[Draw your reader in with an engaging abstract. It is typically a short summary of the document. When you're ready to add your content, just click here and start typing.]

Research summary

Machine Learning

BRAULIO ROLANDO MILLAN CHIN
IRC9A
September 7th 2023

Introduction to Machine Learning

Fundamental concepts of Machine Learning

Concept and characteristics of supervised and unsupervised learning

Supervised learning is a type of machine learning where the algorithm learns from labeled data, which means the input data is paired with corresponding target labels. The primary goal of supervised learning is to learn a mapping function that can accurately predict the target labels for new, unseen data.

The performance of a supervised learning model is typically evaluated using metrics like accuracy, precision, recall, F1-score, or mean squared error, depending on the type of problem (classification or regression).

Unsupervised learning, on the other hand, involves learning patterns and structure from unlabeled data. In unsupervised learning, the algorithm explores the data's inherent structure or relationships without the guidance of predefined labels.

Since unsupervised learning doesn't have access to ground truth labels, evaluating model performance can be more challenging. Evaluation often relies on internal metrics or qualitative assessment.

Concept of the probabilistic model

A probabilistic model is a mathematical framework that represents uncertainty using probability distributions. These models are used to describe and make predictions about random variables and their relationships.

Probabilistic models can be categorized as generative or discriminative. Generative models aim to model the joint probability distribution of both inputs and outputs, allowing them to generate new data samples. Discriminative models, on the other hand, focus on modeling the conditional probability of outputs given inputs and are often used for classification tasks.

Differences between supervised and unsupervised learning

	SUPERVISED	UNSUPERVISED
Objective	*It makes predictions or classifications based on labeled training data.*	*It explores the structure of the input data to identify patterns or group similar data points.*
Data Type	*Requires labeled data, meaning each data point in the training set is paired with a corresponding output label or target value.*	*It does not have access to explicit output labels during training.*
Training Process	*The algorithm is trained to minimize the difference between its predictions and the true labels in the training data. This typically involves optimizing a cost function*	*Aims to find a representation of the data that captures its underlying patterns. Often driven by optimizing some criteria that doesn't rely on labeled targets.*
Evaluation	*Often evaluated using metrics like accuracy, precision, recall, F1-score (for classification tasks), or mean squared error (for regression tasks).*	*Evaluation may rely on internal metrics (e.g., silhouette score for clustering) or qualitative assessment of the discovered patterns or clusters.*
Examples	*Image classification (assigning labels to images), speech recognition (converting spoken words into text), and sentiment analysis (determining the sentiment of text as positive or negative).*	*Clustering (grouping similar data points together), dimensionality reduction (reducing the number of features while preserving essential information), and generative modeling (creating new data samples that resemble the training data distribution).*

Differences between the concept of Regression and Classification

	REGRESSION	CLASSIFICATION
Objective	*To predict a continuous, numerical output or target variable*	*Focuses on assigning data points to predefined categories or classes.*
Output Variable	*Continuous and typically represents a quantity or measurement*	*Categorical and represents class labels or categories*
Prediction vs. Classification	*The model's output is a numerical value, and predictions are made based on these values*	*The model assigns data points to discrete categories or classes. Predictions are categorical, indicating the class to which a data point belongs, such as "Class A" or "Class B."*
Evaluation Metrics	*Mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and R-squared (coefficient of determination).*	*Accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC), depending on the problem and the class distribution.*
Decision Boundary	*There is no decision boundary. The model aims to find a continuous relationship between inputs and outputs.*	*The model learns a decision boundary that separates different classes in the feature space. This boundary is used to classify new data points.*
Examples	*Predicting sales revenue based on marketing spend, estimating the time it takes to complete a task, or forecasting the demand for a product.*	*Spam email detection, image classification, sentiment analysis (classifying text as positive, negative, or neutral), and medical diagnosis.*

Solution to most common problems in Machine Learning

Concept of overfitting

It occurs when a model is trained too well on the training data to the point that it learns not only the underlying patterns but also the noise and random fluctuations in the data. As a result, an overfit model performs exceptionally well on the training data but poorly on new, unseen data.

Concept of overgeneralization (underfitting)

It occurs when a model is too simplistic to capture the underlying patterns in the data, leading to poor performance not only on the training data but also on new, unseen data.

Characteristics of outliers

Outliers are data points that significantly differ from the majority of the data in a dataset, they can represent rare events, errors in data collection, or genuine anomalies in the dataset.

Outliers can skew summary statistics such as the mean and standard deviation. Whether a data point is considered an outlier can depend on the context of the analysis. In some cases, an extreme value might be an anomaly, while in others, it could be a legitimate data point.

Most common solutions for overfitting, overgeneralization, and outliers

Overfitting:

- Reduce model complexity by using fewer features, reducing the number of parameters, or selecting a simpler model architecture.
- Apply regularization techniques like L1 (Lasso) or L2 (Ridge) regularization to penalize large model weights and encourage simpler models.
- Increase the size of the training dataset if possible. More data can help the model generalize better and capture underlying patterns.

Overgeneralization:

- Carefully select and engineer relevant features that provide the model with more meaningful information about the problem.
- Regularization techniques can also be used to reduce underfitting. For example, L2 regularization can encourage the model to make use of all available features.
- Adjust hyperparameters (e.g., learning rate, model depth) to find a better balance between underfitting and overfitting.

Outliers:

- Identify outliers using statistical methods like z-scores, the IQR method, or visualization techniques such as box plots or scatter plots.
- In some cases, outliers can be removed from the dataset if they are deemed to be anomalies or data errors.
- Data transformation techniques like logarithmic or robust scaling can mitigate the influence of outliers.

Process of dimensionality reduction

Dimensionality reduction is a process used in data analysis and machine learning to reduce the number of features or dimensions in a dataset while preserving the essential information.

Start with a dataset that contains a high number of features or dimensions, identify any potential issues like missing values or outliers, ensure that the features are on a similar scale, as many dimensionality reduction techniques are sensitive to the scale of the data. Common scaling methods include standardization (z-score scaling) and min-max scaling.

There are two primary categories of dimensionality reduction techniques: feature selection and feature extraction.

Feature selection methods select a subset of the original features while discarding others. Common techniques include filter methods, wrapper methods, and embedded methods.

Feature extraction methods create new features that are linear or nonlinear combinations of the original features. These methods aim to preserve as much variance or information as possible in a lower-dimensional space. Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) are popular feature extraction techniques.

Depending on the chosen technique, apply dimensionality reduction to the dataset. For feature selection, this means selecting the most relevant features, while for feature extraction, it involves creating new feature representations.

Determine the number of components you want to retain in the reduced dataset. This depends on your specific goals and the trade-off between dimensionality reduction and information loss. Then apply the reduced data to your model.

Dimensionality problem

The dimensionality problem refers to the issues that arise as the number of dimensions in a dataset increases. As the number of dimensions increases, the amount of data required to adequately cover the feature space grows exponentially. High-dimensional data increases the computational cost of various algorithms, and it becomes increasingly difficult for humans to intuitively visualize or comprehend the data.

This can result in a situation where the effective sample size (i.e., the number of data points available for each dimension) becomes very small. This sample size can lead to unreliable statistical inferences.

Bias-variance trade-off

Bias refers to the error introduced by overly simplistic assumptions in the learning algorithm. A model with high bias fails to capture the underlying patterns and relationships on the data. High bias is often associated with a model that is too rigid or has too few parameters,

Variance refers to the error introduced by the model's sensitivity to fluctuations or noise in the training data. A model with high variance overfits the data, meaning it captures not only the underlying patterns but also the noise. High variance is often associated with a model that is too complex or has too many parameters

The bias-variance trade-off arises because as you try to reduce one source of error (bias or variance), you often increase the other. Techniques like cross-validation, grid search, and learning curves can help fine-tune model parameters and find the sweet spot that minimizes both bias and variance.