AWS re:Invent

AIM421-R

How to use unsupervised ML to find patterns, meaning, and anomalies

Tom Faulhaber
Principal Engineer
Amazon SageMaker
Amazon Web Services

AWS re:Invent

aws

Agenda

Let’s get to know each other

Overview of unsupervised learning

Amazon SageMaker algorithms for unsupervised learning

Building autoencoders

Deep dive - you choose

Let’s get to know each other

Supervised learning

Supervised learning requires “labeled” data:

> “For each observation of the predictor measurement(s) x_i, $i=1,\ldots, n$ there is an associated response measurement y_i. We wish to fit a model that relates the response to the predictors”
>
> James, et al., *Introduction to Statistical Learning in R*

Labels are often applied by humans but sometimes can be derived from the data

Labeling is hard and expensive

Sometimes labeling doesn’t make sense or is wrong

The most common goal of supervised learning is predictions

Unsupervised learning

Unsupervised learning operates on unlabeled data

Unsupervised learning algorithms discover patterns or relationships within the data

Types of unsupervised learning include clustering, dimensionality reduction, anomaly detection, text analysis, and encoding

Some forms of unsupervised learning are used as parts of pipelines with other ML algorithms

Amazon SageMaker algorithms for unsupervised learning

Clustering

- K-Means

Dimensionality reduction

- Principal Component Analysis (PCA)

Anomaly detection

- Random Cut Forest

Text

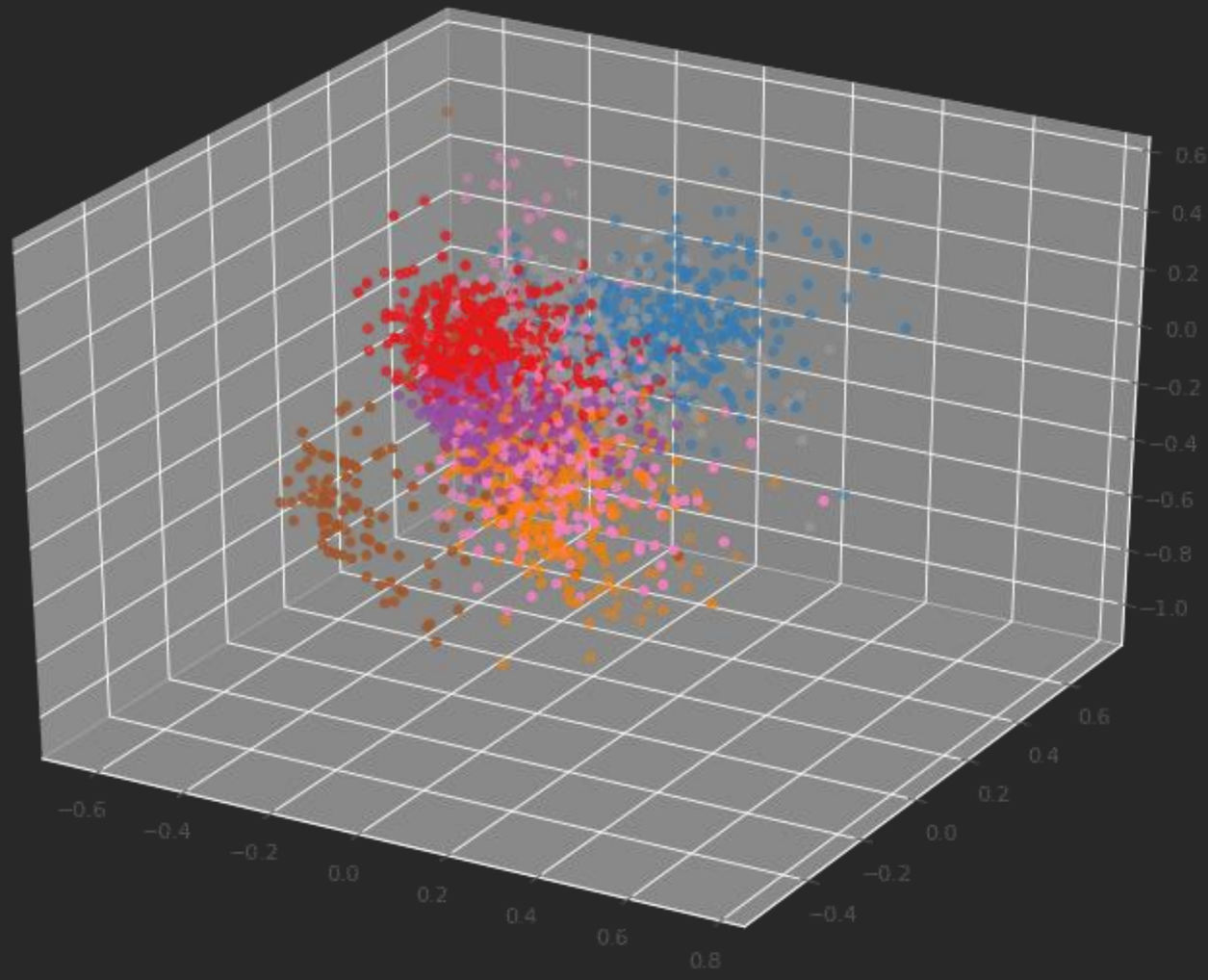
- Latent Dirichlet Allocation (LDA)
- Neural Topic Model
- BlazingText (Word2Vec)
- Object2Vec

K-Means clustering

Finds k clusters of data points within your data

Once you've trained it, you can use inference to see what cluster new data belongs to

Useful for problems like customer segmentation, related product selection, quantization

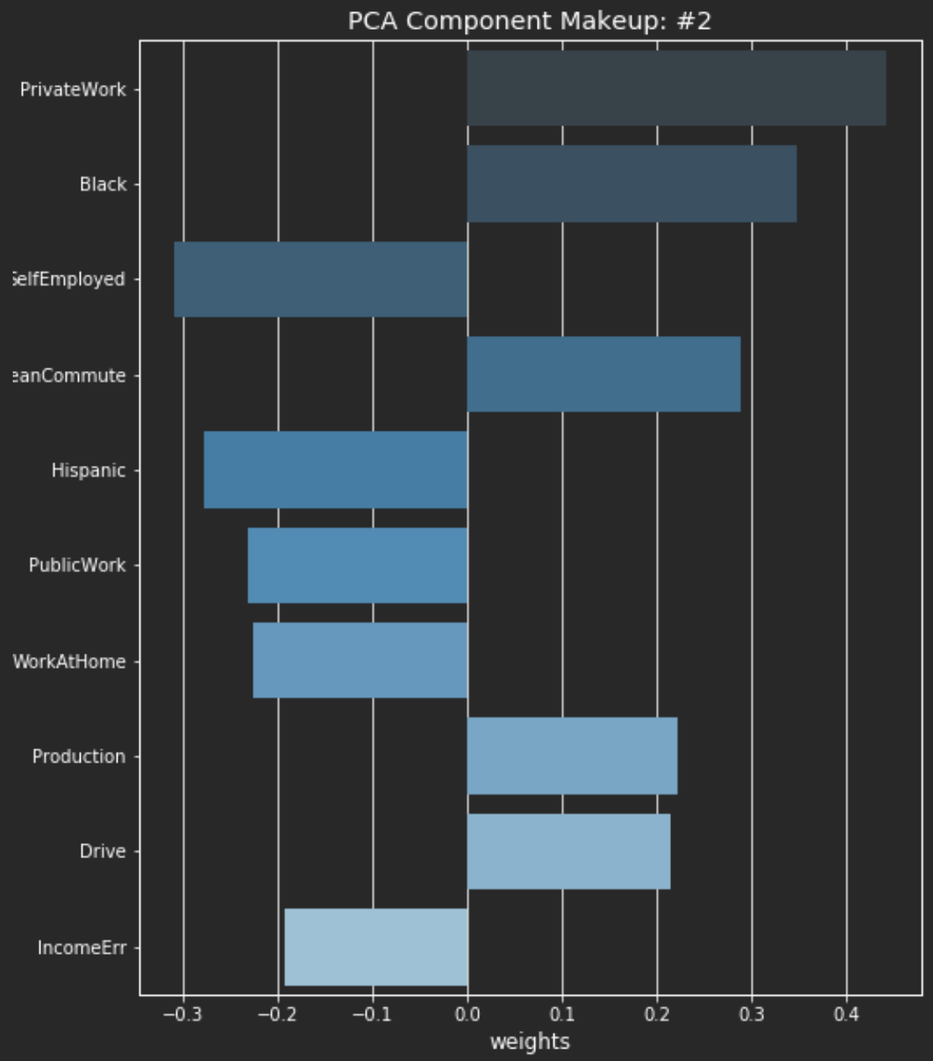


Principal Component Analysis (PCA)

A classic dimensionality reduction technique

Find orthogonal axes with the highest variation; each new axis is a linear combination of the original axes

Used for data visualization, suppressing correlated features, using high-dimensional data in low-dimensional contexts

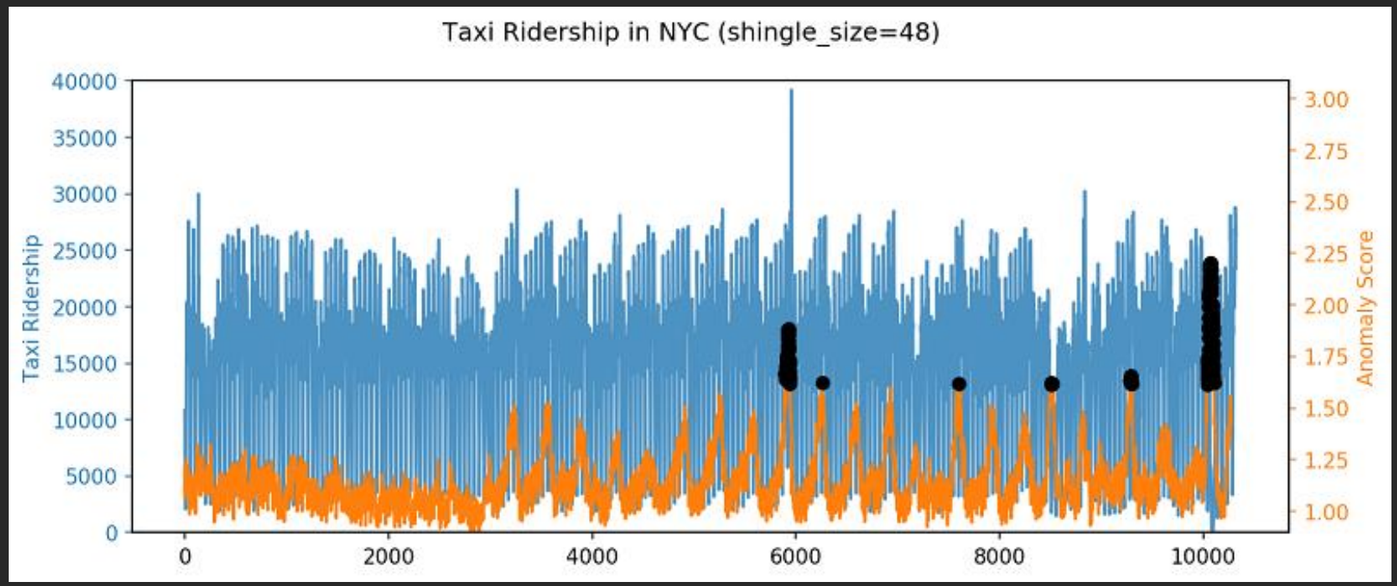


Random Cut Forest

Uses random forest techniques to identify anomalies in datasets

Gives each data point an ”anomaly score”; can be used on time-series data

Useful for identifying potential fraud, cyberattacks, service outages, unexpected developments



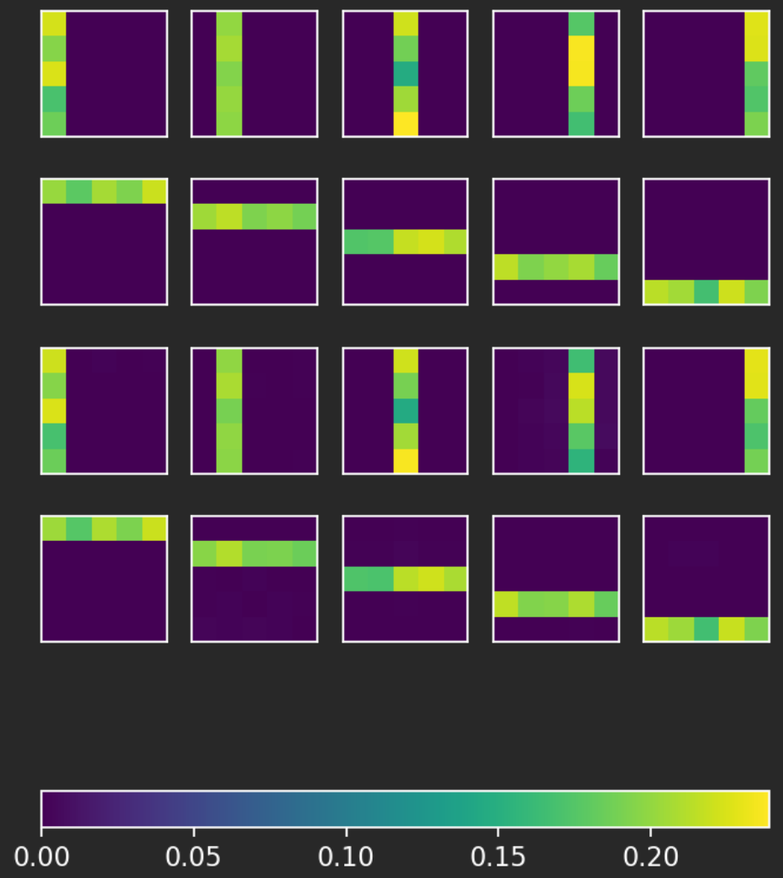
Latent Dirichlet Allocation (LDA)

Uses spectral decomposition to find topics in sets of documents

Spectral LDA is fast, highly scalable, and robust; uses bag-of-words model

Mostly used for documents, but can be used for things like modeling customer cohorts, etc.

Known vs. found topic-word probability distributions

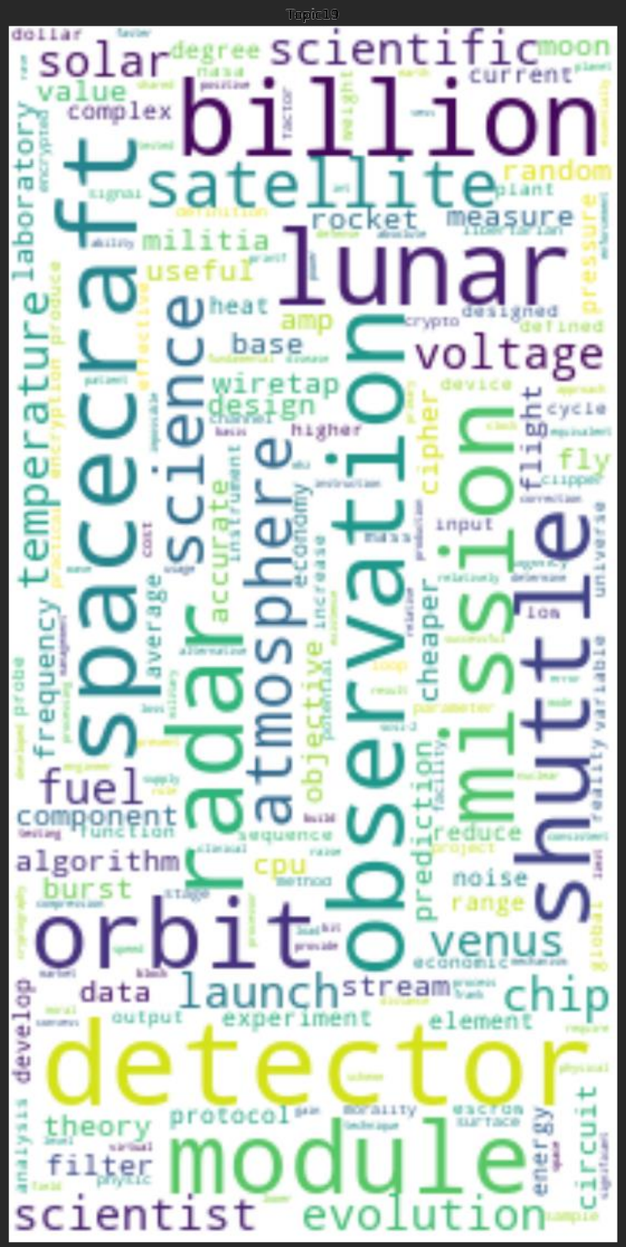


Neural Topic Model

Another algorithm for discovering topic mixtures in document collections

Uses a deep-learning model rather than a pure statistical model

We recommend you try both this and LDA to see which works better in your use case

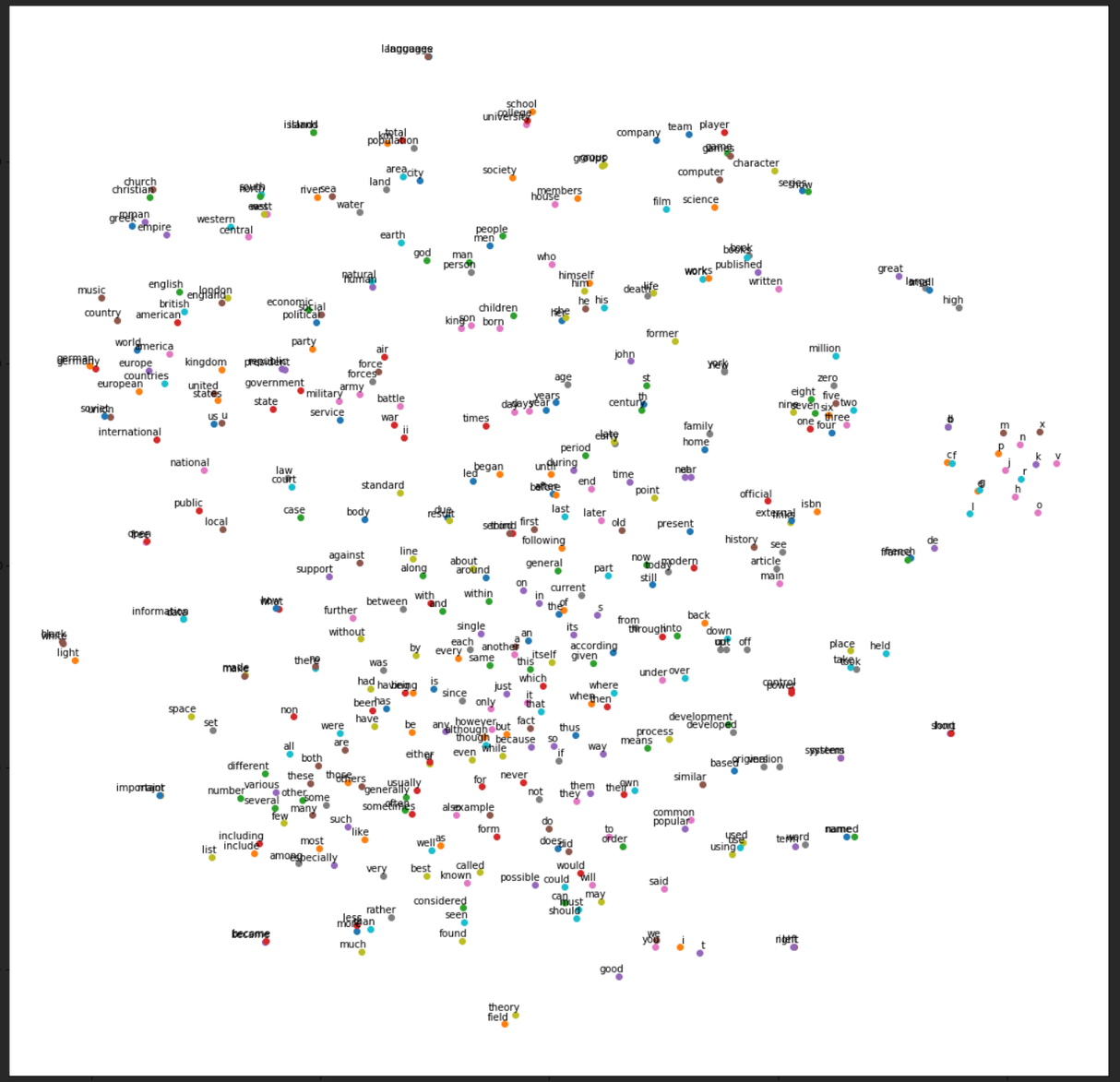


BlazingText (Word2Vec)

Word2Vec converts words in documents into vectors, called embeddings; these preserve semantic relationships

BlazingText can operate at subword level, making it more generalizable

Embeddings are used as input to other NLP algorithms like sentiment analysis

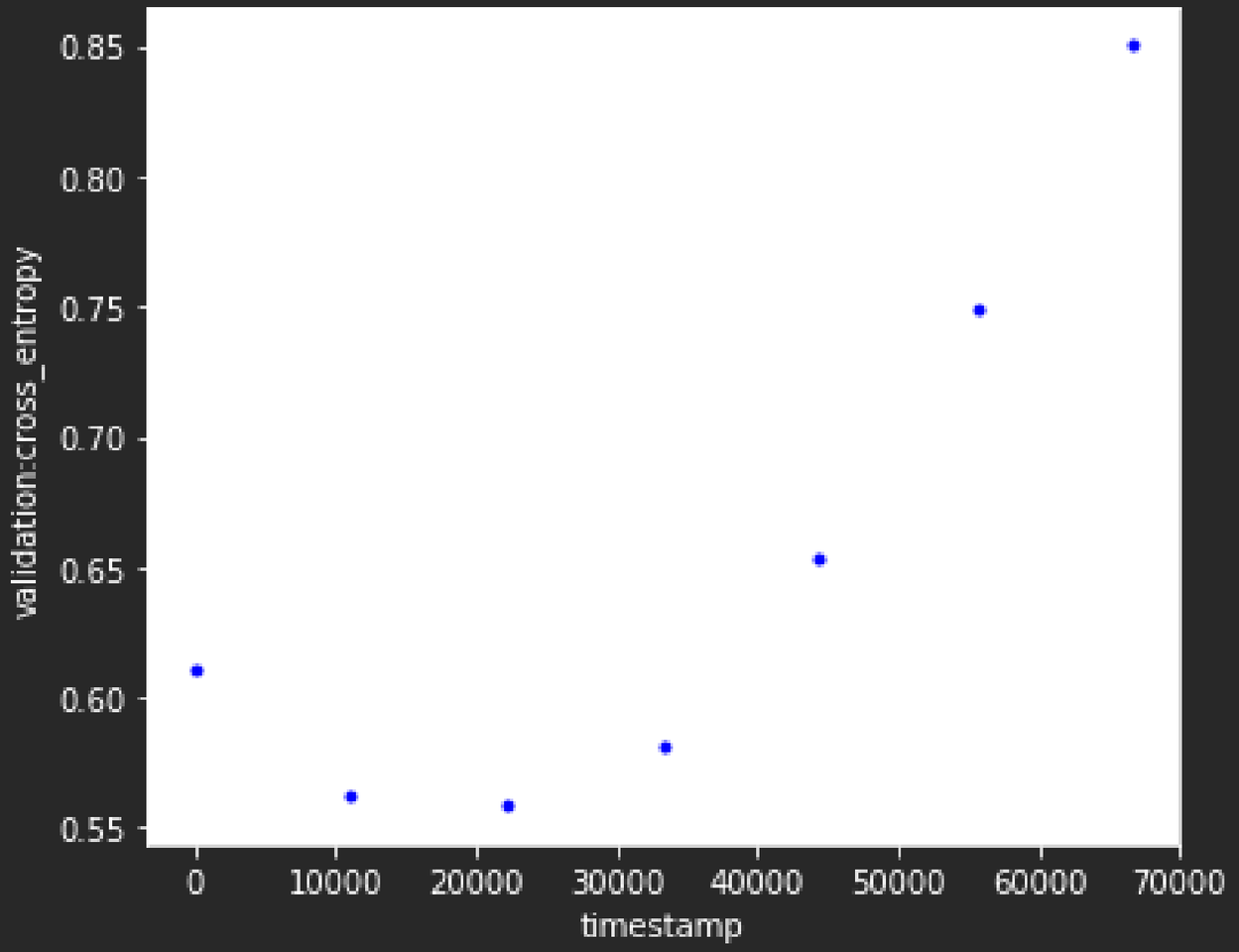


Object2Vec

Object2Vec is designed to learn a much wider variety of embeddings than Word2Vec

Handles both words and sequences of words

Can be used in bioinformatics, signal recognition, etc., in addition to complex document tasks

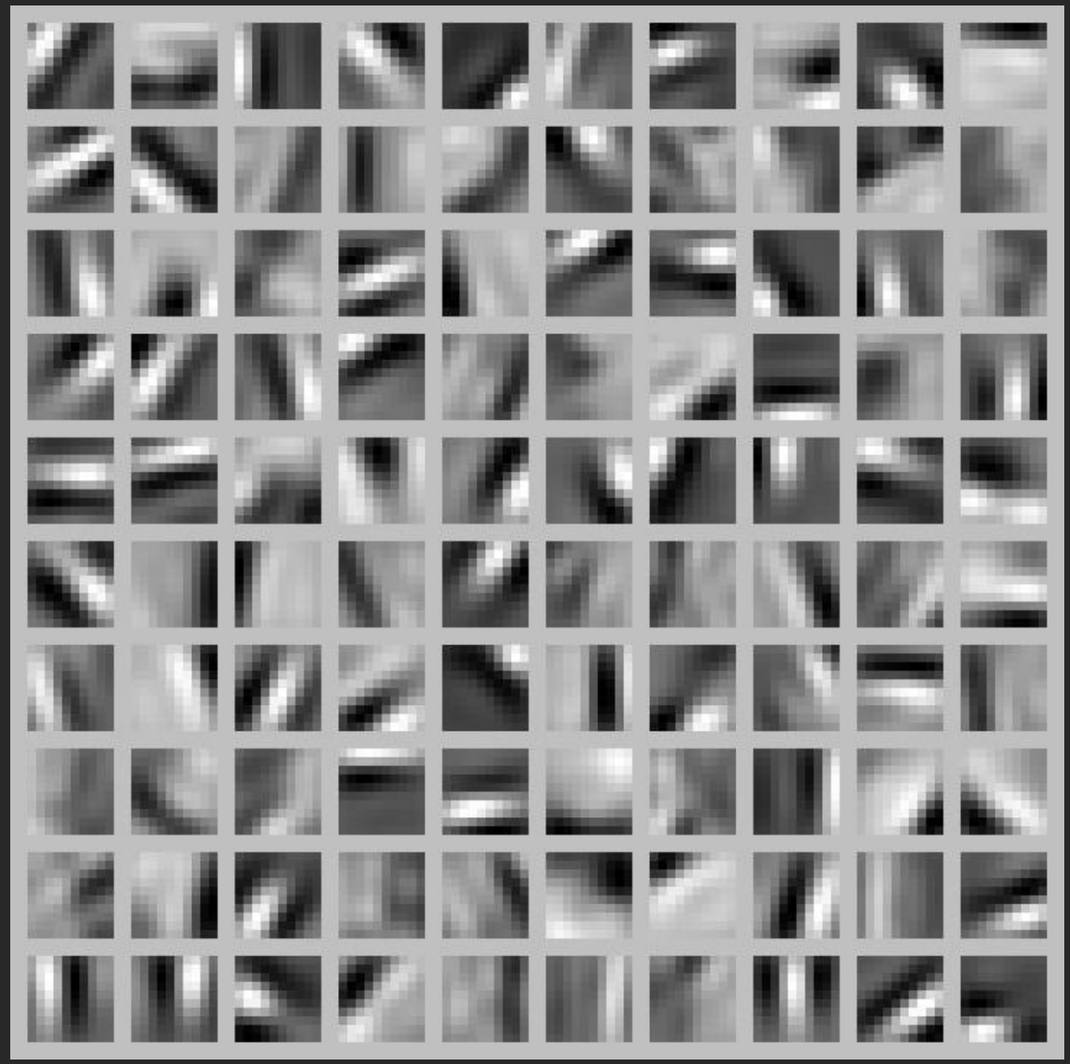


Create your own unsupervised learning algorithm

Use Amazon SageMaker's supported frameworks: PyTorch, MXNet, TensorFlow, Chainer

Autoencoding is a way to build your own embeddings

For example, you can build an autoencoder for images



Deep dive

Related breakouts

AIM-331 Choose the proper algorithm in Amazon SageMaker

AIM-415 Build fraud detection systems with Amazon SageMaker

AIM-319 Amazon SageMaker: Use prebuilt Jupyter notebooks

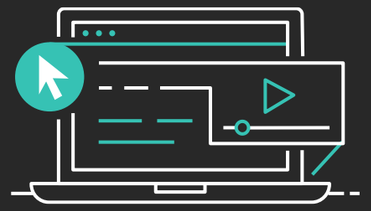
AIM-343 Build computer vision models with Amazon SageMaker

AIM-338 Machine learning with containers and Amazon SageMaker

Learn ML with AWS Training and Certification

The same training that our own developers use, now available on demand

Role-based ML learning paths for developers, data scientists, data platform engineers, and business decision makers



70+ free digital ML courses from AWS experts let you learn from real-world challenges tackled at AWS

Validate expertise with the
AWS Certified Machine Learning - Specialty exam

Visit https://aws.training/machinelearning

aws training and certification

Thank you!

AWS re:Invent

aws

Please complete the session survey in the mobile app.