unsupervised learning

Authors

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland
*Corresponding author: alex.jung@aalto.fi

Abstract

Unsupervised learning is the machine learning (ML) setting in which the data points of a dataset carry features but no label, so nothing says what a correct answer for a data point would be. What is sought instead is structure: clustering collects similar data points into clusters, dimensionality reduction replaces many features by few, and density estimation fits a probabilistic model to the feature vectors. The price of the missing labels is the missing criterion. A validation error tells a supervised learning method when a hypothesis is worse, while the clustering error keeps falling as clusters are added and so cannot say how many there should be. Judging an unsupervised result needs something outside it, such as a downstream learning task or a person inspecting the clusters. Self-supervised learning and semi-supervised learning (SSL) are the other two answers to the cost of labels, differing in what they do with the same unlabeled data points.

Definition

The same weather station at Krems records eight measurements for each day of 2024: three temperatures, precipitation, sunshine duration, humidity, air pressure and wind speed. Nothing in the record says which of them is to be predicted from the others, and no quantity outside the record is attached to a day. What can still be asked of these days is how they arrange themselves: which of them resemble each other, how many of the eight numbers are needed to tell them apart, and which combinations of values are common. Unsupervised learning is the machine learning (ML) setting in which the data points of a dataset carry features but no label, and what is sought is structure among them (Hastie et al., 2009, Sect. 14.1; Sutton and Barto, 2018, Sect. 1.1).

Formally, the input is a dataset $\mathcal{D} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\}$ of feature vectors $\mathbf{x}^{(r)} \in \mathcal{X}$, with no label space in sight. In supervised learning the label of a data point says what the correct answer would have been, and here nothing does, which is why the two settings are called learning with and without a teacher (Hastie et al., 2009, Sect. 14.1). Three learning tasks make up most of the setting, and the days at Krems carry all three.

Clustering partitions the data points into clusters of similar ones. k-means with $k = 2$ clusters, applied to the eight measurements after each is scaled to zero sample mean and unit sample variance, reaches a fixed point after ten iterations. Its two clusters agree with the cold and the warm half of the year

on 90% of the days (see Fig. 1), although no month, season or other label entered the computation. Dimensionality reduction replaces the eight features by fewer: principal component analysis (PCA) delivers two, which carry 64% of the total variance and leave a reconstruction error equal to the number of data points times the sum of the dropped eigenvalues of the sample covariance matrix. Density estimation fits a probabilistic model to the feature vectors: a normal distribution fitted to the temperatures of the first 244 days assigns the remaining days an average log-density of -6.45, against -7.19 for one that keeps the same means but a diagonal covariance matrix.

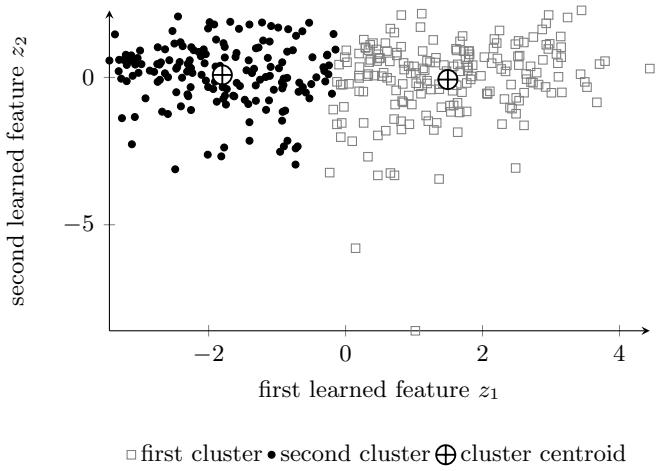

Figure 1: The 366 days of 2024 at Krems, each a data point with eight features and no label, drawn in the two features learned by PCA and marked by the cluster that k-means assigned them. Neither the clusters nor the two features used any quantity beyond the eight measurements. Data generated by `pythondemos/unsupervisedlearning.py`

What the missing labels cost is a criterion. Hastie, Tibshirani and Friedman state it plainly: supervised learning has a clear measure of success, the loss on data points held out of training, and unsupervised learning has no such direct measure (Hastie et al., 2009, Sect. 14.1). The loss minimized by an unsupervised method is intrinsic to the method, so it compares its own outputs and not its fitness for a purpose. The clustering error of k-means on the days at Krems shows what follows: it falls from 2928 for one cluster to 1940, 1651, 1414, 1260 and 1156 as clusters are added, and it would reach zero if every day were its own cluster. The criterion therefore cannot say how many clusters the days have, whereas a validation error does grow once a hypothesis becomes worse. Judging an unsupervised result needs something from outside it: a downstream learning task it should help, or a person inspecting the clusters.

Unsupervised learning is one of three answers to the cost of labels, and the three differ in what they do with the unlabeled data points. It uses them on their

own and gives up the label entirely. Self-supervised learning constructs labels from the data points themselves by withholding some of their features, which turns the same raw collection into an ordinary supervised learning problem. Semi-supervised learning (SSL) keeps a few labeled data points and uses the unlabeled ones to support them. The boundary is a matter of where the label comes from, not of the data: one unlabeled dataset serves all three.

Cross References

clustering, k-means, dimensionality reduction, PCA, probabilistic model, label, dataset, supervised learning, self-supervised learning, SSL, clustering error, validation error

References

1. Sutton and Barto (2018). Reinforcement Learning: An Introduction. 2nd ed., MIT Press, Cambridge, MA, USA.

2. Hastie et al. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed., Springer Science+Business Media, New York, NY, USA.