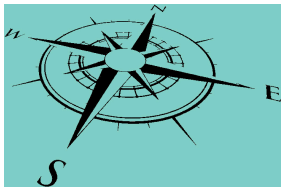
Evaluate Trustworthy AIKR objects implemented by machine learning powered services

This plan defines the role of the AI KR Strategist.

Contents



Artificial Intelligence Knowledge Representation Community Group (AIKR CG)

Stakeholder(s):

Carl Mattocks
Role: CoChair

Vision

For all AI systems to have clearly and transparently documented goals and performance data showing that they are being achieved.

Mission

The mission of an AI Strategist is to define the purpose and goals of AI systems, as well as the KPIs by which we can determine if the system is meeting its goals.

1. Ethical

Ensure AI Systems adhere to pivotal principles, such as, confidentiality, autonomy, accountability and veracity

Performance Indicators

Description	Type	Start Date	End Date
	Target		
	Actual		

2. Machine Learning Evaluation

Evaluate machine learning models

Stakeholder(s)

Artificial Intelligence Knowledge Representation Community Group (AIKR CG)
Role: Community of Interest

1. Trustworthy

Provide the foundation for a trustworthy AIKR

Evaluation metrics are tied to machine learning tasks. Perhaps the easiest metric to interpret is the percent of estimates that differ from the true value by no more than X%.

Performance Indicators

Description	Type	Start Date	End Date
	Target		
	Actual		

2. Track

Track Classification Performance Indicators

Ontological Statement: Classification Accuracy is the ratio of number of correct class label predictions to the total number of input samples data. Ontological Statement: F1 Score measure the Harmonic Mean between precision and recall. The range for F1 Score is [0, 1]. It tells you how precise your classifier is (how many instances it classifies correctly), as well as how robust it is (it does not miss a significant number of instances).

Performance Indicators

6 AUC-ROC Curve

Description	Type	Start Date	End Date
	Target		
	Actual		

Ontological Statement: check performance of multi - class classification AUROC (Area Under the Receiver Operating Characteristics) curve.Ontological Statement: Area Under Curve(AUC) is one of the most widely used metrics for evaluation. It is used for binary classification problem. AUC of a classifier is equal to the probability that the classifier will rank a randomly chosen positive example higher than a randomly chosen negative example. True Positive Rate (Sensitivity) : True Positive Rate is defined as TP/ (FN+TP). True Positive Rate corresponds to the proportion of positive data points that are correctly considered as positive, with respect to all positive data points. False Positive Rate (Specificity) : False Positive Rate is defined as FP / (FP+TN). False Positive Rate corresponds to the proportion of negative data points that are mistakenly considered as positive, with respect to all negative data points.

5 Log-Loss

Description	Type	Start Date	End Date
	Target		
	Actual		

Ontological Statement: Logarithmic loss (related to cross-entropy) measures the performance of a classification model where the prediction input is a probability value between 0 and 1 - Log loss increases as the predicted probability diverges from the actual label Logarithmic Loss or Log Loss, works by penalising the false classifications. It works well for multi-class classification. When working with Log Loss, the classifier must assign probability to each class for all the samples. where, y_ij, indicates whether sample i belongs to class j or not p_ij, indicates the probability of sample i belonging to class j Log Loss has no upper bound and it exists on the range [0, ∞). Log Loss nearer to 0 indicates higher accuracy, whereas if the Log Loss is away from 0 then it indicates lower accuracy. In general, minimising Log Loss gives greater accuracy for the classifier.

2 Accuracy

Description	Type	Start Date	End Date
	Target		
	Actual		

Ontological Statement: Classification Rate or Accuracy is given by the relation: True Positives + True Negatives / All Instances (True & False Positives + True & False Negatives)

4 Per-class accuracy

Description	Type	Start Date	End Date
	Target		
	Actual		

3 Confusion Matrix

Description	Type	Start Date	End Date
	Target		
	Actual		

Ontological Statement: A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class (the types of errors being made) Types :

- True Positives : The cases in which we predicted YES and the actual output was also YES.
- True Negatives : The cases in which we predicted NO and the actual output was NO.
- False Positives : The cases in which we predicted YES and the actual output was NO.
- False Negatives : The cases in which we predicted NO and the actual output was YES. Accuracy for the matrix can be calculated by taking average of the values lying across the “main diagonal” Type StartDate EndDate Description Target Number of True Positives Target Number of False Positives Target Number of True Negatives Target Number of False Negatives Actual [To be determined]

11 "Almost correct" predictions

Description	Type	Start Date	End Date
	Target		
	Actual		

1 Precision Recall

Description	Type	Start Date	End Date
	Target		
	Actual		

Ontological Statement: Precision is the number of correct positive results divided by the number of positive results predicted by the classifier. Ontological Statement: Recall is the number of correct positive results divided by the number of all relevant samples (all samples that should have been identified as positive).

9 Regression Analysis

Description	Type	Start Date	End Date
	Target		
	Actual		

Root Mean Square Error (RMSE) Ontological Statement: Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are.

8 NDCG

Description	Type	Start Date	End Date
	Target		
	Actual		

Ontological Statement: Normalized discounted cumulative gain (DCG) is a measure of ranking quality. In information retrieval, DCG measures the usefulness, or gain, of a document based on its position in the result list.

10 Quantiles of Errors

Description	Type	Start Date	End Date
	Target		
	Actual		

Quantiles (or percentiles), which is the element of a set that is larger than half of the set, and smaller than the other half.

7 F-measure

Description	Type	Start Date	End Date
	Target		
	Actual		

F1 Score is the Harmonic Mean between precision and recall. Ontological Statement: F-measure represents both Precision and Recall it helps to have a measurement that represents both of them. F-measure is calculated using Harmonic Mean (in place of Arithmetic Mean). Ontological Statement: Mean Absolute

Error is the average of the difference between the Original Values and the Predicted Values. It gives us the measure of how far the predictions were from the actual output. Ontological Statement: Mean Squared Error(MSE) takes the average of the square of the difference between the original values and the predicted values.

3. Lawful

Ensure AI Systems comply with all applicable laws and regulations, such as, provision audit data defined by a governance operating model

Performance Indicators

Description	Type	Start Date	End Date
	Target		
	Actual		

4. Ontological Statements

Employ ontological statements when explaining AIKR object audit data, veracity facts and (human, social and technology) risk mitigation factors

Performance Indicators

Description	Type	Start Date	End Date
	Target		
	Actual		

5. Track

Track AIKR object performance outcome via KPI (Key Performance Indicator) based on supervised learning models measurements

Performance Indicators

Description	Type	Start Date	End Date
	Target		
	Actual		

6. Document

Document the vision, values, goals, objectives for one or more AIKR objects

Performance Indicators

Description	Type	Start Date	End Date
	Target		
	Actual		

7. Robust

Ensure AI Systems are designed to handle uncertainty and tolerate perturbation from a likely threat perspective, such as, design considerations incorporate human, social and technology risk factors

Performance Indicators

Description	Type	Start Date	End Date
	Target		
	Actual		

Administrative Information

Start Date: 2020-04-01
End Date:

Publication Date: **2020-04-14**
Source: https://www.stratnavapp.com/StratML/Part2/861566c8-e9be-4642-b52f-f673fa499f4e

Submitter:

Given Name: **Surname:**
Email: **Phone:**

Submitter_861566c8-e9be-4642-b52f-f673fa499f4e