique based on the LASSO. Journal of Computational and Graphical Statistics, 12(3) pp. 531–547. 2003.

[Wikipedia, 2017] Wikipedia, "Gram-Schmidt process," https://en.wikipedia.org/wiki/Gram-Schmidt_process, accessed June 1, 2017.