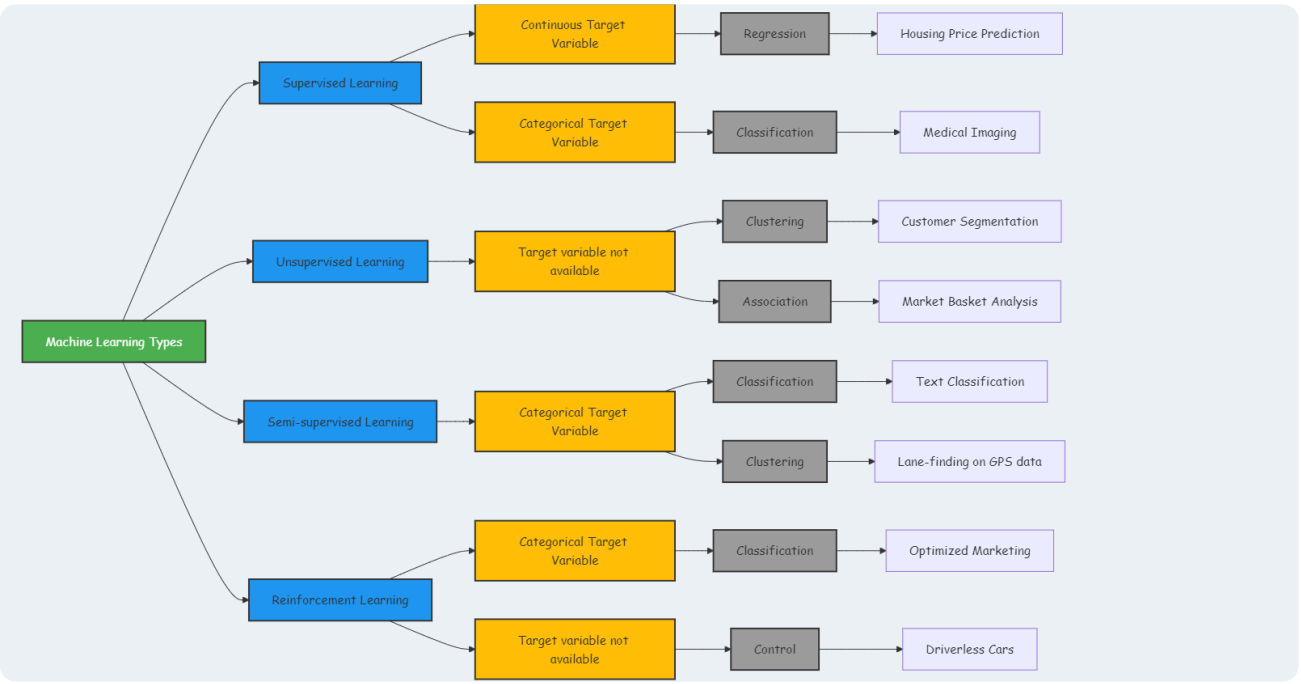
01 ML Basics - KirkYagami X NikhilSharma

Machine Learning:

https://developers.google.com/machine-learning/crash-course/linear-regression

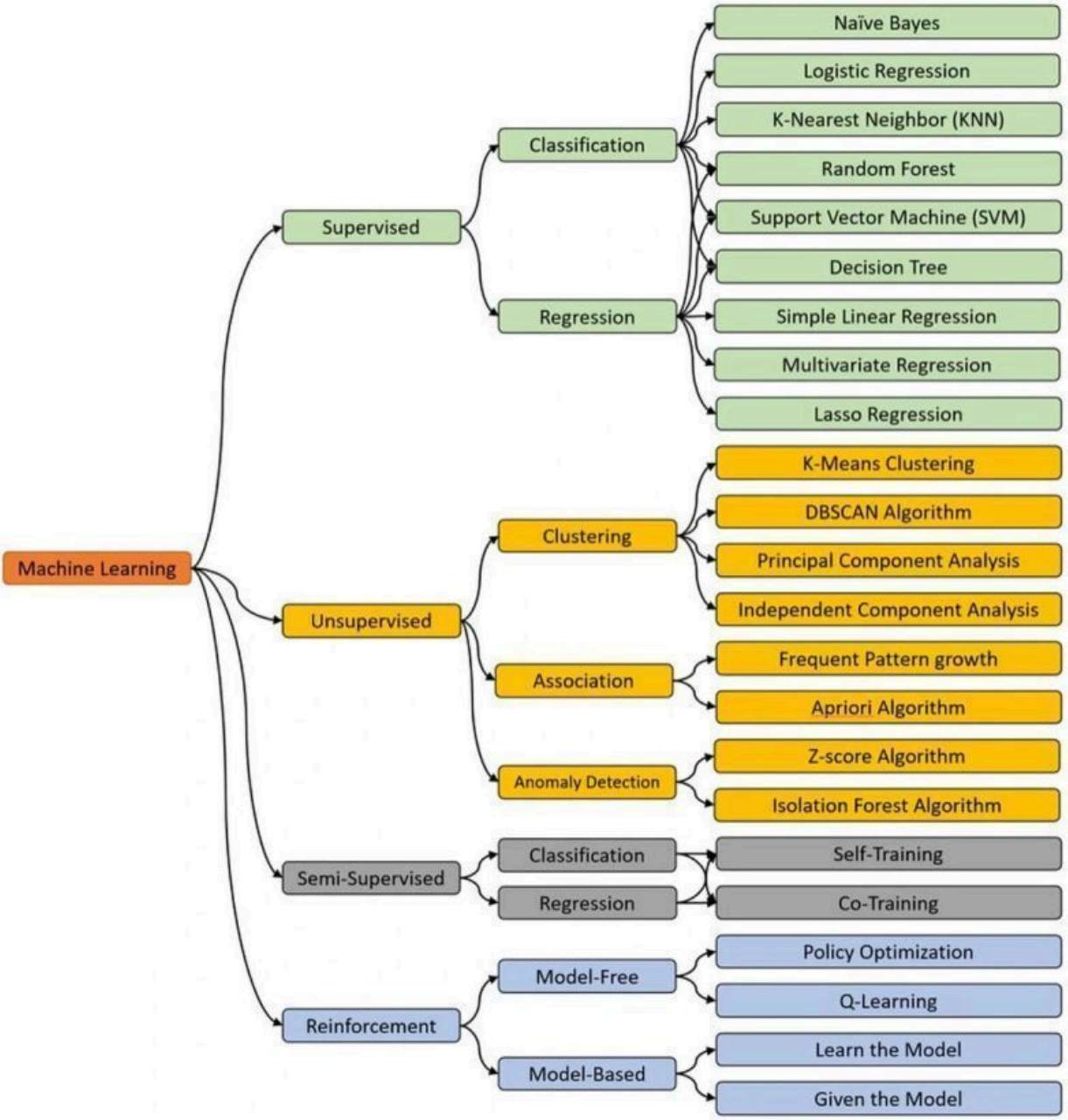
ML and Types:

https://cloud.google.com/learn/what-is-machine-learning?hl=en



Machine Learning Algorithms

(Every data scientist must know)



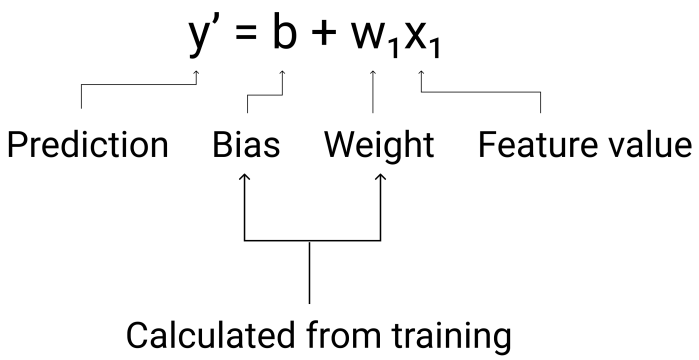
> **Self-supervised learning** consists of **training a model based on the inherent structure of the training data**.

Linear Regression

Core Concept

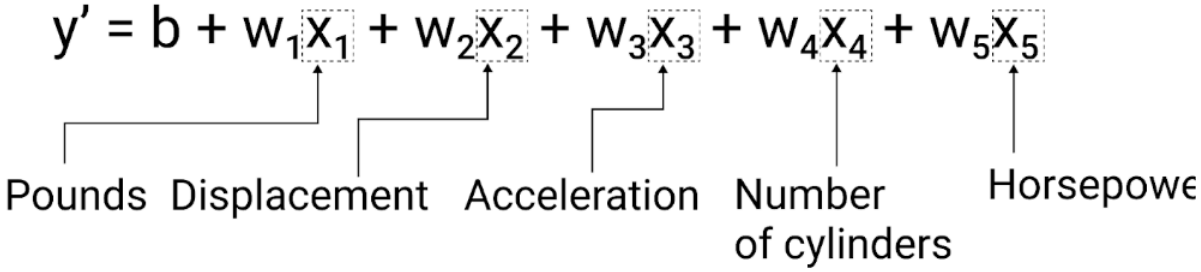
Linear regression establishes a mathematical relationship between a dependent variable (the outcome we want to predict) and one or more independent variables (the features we use to make predictions). At its heart, the model assumes this relationship can be approximated by a linear equation.

The simplest form, simple linear regression, models the relationship between two variables with a straight line:



For example, a model that predicts gas mileage could additionally use features such as the following:

- Engine displacement
- Acceleration
- Number of cylinders
- Horsepower



A model with five features to predict a car's miles per gallon rating.

What is linear regression?

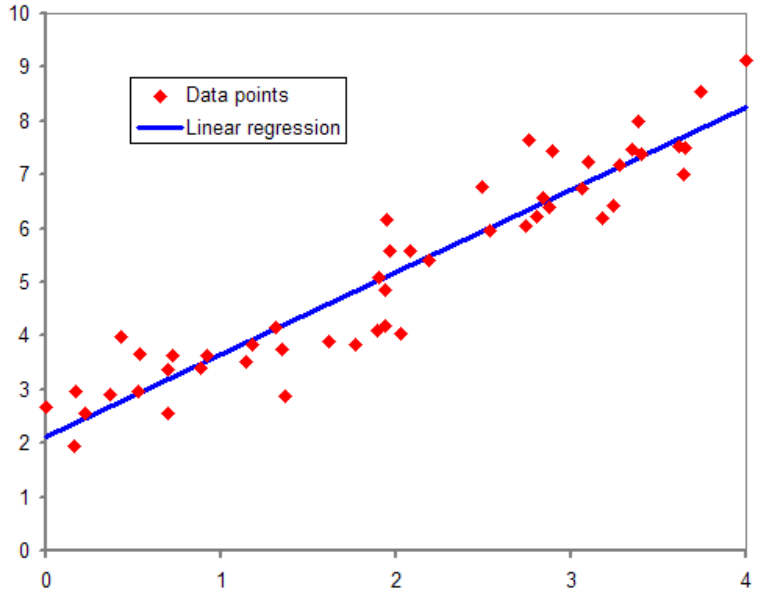
Linear regression is a statistical analysis technique that models the linear relationship between one independent variable and one dependent variable. It predicts this relationship by fitting a linear equation to given data.
Linear regression is the simplest form of regression, and can only model relationships between two variables.

What is a regression line?

A regression line is a straight line used in linear regression to indicate a linear relationship between one independent variable (on the x-axis) and one dependent variable (on the y-axis).
Regression lines may be used to predict the value of Y for a given value of X.

what is line of best fit in regression analysis?

In regression analysis, the line of best fit, also known as the regression line or trend line, is a straight line that best approximates the relationship between two variables on a scatter plot. It minimizes the overall distance between the line and the data points, offering a visual representation of the data's general trend and allowing for predictions.



Key Assumptions

1. **Linearity**: The relationship between independent and dependent variables is linear
2. **Independence**: Observations are independent of each other
3. **Homoscedasticity**: The variance of residuals is constant across all levels of independent variables
4. **Normality**: The residuals follow a normal distribution
5. **No multicollinearity**: For multiple regression, the independent variables should not be highly correlated

How good is the model?

`R2` also called as **coefficient of determination** summarizes the explanatory power of the regression model and is computed from the sums-of-squares terms.

$$\textit{Coefficient of Determination} \rightarrow \quad R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

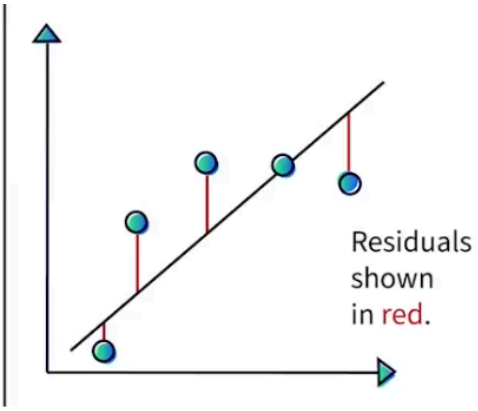
$$\textit{Sum of Squares Total} \rightarrow \quad SST = \sum (y - \bar{y})^2$$

$$\textit{Sum of Squares Regression} \rightarrow \quad SSR = \sum (y' - \bar{y'})^2$$

$$\textit{Sum of Squares Error} \rightarrow \quad SSE = \sum (y - y')^2$$

R^2 describes the proportion of variance of the dependent variable explained by the regression model. If the regression model is "perfect", SSE is zero, and R2 is 1. If the regression model is a total failure, SSE is equal to SST, no variance is explained by regression, and R2 is zero. It is important to keep in mind that there is no direct relationship between high R2 and causation.

- Distances between the actual and predicted values are called residuals.
- Residuals are key to determining the performance of a regression model.



Regression Metrics

- R^2 is more of a relative measure
- Mean square error and root mean square error are absolute measures
- MSE is an absolute number of how much your predicted results deviate from the actual number
- Root mean square error is the square root of MSE

Real-World Examples

1. **Housing Price Prediction**: A real estate company can predict house prices based on features like square footage, number of bedrooms, location, age of the building, etc. For example, in Boston's housing market, each additional square foot might increase a home's value by approximately $300, while an additional bedroom might add $15,000 to the price, controlling for other factors.
2. **Healthcare Cost Prediction**: Hospital systems can model patient care costs based on factors like length of stay, procedures performed, patient age, and comorbidities. A linear regression model might determine that each additional day in the hospital adds approximately $2,000 to a patient's bill, while certain procedures like an MRI add fixed costs.
3. **Crop Yield Forecasting**: Agricultural scientists can predict crop yields based on factors like rainfall, temperature, soil quality, and fertilizer use. A model might show that each inch of rainfall during the growing season increases corn yield by 8 bushels per acre, while each degree increase in average temperature above a threshold decreases yield by 3 bushels per acre.

Sources To Refer:

1. https://medium.com/analytics-vidhya/everything-you-need-to-know-about-linear-regression-750a69a0ea50
2. https://medium.com/@msong507/linear-regression-fundamentals-489e6d60d5cc

Logistic Regression

> Classification is the task of predicting which of a set of classes (categories) an example belongs to.

> It's a classification algorithm, that is used where the response variable is *categorical*. The idea of Logistic Regression is to find a **relationship between features and probability of particular outcome**==.

Logistic regression is a supervised machine learning algorithm widely used for classification. We use logistic regression to predict a binary outcome (1/0, Yes/No, True/False) given a set of independent variables.

What is logistic regression in simple terms?

Logistic regression is a statistical model that estimates how likely a binary outcome will occur, such as in yes/no or true/false scenarios, based on analyzing previous variable data.
Since logistic regression determines a probability, the dependent variable in this model will always be a value between 0 and 1.

What is the difference between linear regression and logistic regression?

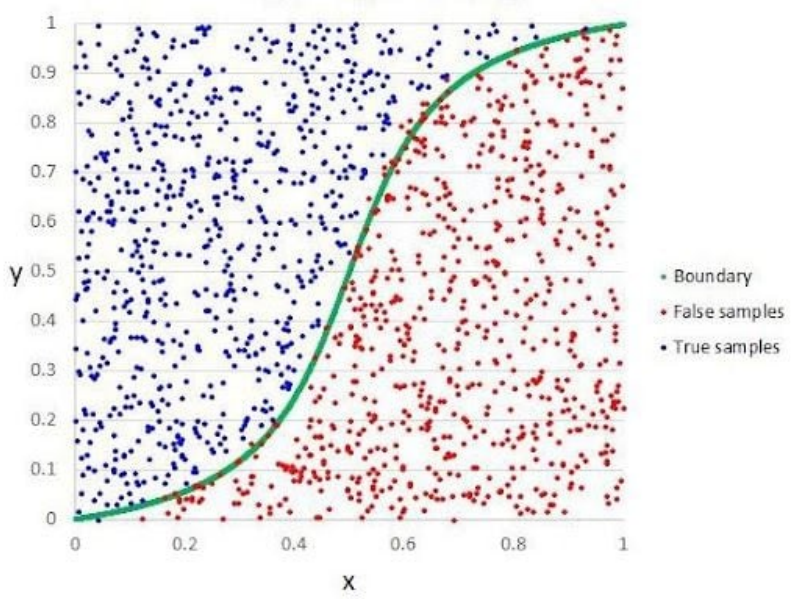
Linear regression can model variable outcomes on a continuous scale, while logistic regression can only model variable outcomes on a discrete or categorical scale.

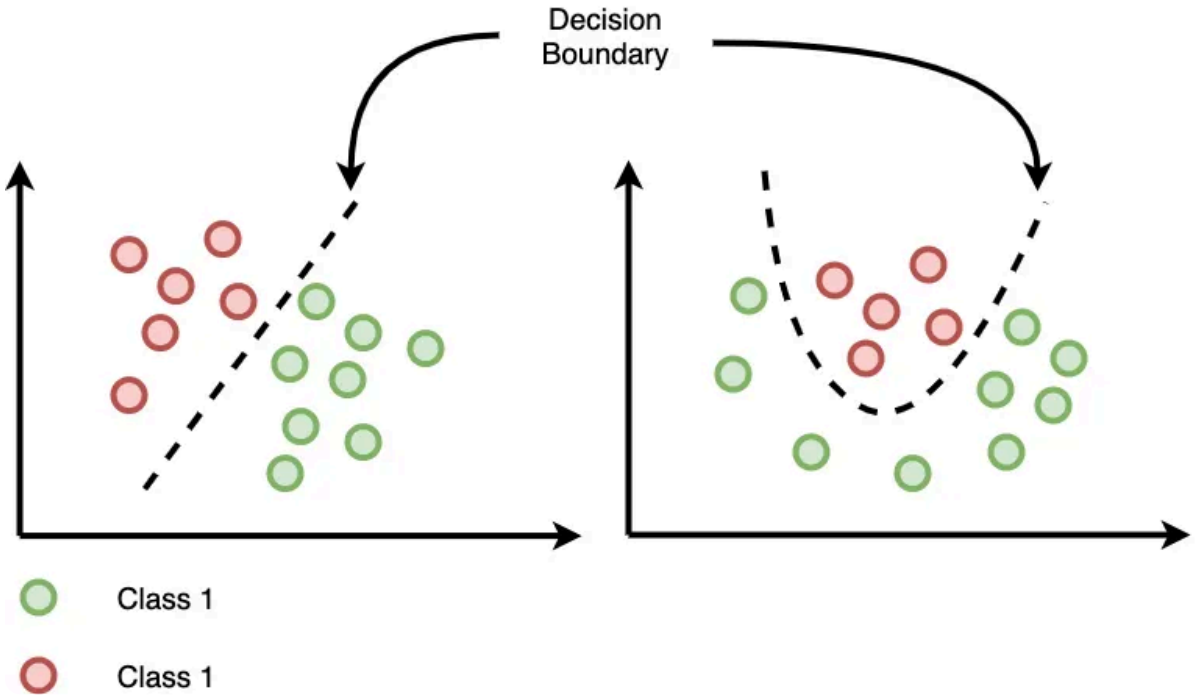
What is logistic regression best used for?

Logistic regression is best used for classification and prediction problems in machine learning. It can be applied to help identify fraud, determine disease likelihood or predict user behavior on websites and apps.

Decision Boundary in Logistic Regression

To predict the class to which data belongs, you can set a threshold which we call the decision boundary. Based upon this threshold, we classify the obtained estimated probability into different classes. Say, if `predicted_value ≥ 0.5`, then classify email as spam else as not spam.
Decision boundaries can be linear or nonlinear. You can also increase the polynomial order to get a complex decision boundary.





In the above diagram, the dashed line can be identified as the decision boundary since we will observe instances of a different class on each side of the boundary. Our intention in logistic regression would be to decide on a proper fit to the decision boundary so that we will be able to predict which class a new feature set might correspond to. The interesting fact about logistic regression is the utilization of the sigmoid function as the target class estimator.

Core Concept

Despite its name, logistic regression is a classification algorithm, not a regression technique. It predicts the probability that an instance belongs to a particular class. The algorithm uses the logistic function (sigmoid) to constrain the output to values between 0 and 1, representing probabilities.
The logistic function:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Where z is the linear combination of features:

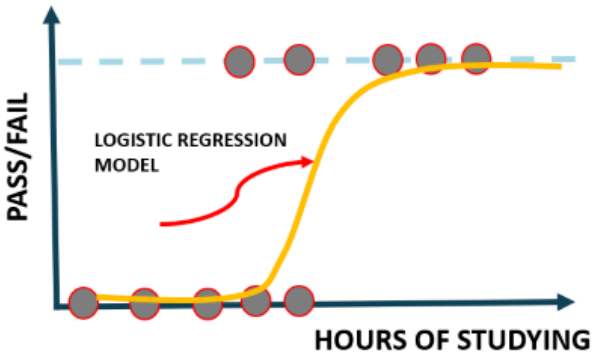
$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

The probability model becomes:

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)}}$$

LOGISTIC REGRESSION: MATH

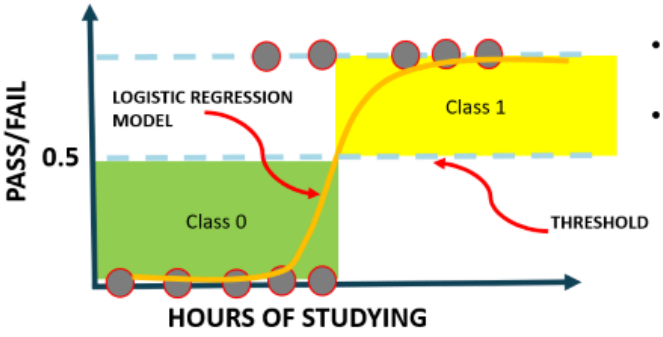
- Linear regression is not suitable for classification problem.
- Linear regression is unbounded, so logistic regression will be better candidate in which the output value ranges from 0 to 1.



- Linear equation:
  - $y = b_0 + b_1 * x$
- Apply Sigmoid function:
  - $P(x) = sigmoid\ (y)$
  - $P(x) = \frac{1}{1+e^{-y}}$
  - $P(x) = \frac{1}{1+e^{-(b_0+b_1*x)}}$

LOGISTIC REGRESSION: FROM PROBABILITY TO CLASS

- Now we need to convert from a probability to a class value which is "0" or "1".



- Linear equation:
  - $y = b_0 + b_1 * x$
- Apply Sigmoid function:
  - $P(x) = sigmoid\ (y)$
  - $P(x) = \frac{1}{1+e^{-y}}$

Evaluation Metrics

1. **Confusion Matrix Components:**
   - True Positives (TP): Correctly predicted positive cases
   - True Negatives (TN): Correctly predicted negative cases
   - False Positives (FP): Incorrectly predicted positive cases (Type I error)
   - False Negatives (FN): Incorrectly predicted negative cases (Type II error)
2. **Accuracy:**

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

3. **Precision:**

$$\text{Precision} = \frac{TP}{TP + FP}$$

4. **Recall (Sensitivity)**:

$$\text{Recall} = \frac{TP}{TP + FN}$$

5. **F1 Score**:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

6. **ROC Curve and AUC**: Plots true positive rate against false positive rate across different thresholds

Real-World Examples

1. **Credit Default Prediction**: Financial institutions use logistic regression to predict the probability of loan default based on factors like credit score, income, debt-to-income ratio, and payment history. For instance, a model might indicate that each 50-point decrease in credit score doubles the odds of default, while a 10% increase in debt-to-income ratio increases default odds by 30%.
2. **Medical Diagnosis**: Doctors can use logistic regression to assess the probability of disease based on patient symptoms, test results, and demographics. For example, a model for heart disease might show that smoking increases the odds ratio by 2.5, while each 10-point increase in systolic blood pressure increases the odds by 1.4, controlling for other risk factors.
3. **Customer Conversion Prediction**: A digital marketing team can predict whether website visitors will convert to customers based on features like time spent on site, pages visited, referral source, and demographic information. The model might reveal that visitors who spend more than 3 minutes on the site are twice as likely to convert, and those who visit the pricing page are 3.5 times more likely to become customers.

Sources To Refer:

1. https://medium.com/analytics-vidhya/a-comprehensive-guide-to-logistic-regression-e0cf04fe738c

---

Classification

Core Concept

Classification is a supervised learning technique where the model learns to assign data points to predefined categories (classes) based on their features. Using labeled training examples, the algorithm learns decision boundaries to predict the class labels of new, unseen instances.

Real-World Examples

1. **Email Spam Detection**: Email providers use classification algorithms to separate spam from legitimate emails based on features like sender information, content keywords, HTML structures, and header patterns. A Naive Bayes classifier might learn that emails containing phrases like "limited offer," "act now," and "free money" in combination with image-heavy content and links to unfamiliar domains have a 95% probability of being spam.
2. **Medical Image Analysis**: Healthcare systems can use classification models to identify conditions in medical images. For example, a convolutional neural network trained on a large dataset of chest X-rays might classify images as showing pneumonia, tuberculosis, normal lungs, or other conditions

with accuracy approaching that of radiologists. The model learns to recognize patterns such as lung opacity patterns, infiltrates, and consolidation that distinguish different conditions.

3. **Fraud Detection in Financial Transactions**: Banks apply classification models to flag potentially fraudulent transactions based on features like transaction amount, location, time, merchant category, and deviation from typical spending patterns. A Random Forest model might learn that transactions that occur in a different country than the cardholder's residence, for high-value amounts, at unusual hours, and at merchants never previously visited have a high probability of being fraudulent.

Evaluation Metrics

Regression Metrics

1. Mean Absolute Error (MAE)

Measures the average magnitude of errors without considering their direction.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Properties:

- Scale-dependent (in same units as target variable)
- Less sensitive to outliers than MSE
- All errors weighted equally
- Measures central tendency of the error

2. Mean Squared Error (MSE)

Average of squared differences between predicted and actual values.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Properties:

- Penalizes larger errors more heavily due to squaring
- Scale-dependent (in squared units)
- More sensitive to outliers
- Differentiable everywhere, making it useful for optimization

3. Root Mean Squared Error (RMSE)

Square root of MSE, returning the error to the original scale of the target variable.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Properties:

- Same units as the target variable
- Easier to interpret than MSE
- Still penalizes large errors more than MAE

4. Coefficient of Determination (R^2)

Proportion of variance in the dependent variable that is predictable from the independent variables.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Where $\bar{y}$ is the mean of the observed values.

Properties:

- Scale-independent (always between 0 and 1, or can be negative for very poor models)
- Represents the proportion of explained variance
- Higher values indicate better fit (1 is perfect)
- Can be misleading for models with many variables or non-linear relationships

5. Adjusted R^2

Modified version of R^2 that adjusts for the number of predictors in the model.

$$\text{Adjusted } R^2 = 1 - \frac{(1 - R^2)(n - 1)}{n - p - 1}$$

Where n is the number of samples and p is the number of predictors.

Properties:

- Penalizes adding unnecessary variables
- More reliable for comparing models with different numbers of predictors
- Can decrease with the addition of irrelevant features

Classification Metrics

1. Confusion Matrix

A table showing predicted vs. actual class counts across all classes:

.	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

2. Accuracy

Proportion of all predictions that were correct.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Properties:

- Easy to understand
- Works well for balanced classes
- Can be misleading for imbalanced classes
- Range: 0 to 1 (higher is better)

3. Precision

Proportion of positive identifications that were actually correct.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Properties:

- Focuses on minimizing false positives
- Important when false positives are costly
- Range: 0 to 1 (higher is better)
- Also called Positive Predictive Value (PPV)

4. Recall (Sensitivity or True Positive Rate)

Proportion of actual positives that were correctly identified.

$$\text{Recall} = \frac{TP}{TP + FN}$$

Properties:

- Focuses on minimizing false negatives
- Important when false negatives are costly
- Range: 0 to 1 (higher is better)
- Also called Sensitivity or True Positive Rate

5. F1 Score

Harmonic mean of precision and recall.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Properties:

- Balances precision and recall
- Single metric that works well for imbalanced classes
- Range: 0 to 1 (higher is better)
- Gives more weight to lower values

6. Area Under the ROC Curve (AUC-ROC)

Measures the area under the curve plotting true positive rate against false positive rate at various threshold settings.

$$\text{AUC} = \int_0^1 TPR(FPR^{-1}(t))dt$$

Where:

- TPR is True Positive Rate (Recall): $TPR = \frac{TP}{TP+FN}$
- FPR is False Positive Rate: $FPR = \frac{FP}{FP+TN}$

Properties:

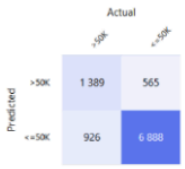
- Threshold-independent evaluation
- Range: 0 to 1 (0.5 = random classifier, higher is better)
- Represents probability that a random positive instance ranks higher than a random negative instance
- Less sensitive to class imbalance than accuracy

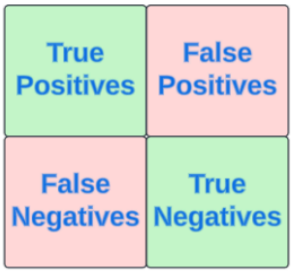
Detailed Analysis of Confusion Matrix:

Two-class Logistic regression

Confusion Matrix

This table gives us a summary of the results when it comes to what the model was supposed to predict and the actual prediction.





Accuracy	0.847
Precision	0.711
Recall	0.6
F1 Score	0.651
AUC	0.901

Accuracy - This tells how often the classifier is right in predicting results.

$$\frac{\text{True Positives + True Negatives}}{\text{True Positives + True Negatives + False Positives + False Negatives}}$$

Precision - This tells to what extend does the model accurately predict results.

Confusion Matrix

The confusion matrix shows:

.	Predicted Positive	Predicted Negative
Actual Positive	True Positives: 1,389	False Negatives: 926
Actual Negative	False Positives: 565	True Negatives: 6,888

h

This gives us the following totals:

- Total examples: 9,768 (1,389 + 926 + 565 + 6,888)
- Actual positives: 2,315 (1,389 + 926)
- Actual negatives: 7,453 (565 + 6,888)
- Predicted positives: 1,954 (1,389 + 565)
- Predicted negatives: 7,814 (926 + 6,888)

Metrics Calculation and Interpretation

1. Accuracy = 0.847

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{1,389 + 6,888}{9,768} = 0.847$$

Interpretation: The model correctly classifies 84.7% of all instances. While this number seems high, it can be misleading if the classes are imbalanced, which they are in this case (many more negative examples than positive).

2. Precision = 0.711

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{1,389}{1,389 + 565} = \frac{1,389}{1,954} = 0.711$$

Interpretation: When the model predicts the positive class, it is correct 71.1% of the time. In other words, about 29% of positive predictions are false alarms.

3. Recall = 0.6

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{1,389}{1,389 + 926} = \frac{1,389}{2,315} = 0.6$$

Interpretation: The model captures only 60% of all actual positive cases. This is particularly important in contexts where missing positive cases is costly (e.g., disease diagnosis, fraud detection).

4. F1 Score = 0.651

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = 2 \times \frac{0.711 \times 0.6}{0.711 + 0.6} = 2 \times \frac{0.4266}{1.311} = 0.651$$

Interpretation: The F1 score provides a balance between precision and recall. At 0.651, it indicates moderate performance when both false positives and false negatives are considered equally important.

5. AUC = 0.901

The Area Under the ROC Curve of 0.901 indicates that the model has excellent discriminative ability.

Interpretation: There is a 90.1% probability that the model will rank a randomly selected positive instance higher than a randomly selected negative instance. This high AUC suggests that the model is good at distinguishing between the two classes, even if its recall is somewhat low at the current threshold.

Overall Model Assessment

The model demonstrates good overall accuracy (84.7%) and excellent discrimination capability (AUC = 0.901). However, the recall of 0.6 indicates that it misses 40% of the actual positive cases.

This pattern suggests a model that:

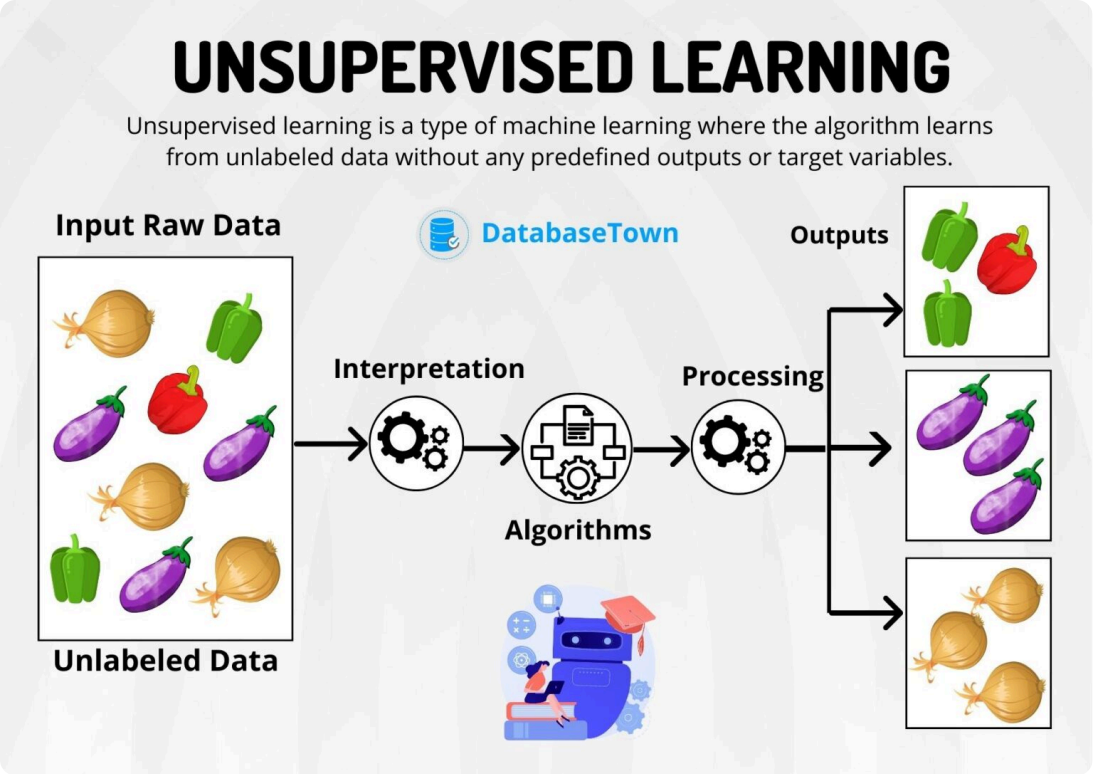
1. Is relatively conservative in making positive predictions (high precision)
2. Misses a substantial portion of positive cases (lower recall)
3. Works with imbalanced data (more negative than positive cases)

Depending on the application, the threshold for classification could be adjusted:

- If missing positive cases is very costly, the threshold could be lowered to increase recall at the expense of precision
- If false positives are costly, the current threshold might be appropriate or could even be raised

The high AUC indicates that the model has the potential to perform well across different threshold settings, giving users flexibility in making this precision-recall tradeoff.

Unsupervised Learning



Applications of Unsupervised Learning

Unsupervised learning finds applications across various domains. Some notable applications include:

1. **Customer Segmentation:** Unsupervised learning algorithms can group customers based on their purchasing behavior, allowing businesses to tailor marketing strategies.
2. **Anomaly Detection:** By identifying abnormal patterns or outliers, unsupervised learning can help detect fraud, network intrusions, or manufacturing defects.
3. **Image and Text Clustering:** Unsupervised learning can automatically group similar images or texts, aiding in tasks like image organization, document clustering, or content recommendation.
4. **Genome Analysis:** Unsupervised learning algorithms can analyze genetic data to identify patterns and relationships, leading to insights in personalized medicine and genetic research.
5. Social Network Analysis: Unsupervised learning can be used to identify communities or influential individuals within social networks, enabling targeted marketing or detecting online communities.

Clustering

Core Concept

Clustering is an unsupervised learning technique that groups similar data points together based on their intrinsic properties. Unlike supervised learning, clustering doesn't use labeled data. Instead, it identifies natural groupings or patterns in data by maximizing within-cluster similarity and minimizing between-cluster similarity.

Key Algorithms

1. K-means Clustering

K-means partitions data into K distinct clusters by minimizing the within-cluster sum of squares:

$$\text{minimize} \sum_{i=1}^{k} \sum_{x \in S_i} |x - \mu_i|^2$$

Where:

- S_i is the set of points in cluster i
- μ_i is the centroid of cluster i
- $|x - \mu_i|^2$ is the squared Euclidean distance

Algorithm steps:

1. Select K initial centroids
2. Assign each data point to the nearest centroid
3. Update centroids based on the mean of assigned points
4. Repeat steps 2-3 until convergence or maximum iterations

Choosing K: Methods include the elbow method, silhouette analysis, and gap statistics

2. Hierarchical Clustering

Builds a tree of clusters (dendrogram) without requiring a pre-specified number of clusters.

Types:

- **Agglomerative** (bottom-up): Starts with each data point as a separate cluster and merges the closest pairs of clusters
- **Divisive** (top-down): Starts with all data in one cluster and recursively splits it

Linkage methods:

- **Single linkage**: Distance between closest elements
- **Complete linkage**: Distance between furthest elements
- **Average linkage**: Average distance between all pairs
- **Ward's method**: Minimizes variance within clusters

3. DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

Groups together points that are closely packed while marking points in low-density regions as outliers.

Key parameters:

- ϵ (eps): Maximum distance between two points to be considered neighbors
- MinPts: Minimum number of points required to form a dense region

Point classifications:

- **Core points**: Have at least MinPts points within distance ϵ
- **Border points**: Within distance ϵ of a core point but have fewer than MinPts neighbors
- **Noise points**: Neither core nor border points

Evaluation Metrics

1. **Silhouette Coefficient**:

$$s(i) = \frac{b(i) - a(i)}{\max a(i), b(i)}$$

   Where $a(i)$ is the average distance to points in the same cluster, and $b(i)$ is the average distance to points in the nearest different cluster.

2. **Davies-Bouldin Index**: Measures average similarity between clusters
3. **Calinski-Harabasz Index**: Ratio of between-cluster dispersion to within-cluster dispersion

4. **Dunn Index**: Ratio of smallest inter-cluster distance to largest intra-cluster distance

Real-World Examples

1. **Customer Segmentation**: A retail company can cluster customers based on purchasing behavior (frequency, monetary value, recency), demographics, and browsing patterns. For example, analysis might reveal five natural customer segments: "High-Value Loyalists" who purchase frequently and spend significantly; "Discount Hunters" who only purchase during sales; "New Enthusiasts" who recently began shopping but show strong engagement; "Occasional Buyers" with infrequent patterns; and "At-Risk Customers" whose engagement is declining.
2. **Anomaly Detection in Network Security**: IT security teams can use clustering to identify abnormal network traffic patterns that might indicate security breaches. DBSCAN might be particularly useful here as it can identify outliers (potential security threats) that don't fit within normal traffic clusters. For instance, a sudden cluster of high-volume data transfers to unusual international IP addresses outside working hours might be flagged as potential data exfiltration.
3. **Document Classification**: A news organization can automatically group articles into topics without predefined categories. The clustering algorithm might naturally discover categories like "Politics," "Sports," "Entertainment," and "Business" based on word frequencies and semantic similarities in the text. This can help organize large document archives or enable content discovery features for readers.