Notes by [Izam Mohammed]

Classification

It is categorizing data into predefined classes or categories based on their features.

Terms

- *Classes: The distinct labels or target classes that the data points are assigned to*
- *Label: A class or category assigned to a data point*
- *Decision Boundary: a boundary that separates different classes in a feature space*
- *Binary Classification: A classification problem with only 2 classes*
- *Multiclass Classification: A classification problem with more than two classes*
- *Multilabel Classification: A classification problem where each data point can belong to multiple classes simultaneously*
- *Imbalanced dataset: A dataset where one class significantly outnumbers the others.*

Evaluation

- *Accuracy* -> $(TP + TN) / (TP + TN + FP + FN)$
- *Precision* -> $TP / (TP + FP)$
- *Recall - Sensitivity, True Positive Rate* -> $TP / (TP + FN)$
- *FPR - False Positive Rate* -> $FP / (FP + TN)$
- *Specificity - True Negative Rate* -> $TN / (FP + TN)$
- *F1 - score* -> $2 * (Precision * Recall) / (Precision + Recall)$
- *Confusion Metrix*
- *ROC curve - Receiver Operating Characteristic*
- *ROC AUC - ROC Area Under Curve*
- *PR AUC - Area Under the Precision-Recall curve*
- *MCC - Mathew's Correlation Coefficient* -> $(TP * TN - FP * FN) / \sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}$
- *Log loss - Binary Cross Entropy* -> $-(1/n) * \Sigma (y_i * \log(p_i) + (1 - y_i) * \log(1 - p_i))$
- *Categorical Cross entropy* -> $-\Sigma (y_i * \log(p_i))$
- *Hinge Loss* -> $\max(0, 1 - y_i * f(x_i))$

Algorithms

- *Logistic Regression*
  - *Logistic Function (Sigmoid)*
  - *Maximum Likelihood estimation*
  - *Odds*
  - *Odds ratio*
  - *Logit*
  - *Newton-Raphson Method*
  - *Regularized Logistic Regression (Ridge, Lasso)*
  - *Confidence intervals*
  - *Coefficients*
  - *Threshold*

- *Link Function*
- *Regularization*
- *Multinomial Logistic Regression*
- *Ordinary Regression*
- *Log Likelihood*
- *Deviance*
- *Deviance Residuals*
- *Akaike Information Criterion (AIC)*
- *Wald Test*
- *C- Static (Concordance Static)*
- *ROC and AUC*
- *Penalty*

- *KNN*
    - *Nearest Neighbors*
    - *Distance Metric (Euclidean, Manhattan, Minkowski, Mahalanobis)*
    - *k-Value*
    - *Majority Vote*
    - *Lazy Learning*
    - *Curse of Dimensionality*
    - *Weighted k-NN*
    - *Data Scaling*
    - *Local vs Global Behaviour*
    - *Decision Boundary*
    - *Parzen Window*
    - *Local Outlier Factor(LOF)*
    - *Radius Neighbors Classifier*
    - *Ball Tree and KD-Tree*
    - *Distance Weighting*
    - *k-D tree*
    - *Cover Tree*
    - *Voronoi Tessellation*
    - *Reverse Nearest Neighbors*
    - *Dynamic Time Warping (DTW)*
    - *Instance-based learning*

- *Naive Bayes*
    - *Bayes' Theorem*
    - *Conditional Independence*
    - *Prior Probability*
    - *Posterior Probability*
    - *Probability Distribution (Gaussian, Multinomial)*
    - *Laplace Smoothing (Adaptive Smoothing)*
    - *Text Classification*
    - *Spam Detection*
    - *Bag of Words*
    - *Feature Independence Assumption*
    - *Maximum Likelihood Estimation (MLE)*

- *Bayesian Inference*
- *Text Mining*
- *Laplace Estimation*
- *Log-odd ratio*
- *Bayesian Network*
- *Class Conditional Independence*
- *Bernoulli Naive Bayes*
- *Gaussian Naive Bayes*
- *Multinomial Naive Bayes*
- *MAP Estimation (Maximum A Posteriori)*
- *Bayesian Information Criterion (BIC)*

Parameter

- *Smoothing Parameter (Alpha): Controls additive smoothing (Laplace smoothing).*
- *Fit Prior: Whether to learn class prior probabilities.*
- *Class Prior: User-specified class prior probabilities.*

Logistic Regression

Advantages

- *Simple and interpretable*
- *Works well in small sample sets*
- *Can handle both binary and multiclass problems*
- *Robust to outliers*

Disadvantages

- *Assumes a linearity*
- *It does not capture complex relations*
- *Multicollinearity*
- *Vulnerable to overfitting*

Types

- *Binary Logistic regression*
- *Multinomial logistic regression*
- *Ordinal logistic regression: used when the dependent variable is ordinal*
- *Penalized Logistic regression: Lasso and ridge*
- *Logistic regression with interaction terms: Add interaction terms to the independent variables. Eg:- Multiply 2 columns*

Terms

- *Logistic Function (Sigmoid) : f(x) = 1 / (1 + exp(-x))*
- *Maximum Likelihood estimation: Find the best fitting model parameters that maximize the likelihood of the observed data.*
- *Odds: the likelihood of an event occurring compared to the likelihood of it not occurring. In binary outcome - odds= P(Y=1) / (1- P(Y-1))*

- *Odds ratio: Measure of the change in the odds of an event occurring due to the one-unit change in a predictor variable*
- *Odds Radio = exp (Coefficient of the predictor variable)*
- *Logit: Natural logarithm (log base e) of the odds of the dependent variable*
  logit (p) = ln (p / (1 - p))
  Where p is the probability of the binary outcome 1 and 1-p is the binary outcome 0
- *Newton-Raphson Method: It is an iterative optimization technique used to find the maximum likelihood estimates of the parameters in logistic regression*
- *Regularized Logistic Regression (Ridge, Lasso): In ridge and lasso logistic regression, it adds a penalty term to the logistic regression cost function that discourages large values of coefficients*
- *Confidence intervals: Provides a range of values within which we can be reasonably confident that the true parameter values lie.*
  Confidence Interval=Estimated Parameter±(Critical Value×Standard Error)
- *Coefficients: represent the strength and direction of the relation between predictor variables and the log odds of the binary outcome*
- *Threshold: It is the decision boundary, usually considered as 0.5*
- *Link Function: connects the linear combination of predictor variables. Usually use sigmoid in logistic regression*
- *Regularization: It is adding a penalty to prevent overfitting. Ridge adds an L2 penalty (squared) and lasso adds an L1 penalty (absolute) to the cost function.*
- *Multinomial Logistic Regression: Allows more than 2 categories in the dependent variable*
- *Ordinary Regression: refers to linear regression*
- *Log Likelihood: yi . ln (pi) + (1 - yi) . ln(1 - pi)*
  Where yi is the actual outcome and pi is the predicted probability
- *Deviance: Measure the goodness of fit in logistic regression*
  Deviance = - 2 [ln (L full_mode) - ln(L null_model)]
  L refers to the log likely hood
  The null model is intercept-only, The probability of the outcome being 1 in the null model is the overall proportion of 1s in the data
- *Deviance Residuals: the difference between the observed outcome and the predicted outcome based on the model*
  Deviance Residuali = 2 [yi ln (yi / pi) + (1 - yi) ln ((1 - yi) / (1 - pi))]
  Where yi is the observed outcome for the ith observation (0 or 1)
  pi is the predicted probability of the outcome being 1 in the observation
- *Akaike Information Criterion (AIC): Model Selection criterion used to compare different models. Lower AIC indicates a better trade-off between model fit*
  AIC = - 2 x Log likelihood + 2 x Number of parameters
  Where the number of parameters refers to the intercept term + number of coefficients
- *Wald Test: Used to assess the significance of individual coefficients in logistic regression*
  W = ((specific coeff - Null hypothesis) / Standard Error)
  Where the Null hypothesis will usually be 0
- *C- Static (Concordance Static): Associated with ROC*
- *ROC and AUC: ROC stands for (Reciever operating Characteristic), it is a graphical tool for evaluating the performance of binary classification models*

AUC refers to the Area Under the Curve of ROC. It quantifies the model's ability to discriminate between the 2 outcome classes

- *Penalty: Refers to a regularization term added into the cost function*

Hyperparameters

- *Penalty (Regularization Type): L1 or L2 regularization.*
- *C (Inverse of Regularization Strength): Controls the trade-off between fitting the training data and preventing overfitting.*
- *Solver: Optimization algorithm (e.g., 'liblinear', 'newton-cg', 'sag', 'lbfgs').*
    - *Liblinear (Library for Large Linear Classification): suited for binary classification*
    - *Newton-cg (Newton-Conjugate Gradient): uses the Newton-Raphson method, well suited for multi-class problems*
    - *Sag(Stochastic Average Gradient): designed for large datasets, uses SGD*
    - *Lbfgs (Limited-memory Broyden–Fletcher–Goldfarb–Shanno): Works well with many features. It is also a memory solver*
- *Multi_class: Specifies the strategy for multiclass classification ('ovr' or 'multinomial').*
- *Class Weight: Optional weights for classes.*
- *Max Iterations: Maximum number of iterations for the solver.*
- *Dual: Formulation ('True' or 'False') for the dual problem*

KNN

Advantages

- *Simple to understand and implement*
- *No training period*
- *No assumption about data distribution*
- *Robust to noisy data*
- *Effective multiclass classification*

Disadvantages

- *Sensitivity to feature scaling*
- *The optimal value for k*
- *Curse of dimensionality*
- *Not suitable for large dataset*

Terms

- *Nearest Neighbors: data points from the training dataset that are closest to the given input data point in features*
- *Distance Metric (Euclidean, Manhattan, Minkowski, Mahalanobis): Math formula used to measure the distance between 2 data points*
- *k-Value: Number of nearest neighbors*
- *Majority Vote: predicts by taking the majority vote from the class labels*
- *Lazy Learning: KNN is a lazy learner because it doesn't build a model during the training*
- *Curse of Dimensionality: challenges arise as the number of features increases*
- *Weighted k-NN: Assign different weights to the nearest neighbors while predicting*

- ➢ *Local vs Global Behaviour: KNN exhibits local behavior as it considers only a subset of nearby points to make predictions. In a decision tree, exhibit global behavior by considering all of them*
- ➢ *Decision Boundary: Boundary that separates different classes in feature space*
- ➢ *Parzen Window: Technique that can be used to estimate the probability density function, used for density-based classification or regression*
- ➢ *Local Outlier Factor(LOF): An algorithm used to detect anomalies and outliers*
- ➢ *Radius Neighbors Classifier: A variation of KNN where we only specify the radius instead of k neighbors*
- ➢ *Ball Tree and KD-Tree: Data structure can be used to increase the performance of KNN*
- ➢ *Distance Weighting: In weighted KNN, the contribution of each neighbor to the majority vote is weighted based on their distance.*
- ➢ *Cover Tree: A data structure design for KNN*
- ➢ *Voronoi Tessellation: divides the features into regions, each corresponding to different data points.*
- ➢ *Reverse Nearest Neighbors: Instead of finding the neighbor, find a data point in which the given data point is one of their nearest neighbors*
- ➢ *Dynamic Time Warping (DTW): It is a distance metric used in time series data*
- ➢ *Instance-based learning: KNN is an instance-based learning algorithm*

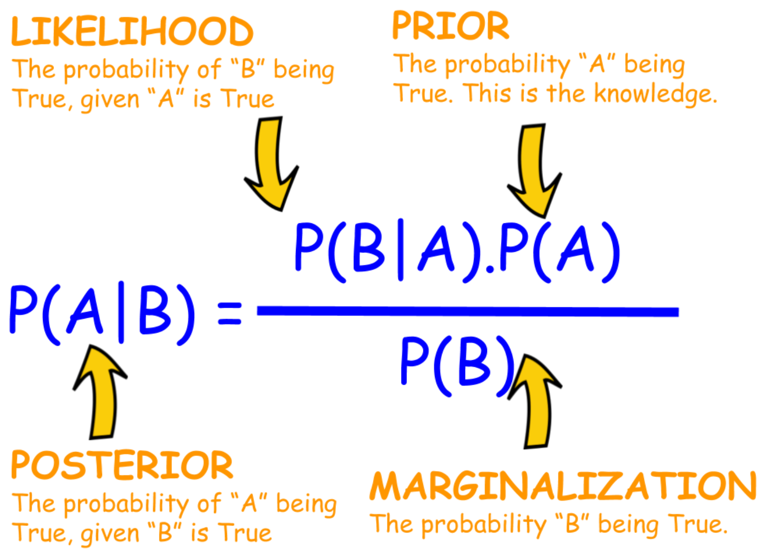
Hyperparameters

- ➢ *Number of Neighbors (k): Number of nearest neighbors to consider.*
- ➢ *Distance Metric: The distance metric used (e.g., Euclidean, Manhattan).*
  - *Euclidean (L2 Norm) :* $\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2 + \ldots}$
  - *Manhattan (L1 Norm):* $|(x_1 - x_2)| + |(y_1 - y_2)| + |(z_1 - z_2)| + \ldots$
  - *Minkowski:* $\sqrt[p]{(|x_1 - x_2|)^p + (|y_1 - y_2|)^p + (|z_1 - z_2|)^p + \ldots}$
  - *Mahalanobis:*
    $$\sqrt{(x_1 - x_2)^2 \cdot \Sigma_x^{-1} \cdot (x_1 - x_2) + (y_1 - y_2)^2 \cdot \Sigma_y^{-1} \cdot (y_1 - y_2) + \ldots}$$
- ➢ *Weighting: How to weight neighbors ('uniform' or 'distance').*
- ➢ *Algorithm: Algorithm used for nearest neighbor search (e.g., 'ball_tree', 'kd_tree', 'brute').*
- ➢ *Leaf Size (for 'ball_tree' or 'kd_tree'): Size of leaves in the data structure.*
- ➢ *P (for Minkowski distance): Power parameter for Minkowski distance.*

Types

- ➢ *Standardized KNN: Normal*
- ➢ *Weighted KNN: have higher weight for nearest value*
- ➢ *Radius Neighbour: check the radius of the point*
- ➢ *KNN for imbalance: modify the distance metric*

Naive Bayes



Advantages

- ➢ *Works well with a small dataset*
- ➢ *Handle high-dimensional data*
- ➢ *Low computation power compared to others*
- ➢ *Robust to irrelevant features*

Disadvantages

- ➢ *Sensitivity of data quality*
- ➢ *Zero probability problem*
- ➢ *Class Imbalance*
- ➢ *Ineffective for regression*

Types

- ➢ *Gaussian: Fon continuous*
- ➢ *Multinomial: for discrete*

Terms

- ➢ *Bayes' Theorem: Fundamental theorem for naive Bayes classifier*
- ➢ *Conditional Independence: Assumes features are independent of each other in the given class*
- ➢ *Prior Probability: The probability of a class occurring before considering any evidence.*
- ➢ *Posterior Probability: The probability of a class occurring after considering the evidence.*
- ➢ *Probability Distribution (Gaussian, Multinomial): types of probability distributions*
- ➢ *Laplace Smoothing (Adaptive Smoothing): A technique to handle "zero probability problem"*
- ➢ *Bag of Words: A representation of text data where the order of words is ignored.*
- ➢ *Maximum Likelihood Estimation (MLE): A method used to estimate probability in naive Bayes by counting the occurrences of events*

- *Text Mining: The process of extracting information from text data*
- *Laplace Estimation: Another name for Laplace Smoothing*
- *Log-odd ratio: A measure used in text classification*

$$\text{ODDS} = \frac{\text{Probability of winning}}{\text{Probability of losing}} = \frac{p}{1-p}$$

Where p = probability of winning (event occurring)

- *Bayesian Network: A graphical model representing probabilistic relationships among a set of variables*
- *Class Conditional Independence: Features are conditionally independent in the given class*
- *Bernoulli Naive Bayes: A variant of naive Bayes for binary data*
- *MAP Estimation (Maximum A Posteriori): An approach in Bayesian statistics that estimates the most likely values for model parameters.*
- *Bayesian Information Criterion (BIC): A model selection criterion that penalizes model complexity.*

Hyperparameters

- *Smoothing Parameter (Alpha): Controls additive smoothing (Laplace smoothing).*
- *Fit Prior: Whether to learn class prior probabilities.*
- *Class Prior: User-specified class prior probabilities.*