How Machine Learning Contributes to Solve Acoustical Problems[1]

Marie A. Roch, Peter Gerstoft, Bozena Kostek, and Zoi-Heleni Michalopoulou

What Is Machine Learning?

Machine learning is the process of learning functional relationships between measured signals (called percepts in the artificial intelligence literature) and some output of interest. In some cases, we wish to learn very specific relationships from acoustics. Examples with direct commercial applications include selecting or recognizing music (Schedl et al., 2014) and identifying the language of a speaker (e.g., Zissman, 1996) for call center routing.

Alternatively, we may be interested in an exploratory analysis such as discovering relationships between animal-produced sounds and potential call categories that may carry signaling information (e.g., Sainburg et al., 2020). Machine learning can be used to discover information about the physical world such as determining the distance to a source based on pressure levels in a vertical line array (Niu et al., 2017) or solving inversion problems to find geoacoustic parameters of a seabed (Benson et al., 2000).

This article provides a high-level introduction to machine learning with a limited number of techniques that are explained conceptually. Most of our examples will use the vowel data of Peterson and Barney (1952). They showed that vowels could be relatively well identified by formant frequencies, harmonics of voiced speech that are amplified by resonances in the vocal tract. These data were selected because they provide an example of a real acoustics problem that can be solved in a low-dimensional space suitable for two-dimensional figures.

For readers desiring a more quantitative introduction to machine learning, we recommend the review by Bianco et al. (2019) that focuses on machine learning and acoustics or one of the many excellent book-length treatments of machine learning (e.g. Bishop, 2006; Hastie et al., 2009; Goodfellow et al., 2016).

Types of Machine Learning

Machine learning can be broadly separated into the major categories of supervised and unsupervised learning (Russell and Norvig, 2021). Other forms of machine learning exist but have not been used as extensively in acoustics, such as reinforcement learning (e.g., Shah et al., 2021; Wang et al., 2018) and so are beyond the scope of this article.

In supervised learning, the machine learning algorithm or learner, is presented with examples of what is to be learned and labels that consist of values or categories for each example. An example of this is seen in the work of Godino-Llorente and Gomez-Vilda (2004) where the goal was to learn to detect specific pathologies of the vocal folds from recordings of vowels.

In contrast, unsupervised learning attempts to learn from examples that do not have labels. Xi et al. (2004) trained probability models for individual musical recordings. Similarity between pairs of songs was measured by seeing how well each song's model scored the other. Clustering these scores separated songs by genre without the algorithm ever knowing the type of music.

Features

Regardless of the type of machine learning, all algorithms require transformation of the input data into features, a representation of the input signal that is conducive to solving the machine learning problem. Traditionally, these features are selected by experts using knowledge about the problem domain. For example, Peterson and Barney (1952) recognized that

1 For additional information on machine learning in acoustics, see the special issue of The Journal of the Acoustical Society of America *at* acousticstoday.org/JASA-machine-learning.

https://doi.org/10.1121/AT.2021.17.4.48

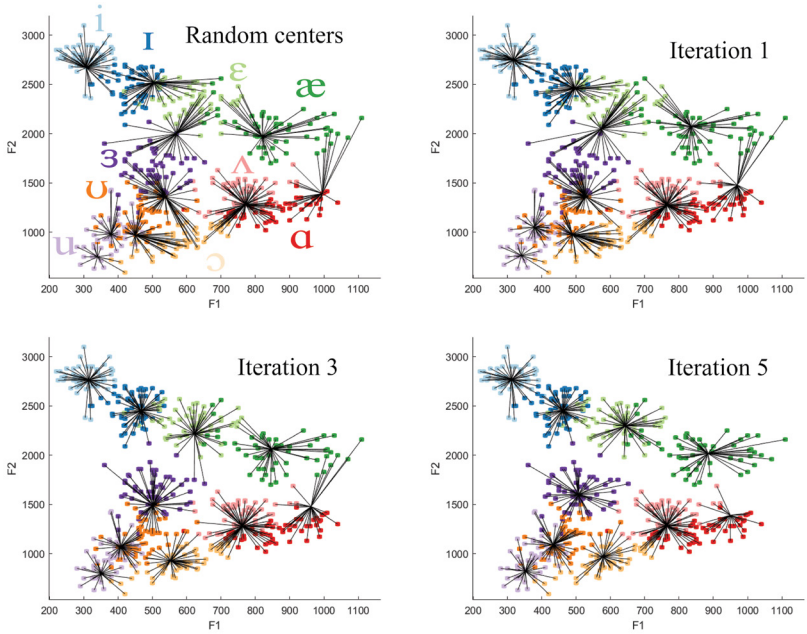


Figure 1. *Learning vowels using* k*-means. Vowel data are formant frequencies (harmonics that resonate in the vocal tract) produced by female adults (Peterson and Barney, 1952). F1, first formant frequency; F2, second formant frequency.* ***Dots*** *represent formant measurements of vowels and are color coded by vowel and labeled with international phonetic alphabet symbols. Ten representative points were chosen at random, and data were partitioned based on proximity to the closest point as shown by the* ***black lines*** *from each vowel to the closest representative point. New representative points were computed from the average of each data partition. The process was repeated until a convergence criterion was met. By the fifth iteration, most partitions contained points that were primarily from 1 of the 10 vowels.*

measurement of formant frequencies was sufficient for characterizing vowels.

In modern machine learning, there is a trend toward automated feature discovery. In many cases, the input to the model is spectrograms, at which point many of the machine learning techniques developed for image recognition become applicable. However, one does need to remember that spectrograms are not images; they represent sound that has different properties than light. For example, in images, occlusion by an object in the foreground usually prevents one from observing what is behind it. In contrast, in acoustics, two signals that overlap in time and frequency may still be recoverable if they have a strong structure such as overlapping frequency modulations with harmonics.

One promising example of an acoustics-based approach to feature learning proposed by Ravanelli and Bengio (2018) learns sets of band-pass filters that are automatically adjusted to maximize classification performance. It automatically learns the ranges of frequencies that are important to a classification problem. Other types of learned feature representations that are discussed involve finding a reduced dimension representation of the signal, a so-called manifold of the signal.

Unsupervised Learning

Unsupervised learners attempt to associate or cluster examples that are similar to one another. Although there are many different types of unsupervised learners, one of the easiest to understand is the *k*-means algorithm (Bianco et al., 2019). In this approach (**Figure 1**), one decides a priori that there are *k* different types of things in a dataset, and the goal is to find *k* representative vectors in the feature space that approximate the data. The initial *k* vectors are drawn randomly from the data. Data are partitioned based on the representative vector to which they are closest. New representative vectors are picked by averaging all of the items in each partition, and the process is repeated until a convergence criterion is

met. This technique has many applications in acoustics and is the basis for the code books that provide increased transmission capacity by transmitting a representative vector index instead of the vector.

Hierarchical Clustering

Classic methods to partition data hierarchically are top-down and bottom-up processing (Hastie et al., 2009). In top-down processing, all examples start in the same group and the group is partitioned into two subgroups in a way that maximizes their dissimilarity. In the formant data (**Figure 1**), we might select an outlier example (e.g., rightmost example of æ) and split the set of vowel data into groups that are closest to the outlier versus vowels versus those more similar to the remaining examples. This process is repeated on each group until a stopping criterion is met.

Conversely, in bottom-up processing, all elements start in their own group and the two groups that are the most similar are merged together. Returning to the formant data (**Figure 1**), we would merge points that are closest to one another in the formant space. This would be repeated until all the vowels were in a single group.

Either method produces a hierarchical tree. Branches of the tree can be assigned to partitions if desired (Hastie et al., 2009). Using these types of methods produces clusters that do not require the number of partitions to be known a priori.

Low-Dimensional Representations of Data

Manifold learning is a dimension reduction technique that may be used either as a feature extraction step or as a precursor to a clustering algorithm. For example, the spectra of vowels consist of many frequencies. Yet, as seen in **Figure 1,** the vowels can be reasonably well represented by a manifold consisting of the first two formants.

Principal components analysis (PCA; Bianco et al., 2019) is a classical method that can be used to reduce the dimensionality of feature spaces. PCA reorients the axes of the example space so that each subsequent axis accounts for less of the variance of the dataset. Because each new axis accounts for progressively less of the data's variability, some axes can be dropped and the new reduced-dimension PCA space can provide a good approximation of the dataset.

Two popular alternative approaches are t-distributed stochastic neighbor embedding (t-SNE; van der Maaten and Hinton, 2008) and uniform mapping and projection (UMAP; McInnes et al., 2018). These nonlinear methods work by matching points in a high dimensional space with an equal number of points in a low-dimensional space. Both attend to the local neighborhoods about points and attempt to align the distribution of points in the high- and low-dimensional spaces using information theoretic measures. UMAP tends to better preserve gaps between clusters.

Supervised Learning

The task of a supervised learner is to estimate a relationship based on labeled examples. In the case of regression problems, the mapping is a function, whereas classification problems partition the feature space into regions associated with categories. There are many different types of supervised classifiers, but they all do one of two things. They either learn the distribution of data or learn boundaries between different types of data.

Distributional Learners

In classification problems, distributional learners attempt to learn the class distributions from training examples. This is known as the posterior distribution and is the probability of a specific class given evidence in the form of features. In our formant data, it is the probability of a specific vowel given the formant measurements. Category decisions are made by examining the posterior probability for each class and selecting the class associated with the highest one. This is known as a Bayes classifier (Hastie et al., 2009) and is optimal when the posterior distributions are correct. As learned distributions are approximations, this assumption is rarely met.

The posterior distribution can be difficult to estimate. It is common to solve an equivalent maximization. The posterior can be replaced by the product of the probabilities of evidence given the assumption of a specific class (the class-conditional probability) and the probability of the class occurring (the prior probability). An example of a prior probability is someone saying "Hello" at random versus the class-conditional probability of someone saying "Hello" when greeting someone.

Mixture models (Hastie et al., 2009) are an example of a distribution learner that use a linear combination of simple parametric distributions (e.g., Gaussians) to model complex distributions. Each distribution in the model has a weight that controls its contribution to the complex distribution.

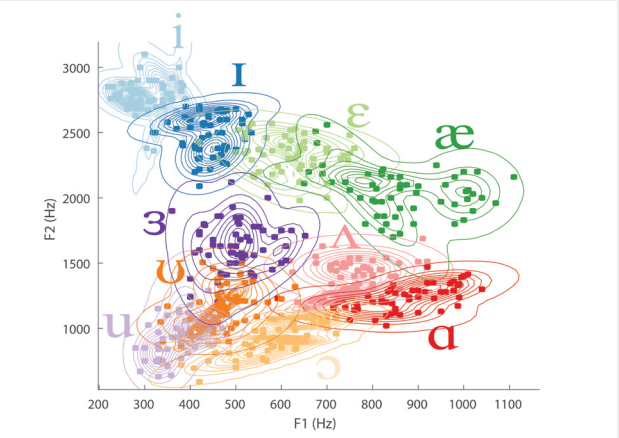


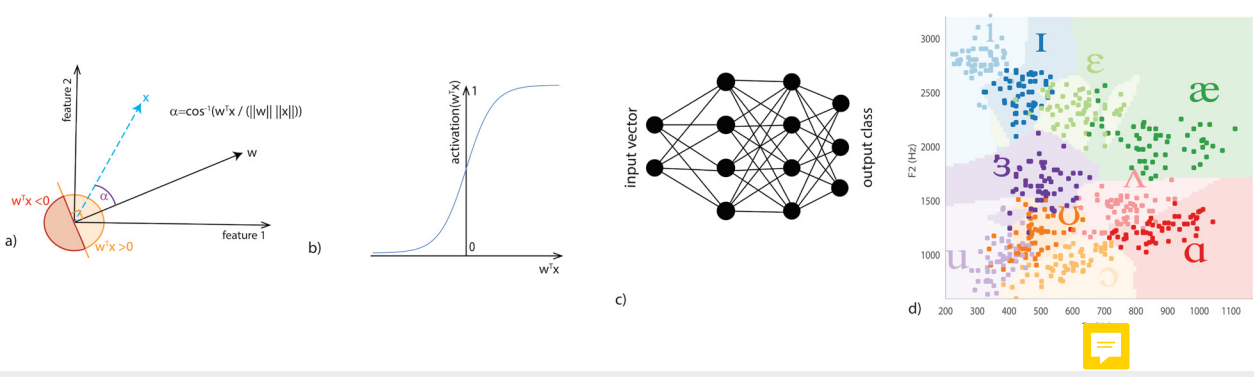
Figure 2. *Gaussian mixture models of the formant data. Isocontours show the probability surfaces for each vowel modeled with one to three Gaussians. The number of Gaussians was selected by maximizing the between cluster variance to the within cluster variance.*

Training the models requires estimating the mixture weights and the parameters of each parametric distribution. This can be done with an iterative procedure that alternates between determining the expected value of the mixture weights and improving the mixture parameters through maximum likelihood estimation (Bianco et al., 2019). In this type of supervised learning, we learn the distribution of each class separately. In the formant data, we have trained one model for each vowel. The training of each model is a form of unsupervised learning because we do not label the variations within specific vowels. **Figure 2** shows lines of equal probability (isocontours) for each vowel, and these distributions could not easily be modeled with a single parametric distribution. If we wanted to classify new data, we would compute the probability of the formants for each model (class conditional) times the prior probability and select the vowel class that produced the highest probability.

Decision Boundary Learners

In contrast, decision boundary learners attempt to find curves or planes that best separate the data. Artificial neural networks are one such method inspired by neurons in the animal kingdom. Cowan and Sharp (1988) discuss early work in this area. In the 1940s, Pitts and McCulloch showed that these networks could be used to represent simple logic functions. Rosenblatt's 1953 work demonstrated that parameters of artificial neurons could be estimated from the training data. Interest in neural networks declined in the latter half of the 20th century due to networks frequently learning their training data too well.

Figure 3. *A feedforward artificial neural network.* ***a:*** *Each neuron computes the dot product of an input vector* x *and a learned weight vector* w*. The product is proportional to the cosine of the angle between the two vectors and is positive if the angle between them is less than 90°. Consequently, the sign of the dot product indicates to which side of the line perpendicular to* w *the vector* x *falls on.* ***b:*** *The dot product is the input to a differentiable nonlinear function called the activation function. Shown here is the sigmoid function that maps the dot product smoothly from 0 to 1.* ***c:*** *Neural networks consist of a series of nodes that each performs the steps in* ***a*** *and* ***b****, with their outputs forming a new input for the next layer. Learning in a neural network is the process of establishing weights that will produce the desired result and is accomplished my minimizing a loss function that measures the difference between the desired output and the produced one on training data.* ***d:*** *Partitions induced by a neural network trained on the vowel formant data.*



This process is called overfitting and results in a poor ability to generalize the learned function to new data.

In the early twenty-first century, the convergence of large datasets, regularization methods to prevent overfitting, and inexpensive parallel hardware (video cards) led to a resurgence of interest in neural networks (LeCun et al., 2015). Each node in a neural network takes a set of input values and combines them by taking the dot product between the inputs and a set of learned weights (**Figure 3a**). This step is similar to classical linear discriminant analysis although the weight vector is learned differently.

The dot product is then transformed (**Figure 3b**) by a differentiable nonlinear function called an activation (e.g., a sigmoid function or one that sets negative values to zero). Nodes are arranged into layers, and in a classic feed-forward network, the outputs of one layer serve as inputs to the next layer (**Figure 3c**). Thus, one can think about each node as making a local decision about which side of a hyperplane its inputs fall on and propagating this knowledge to subsequent nodes.

The final layer is responsible for the prediction and either outputs a predicted value for regression problems or a category (**Figure 3d**), frequently represented as a vector representing the probability of belonging to each class. The recent interest in deep networks, networks that have many layers, is due to these networks repeated ability to provide significant advances in the state of the art across a wide range of problem domains (LeCun et al., 2015) as well as transfer learning, which utilizes knowledge from an already trained network to a new dataset, similar but far from being identical to the original one.

Neural network training is usually accomplished by an iterative procedure called backpropagation (Goodfellow et al., 2016; Bianco et al., 2019). At each step, training examples are presented to the network. For each example, the loss, a measure of deviation from the intended result, is computed. The derivative (gradient) of the loss function with respect to a node's weights indicates the direction in which changing the weight vector would create the largest increase in loss. To decrease the loss, the weights can be modified by a small amount in the opposite direction (gradient descent). This technique can be "backpropagated" through the network, computing the loss gradient at each node and permitting adjustments to weights in layers other than the last one. The training process is repeated until a convergence criterion is met. Backpropagation depends on many factors, and a thorough discussion can be found in Goodfellow et al. (2016).

In acoustics, neural networks have provided advances in speech recognition (Hinton et al., 2012), room localization (Chakrabarty and Habets, 2019), direction of arrival estimation (Ozanich et al., 2020), bioacoustics (Stowell et al., 2019), sea bed classification (Frederick et al., 2020), and increasing speech intelligibility (Healy et al., 2019) among many other areas.

Two popular forms of neural networks are convolutional neural networks (CNNs) and recurrent neural networks (RNNs) (Goodfellow et al., 2016). Convolutional networks are used to recognize local structure in both the input space as well as in hidden layers that contain abstract representations of information needed to make a decision. Convolutional layers learn matched filters that are moved across the input or intermediary data. These outputs are filtered again and combined in subsequent layers and may have operations to reduce the dimension (pooling). A final set of feed-forward layers perform classification or regression. **Figure 4a** shows an application of a convolutional network to the problem of detecting a type of contact call produced by endangered North Atlantic right whales (*Eubalaena glacialis*). The first set of learned filters are shown. Some of these filters produce strong outputs when calls are present and others when calls are absent and serve as features for subsequent layers of the network.

Recurrent neural networks introduce dependencies among subsequent inputs, allowing the network to learn the temporal structure (**Figure 4b**). They are commonly used in time-varying acoustic problems where signal evolution is important, such as speech recognition (e.g., Amodei et al., 2016). A drawback to these types of units is that information decays at each time step, and it is difficult to learn long-term dependencies. There are several variants of this architecture such as gated recurrent units (GRUs) and long-short time memory units (LSTMs) that allow network nodes to learn concepts such as when the input is relevant or when the input history state should be cleared (Goodfellow et al., 2016). It is common to combine convolutional and recurrent networks, with the convolutional network acting as a feature extractor and

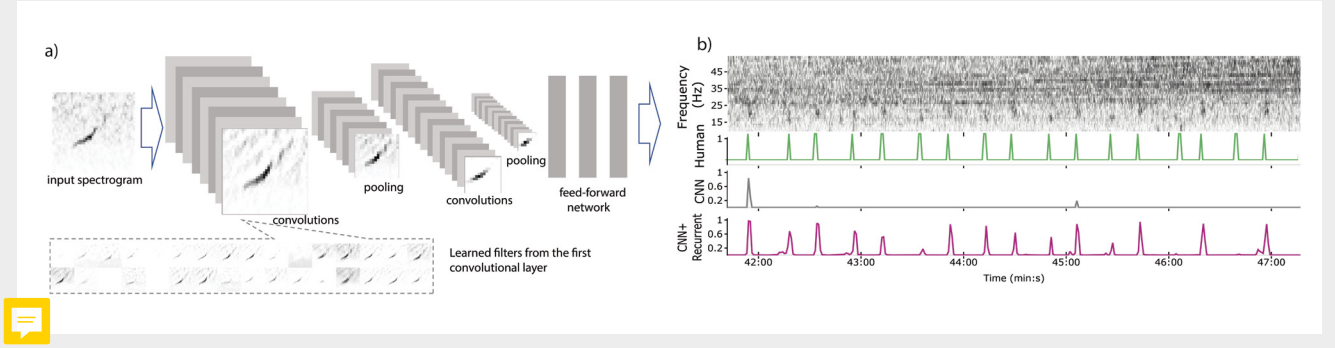


Figure 4. *Convolutional and recurrent neural networks.* ***a:*** *Convolutional neural network for recognizing endangered North Atlantic right whale (*Eubalaena glacialis*) upcalls. Spectrograms were presented to the network that learned convolutional filters representing examples of both upcall present and upcall absent spectrogram patches. Each filter is convolved with the input to create an output that has high values in areas that are similar to the filter. Max pooling takes the maximal value of the convolutional output over a small area, effectively downsampling the convolutional output and decreasing the importance of exact position within the spectrogram. In this network, there are two convolutional and two pooling layers, followed by a traditional multilayer feedforward network that classifies the representation of the input spectrogram extracted by the convolutional layers. Adapted from Shiu et al., 2020, with permission.* ***b:*** *Example of using recurrent networks to exploit context. The spectrogram shows the 20-Hz song of a fin whale (*Balaenoptera physalus*) in the presence of heavy shipping noise. The annotations by a human analyst are shown beneath, followed by a convolutional neural network (CNN) and a hybrid CNN with a recurrence layer. The CNN misses most song notes under these challenging conditions. The hybrid network has learned the song pattern and can better pick up weak notes of the song. Adapted from Madhusudhana et al., 2021, with permission.*

the recurrent network capturing temporal relationships between the features.

Bias and Variance

Most nontrivial problems have an inherent confusability that cannot be resolved regardless of the learner. This is called the Bayes error. Learning algorithms do not usually achieve the Bayes error, and additional error can be attributed to two sources, bias and variance (Hastie et al., 2009). Bias is the additional error that can be attributed to a learner not being capable enough to learn the distributions or separating boundary. For example, in the formant data, the vowels ʊ (book) and u (boot) contain examples that cannot be separated by a linear curve. Any classifier incapable of producing more complex boundaries would inherently be unable to model the separating boundary correctly.

The second component of error is variance. Variance reflects the error that is due to a specific training set. When learners are very sensitive to small changes in the training data, they have high variance. More complex learners tend to have a higher variance, and regularization strategies are a method to mitigate for this.

A common method to reduce the amount of variance error is to train multiple classifiers and make a decision based on a function of their decisions such as a vote. This is an effective method of reducing variance and is the basis of ensemble learning methods such as random forest (Breiman, 2001). Random forest makes decisions based on a set of decision trees, classifiers that make a series of branching decisions that depend on values of features, much like the popular children's game of 20 questions: "Are you thinking of an animal?"/"Yes"/"Is it large?"

Evaluating Learning

One of the goals of researchers using machine learning algorithms to solve applied problems is to ensure that the algorithms are actually useful in novel environments. As such, one needs to take care when evaluating how well an algorithm performs.

Unsupervised Learning Metrics

In unsupervised learning, there are intrinsic and extrinsic measurements of performance. Intrinsic measurements on clustered data examine the quality of data partitions and usually use some variant of measuring the similarity within a cluster versus the similarity between clusters.

The silhouette algorithm is a popular measure that has been demonstrated to correlate well with human intuition on two-dimensional clustering tasks (Lewis et al., 2012). In contrast, extrinsic measurements examine the similarity between partitionings. This can be done to examine different clusterings of the same or similar data for consistency or to determine how well the cluster analysis agrees with an established partitioning method such as human analysts. The adjusted Rand index is one popular extrinsic measure that compares whether or not pairs of examples are consistently assigned to the same cluster or different ones (Hubert and Arabie, 1985).

Supervised Learning Metrics

In supervised learning scenarios, the main question is whether or not the learner will perform well on novel data. Learners should never be evaluated on the data that were used to train them. *N*-fold cross-validation (Bianco et al., 2019) is a commonly used strategy that splits data into *N* partitions. *N* different trials are conducted, with *N*–1 partitions contributing training data and the remaining partition being used for testing. When developing a classification algorithm, it is rare to produce something that works satisfactorily on the first try. Typically, there are a series of experiments where the model parameters are adjusted such that the model performs better on the test data. Such adjustments can be seen as a weak form of training, and as a consequence, it is recommended to have a held-out dataset that is not evaluated until one is satisfied with the learning algorithm.

Various metrics have been used for evaluating supervised learning. A common detection task is to determine if a type of signal is present within a fixed time bin. There are two types of error for this task. False positives or false alarms occur when a bin is mistakenly reported as an occurrence of the signal of interest. False negatives or misses occur when a signal of interest is not reported within the bin. Whether or not a signal is reported is dependent on the threshold. For example, if a neural network produced a probability score, we might set a lower threshold if our goal was to find all instances of a signal and a higher threshold if our goal was to minimize the nuisance caused by false alarms. Various plots have been proposed to visualize this variability. The receiver operating curve (ROC; Fawcett, 2006) plots the false positive rate versus the true positive rate at different thresholds.

Useful variants of this are the detection error tradeoff (DET) curve (Martin et al., 1997) and the precision recall (PR) curve (Davis and Goadrich, 2006). The DET curve makes the assumption that scores are normally distributed and scales the plots using a standard normal deviate. This has the desirable property of separating curves that are close together in ROC space, making it easier to compare systems. DET curves can also add penalties for different types of error, making it easier to see how the performance varies with respect to specific operational goals.

PR curves plot the percentage of detections that are correct (precision) against the percentage of target signals that were correctly detected (recall). PR curves are independent of the number of signal-absent bins. For rare signals in a long time series, PR curves offer a significant advantage. The number of correctly classified signal-absent cases plays a role in ROC/DET curves and can result in low false-positive rates even when the false-positive count greatly exceeds the number correctly detected signals. PR curves also offer the advantage of not requiring detections to be reported on fixed time bins. The F1 metric is the harmonic mean of the precision and recall at a specific operating point and can be used to summarize a point on the PR curve into a single number.

For classification tasks with multiple categories, the error rate is commonly reported, and confusion matrices are frequently used to visualize the results. The rows of a confusion matrix represent actual categories, whereas the columns represent predicted categories. Counts or percentages summarize how well the system functions, with correct classifications being shown along the diagonal of the matrix.

Finally, regression tasks use some measurement of how far the prediction is from the desired target. The squared error distance is a common measurement.

Discussion

One of the drawbacks of many machine learning techniques is the so-called "black box" syndrome, where the predictions of a learner are not interpretable by the user. Some methods, such as the aforementioned decision trees, have the quality of being explainable, which can be very helpful when trying to understand why classification failed. Most techniques, such as deep neural networks that have millions of parameters, are very difficult to interpret, and correcting errors usually requires expert

insight as to the root cause of a problem. Improving the ability to explain such models is an open area of research (see Linardatos et al., 2021 for a review). Strategies for understanding why models make the predictions they do are varied but include techniques such as drawing attention to portions of the input signal that were responsible for strong activations of a neural network.

Another exciting avenue of machine learning research is to utilize systems that take advantage of physical knowledge. An example of this can be seen in Raissi et al. (2019) who trained deep neural networks with priors that were grounded in the physics of problem domains. One can envision acoustic systems that have prior knowledge about e.g., transmission loss and channel characteristics, and such systems may be a promising area for future research.

We hope that this "gentle" introduction to machine learning will inspire readers to dig deeper into the possible uses of machine learning in their own acoustics problems. *The Journal of the Acoustical Society* published a special issue in 2021 on the use of machine learning in acoustics, and this collection of papers provides a wide range of example applications including medical applications, speech, oceanography, bioacoustics, and music. We hope that this collection stimulates the wider adoption of machine learning within the field of acoustics. There are a growing number of published acoustics papers that use these techniques, and it is likely that machine learning will become a valuable component in the acoustician's toolkit.

Acknowledgments

This work was supported by Office of Naval Research Awards N00014-17-1-2867 and N00014-20-1-2029. We thank Arthur Popper for his valuable suggestions in the development of this manuscript.

References

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., and Chen, J. (2016). Deep speech 2: End-to-end speech recognition in English and Mandarin. In *Proceedings of The 33rd International Conference on Machine Learning*, New York, NY, June 19-24, 2016, vol. 48, pp. 173-182.

Benson, J., Chapman, N. R., and Antoniou, A. (2000). Geoacoustic model inversion using artificial neural networks. *Inverse Problems* 16, 1627-1639. https://doi.org/10.1088/0266-5611/16/6/302.

Bianco, M. J., Gerstoft, P., Traer, J., Ozanich, E., Roch, M. A., Gannot, S., and Deledalle, C. A. (2019). Machine learning in acoustics: Theory and applications. *The Journal of the Acoustical Society of America* 146, 3590-3628. https://doi.org/10.1121/1.5133944.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer-Verlag, New York, NY.

Breiman, L. (2001). Random forests. *Machine Learning* 45, 5-32. https://doi.org/10.1023/a:1010933404324.

Chakrabarty, S., and Habets, E. A. P. (2019). Multi-speaker DOA estimation using deep convolutional networks trained with noise signals. *IEEE Journal of Selected Topics in Signal Processing* 13, 8-21. https://doi.org/10.1109/JSTSP.2019.2901664.

Cowan, J. D., and Sharp, D. H. (1988). Neural nets and artificial intelligence. *Daedalus* 117, 85-121.

Davis, J., and Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In *Proceedings of the International. Conference on Machine Learning*, Pittsburgh, PA, June 25-29, 2006, pp. 233-240.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters* 27, 861-874. https://doi.org/10.1016/j.patrec.2005.10.010.

Frederick, C., Villar, S., and Michalopoulou, Z.-H. (2020). Seabed classification using localized forward modeling and deep learning. *The Journal of the Acoustical Society of America* 148, 2730. https://doi.org/10.1121/1.5147584.

Godino-Llorente, J. I., and Gomez-Vilda, P. (2004). Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors. *IEEE Transactions on Biomedical Engineering* 51, 380-384. https://doi.org/10.1109/TBME.2003.820386.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. The MIT Press, Cambridge, MA.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, New York, NY.

Healy, E. W., Delfarah, M., Johnson, E. M., and Wang, D. (2019). A deep learning algorithm to increase intelligibility for hearing-impaired listeners in the presence of a competing talker and reverberation. *The Journal of the Acoustical Society of America* 145, 1378. https://doi.org/10.1121/1.5093547.

Hinton, G., Li, D., Dong, Y., Dahl, G. E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., and Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine* 29, 82-97. https://doi.org/10.1109/MSP.2012.2205597.

Hubert, L., and Arabie, P. (1985). Comparing partitions. *Journal of Classification* 2, 193-218. https://doi.org/10.1007/bf01908075.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436-444. https://doi.org/10.1038/nature14539.

Lewis, J. M., Ackerman, M., and de Sa, V. R. (2012). Human cluster evaluation and formal quality measures: A comparative study. In *Proceedings of the 34th Annual Meeting of the Cognitive Science Society*. Sapporo, Japan, August 1-4, 2012, pp. 1870-1875.

Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2021). Explainable AI: A review of machine learning interpretability methods. *Entropy* 23, 18.

Madhusudhana, S., Shiu, Y., Klinck, H., Fleishman, E., Liu, X., Nosal, E. M., Helble, T., Cholewiak, D., Gillespie, D., Širović, A., and Roch, M. A. (2021). Improve automatic detection of animal call sequences with temporal context. *Journal of the Royal Society Interface* 18, 20210297. https://doi.org/10.1098/rsif.2021.0297.

Martin, A., Doddington, G., Kamm, T., Ordowski, M., and Przybocki, M. (1997). The DET curve in assessment of detection task performance. In *Proceedings of the 5th European Conference on Speech Communication and Technology*, Rhodes, Greece. September 22-25, 1997, vol. 4, pp. 1895-1898.

McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. Available at https://arxiv.org/abs/1802.03426. Accessed June 1, 2020.

Niu, H., Reeves, E., and Gerstoft, P. (2017). Source localization in an ocean waveguide using supervised machine learning. *The Journal of the Acoustical Society of America* 142, 1176. https://doi.org/10.1121/1.5000165.
Ozanich, E., Gerstoft, P., and Niu, H. (2020). A feedforward neural network for direction-of-arrival estimation. *The Journal of the Acoustical Society of America* 147, 2035-2048. https://doi.org/10.1121/10.0000944.
Peterson, G. E., and Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America* 24, 175-184.
Raissi, M., Perdikaris, P., and Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics* 378, 686-707. https://doi.org/10.1016/j.jcp.2018.10.045.
Ravanelli, M., and Bengio, Y. (2018). Speaker recognition from raw waveform with SincNet. In *Proceedings of 2018 IEEE Spoken Language Technology Workshop (SLT 2018)*, Athens, Greece, December 18-21, 2018, pp. 1021-1028.
Russell, S. J., and Norvig, P. (2021). *Artificial Intelligence: A Modern Approach*. Prentice Hall, Upper Saddle River, NJ.
Sainburg, T., Thielk, M., and Gentner, T. Q. (2020). Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. *PLoS Computational Biology* 16, e1008228. https://doi.org/10.1371/journal.pcbi.1008228.
Schedl, M., Gómez, E., and Urbano, J. (2014). Music information retrieval: Recent developments and applications. *Foundations and Trends in Information Retrieval* 8, 127-261. https://doi.org/10.1561/1500000042.
Shah, T., Zhuo, L., Lai, P., Rosa-Moreno, A. D. L., Amirkulova, F., and Gerstoft, P. (2021). Reinforcement learning applied to metamaterial design. *The Journal of the Acoustical Society of America* 150, 321-338. https://doi.org/10.1121/10.0005545.
Shiu, Y., Palmer, K. J., Roch, M. A., Fleishman, E., Liu, X., Nosal, E. M., Helble, T., Cholewiak, D., Gillespie, D., and Klinck, H. (2020). Deep neural networks for automated detection of marine mammal species. *Scientific Reports* 10, 607. https://doi.org/10.1038/s41598-020-57549-y.
Stowell, D., Wood, M. D., Pamuła, H., Stylianou, Y., and Glotin, H. (2019). Automatic acoustic detection of birds through deep learning: The first Bird Audio Detection challenge. *Methods in Ecology and Evolution* 10, 368-380. https://doi.org/10.1111/2041-210x.13103.
van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, 2579-2605.
Wang, C., Wang, Z., Sun, W., and Fuhrmann, D. R. (2018). Reinforcement learning-based adaptive transmission in time-varying underwater acoustic channels. *IEEE Access* 6, 2541-2558. https://doi.org/10.1109/ACCESS.2017.2784239.
Xi, S., Changsheng, X., and Kankanhalli, M. S. (2004). Unsupervised classification of music genre using hidden Markov model. In *Proceedings of the 2004 IEEE International Conference on Multimedia and Expo*, Taipei, Taiwan, June 27-30, 2004, vol. 3, pp. 2023-2026.
Zissman, M. A. (1996). Comparison of four approaches to automatic language identification of telephone speech. *IEEE Transactions on Speech and Audio Processing* 4, 31. https://doi.org/10.1109/TSA.1996.481450.

About the Authors

Marie A. Roch
marie.roch@sdsu.edu

Department of Computer Science
San Diego State University
5500 Campanile Drive
San Diego, California 92182-7720, USA

Marie A. Roch received her PhD from the University of Iowa, Iowa City. Before joining San Diego State University, San Diego, California, as a professor of computer science, she was at the AT&T Bell Telephone Laboratories, Murray Hill, New Jersey. Her research interests are in animal bioacoustics for conservation and mitigation, communication, and behavior. For more information, see roch.sdsu.edu.

Peter Gerstoft
pgerstoft@ucsd.edu

Scripps Institution of Oceanography
University of California, San Diego
9500 Gilman Drive
La Jolla, California 92093-0238, USA

Peter Gerstoft received his PhD from the Technical University of Denmark, Kongens Lyngby, Denmark, in 1986. Since 1997, he has been with the University of California, San Diego, La Jolla. His current research interests are signal processing and machine learning applied to acoustic, seismic, and electromagnetic signals. For more information, see noiselab.ucsd.edu. His work has been featured on *The Late Show with Stephen Colbert*.

Bożena Kostek
bozenka@sound.eti.pg.gda.pl

Faculty of Electronics, Telecommunications and Informatics
Gdansk University of Technology
ul. Narutowicza 11/12
80-233 Gdansk, Poland

Bożena Kostek is a professor at the Gdansk University of Technology, Gdansk, Poland. She is a corresponding member of the Polish Academy of Sciences and a Fellow of the Acoustical Society of America and of the Audio Engineering Society. She has published more than 600 scientific papers and has led many research projects. She is the recipient of many prestigious awards, including two first prizes from the Prime Minister of Poland, several prizes from the Minister of Science, and an award from the Polish Academy of Sciences.

Zoi-Heleni (Eliza) Michalopoulou
michalop@njit.edu

Department of Mathematical Sciences
New Jersey Institute of Technology
University Heights
Newark, New Jersey 07102-1982, USA

Zoi-Heleni (Eliza) Michalopoulou received her PhD from Duke University, Durham, North Carolina. She is a professor at the New Jersey Institute of Technology, Newark. She is a Fellow of the Acoustical Society of America and a Senior Member of the IEEE. Her research interests include ocean acoustics, Bayesian modeling, inverse problems, array signal processing, and machine learning.