Berkeley Engineering | BerkeleyHaas

# PROFESSIONAL CERTIFICATE IN MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

## Office Hour #12 with Matilde D'Amelio

June 9, 2022 at 9 pm UTC

# Practical Application 2

# Classification Models

A **supervised machine learning** algorithm is one that relies on labeled input data to learn a function that produces an appropriate output when given new unlabeled data.

*When we see a pig, we shout "pig!" When it's not a pig, we shout "no, not pig!" After doing this several times with the child, we show them a picture and ask "pig?" and they will correctly (most of the time) say "pig!" or "no, not pig!" depending on what the picture is. That is supervised machine learning*

# K-Nearest Neighbors

The k-nearest neighbors algorithm is a non-parametric, supervised learning classifier, which uses **proximity** to make classifications or predictions about the grouping of an individual data point. It works off the assumption that **similar points can be found near one another**.

A class label is assigned on the basis of a **majority vote**—i.e. the label that is most frequently represented around a given data point is used. “Majority voting” technically requires a majority of greater than 50%, which primarily works when there are only two categories. When you have multiple classes—e.g. four categories, you don’t necessarily need 50% of the vote to make a conclusion about a class; you could assign a class label with a vote of greater than 25%.

# K-Nearest Neighbors Applications

- **Data preprocessing**: Datasets frequently have missing values, but the KNN algorithm can estimate for those values in a process known as missing data imputation.
- **Recommendation Engines**: the a user is assigned to a particular group, and based on that group's user behavior, they are given a recommendation.
- **Finance**: a can help banks assess risk of a loan to an organization or individual. It is used to determine the credit-worthiness of a loan applicant. Furthermore, it can be uses in stock market forecasting, currency exchange rates, trading futures, and money laundering analyses.
- **Healthcare:** making predictions on the risk of heart attacks and prostate cancer. The algorithm works by calculating the most likely gene expressions

.

## Group Discussion

**What can be applications of classification models?**

.

# Classifier Metrics

Using different metrics for performance evaluation:

- Accuracy
- Confusion matrix
- Precision
- Recall
- AUC-ROC

# Accuracy

Accuracy simply measures how often the classifier correctly predicts. We can define accuracy as the ratio of the number of correct predictions and the total number of predictions.

Accuracy is useful when the target class is ***well balanced*** but is not a good choice for the unbalanced classes. Imagine the scenario where we had 99 images of the dog and only 1 image of a cat present in our training data. Then our model would always predict the dog, and therefore we got 99% accuracy. In reality, Data is always imbalanced for example Spam email, credit card fraud, and medical diagnosis. Hence, if we want to do a better model evaluation and have a full picture of the model evaluation, other metrics such as recall and precision should also be considered.

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

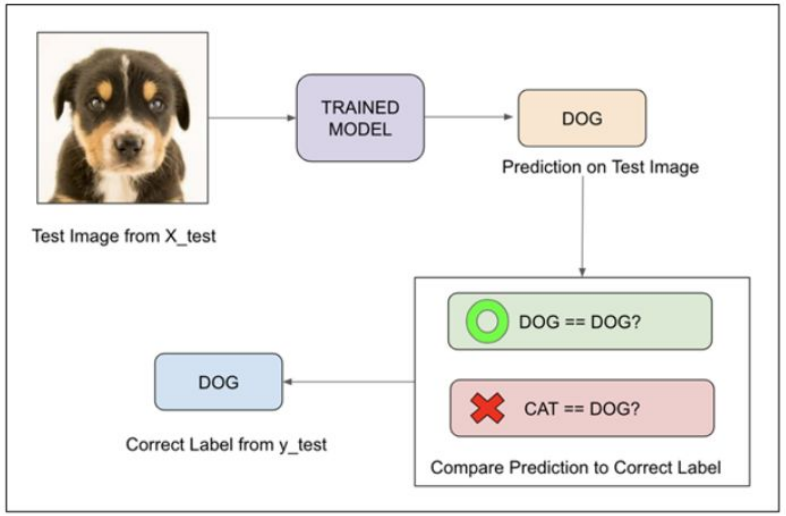



Image Source – https://www.kdnuggets.com/2020/04/performance-evaluation-metrics-classification.html

# Confusion Matrix

Confusion Matrix is a performance measurement for the machine learning classification problems where the output can be two or more classes. It is a table with combinations of predicted and actual values.

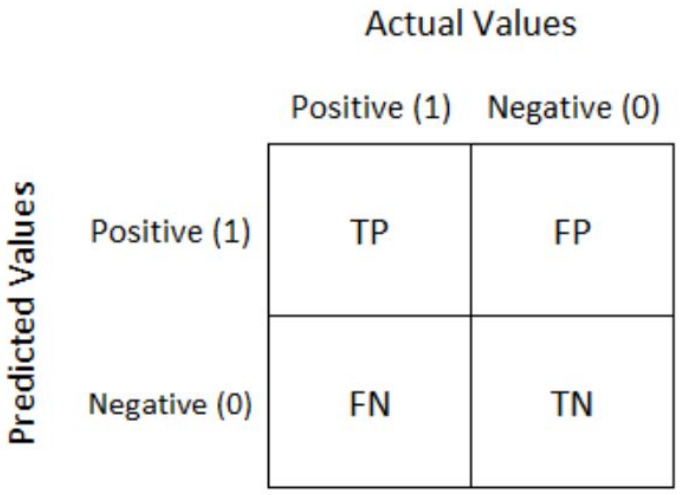



Image Source—https://www.roelpeters.be/glossary/what-is-a-confusion-matrix/

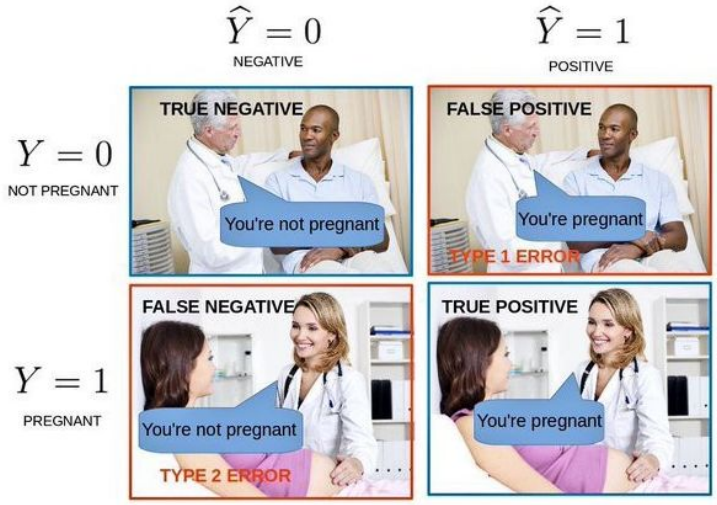



Image Source- https://dzone.com/articles/understanding-the-confusion-matrix

# Precision

Precision explains **how many of the correctly predicted cases actually turned out to be positive**. Precision is useful in the cases where False Positive is a higher concern than False Negatives.

The importance of *Precision is in music or video recommendation systems, e-commerce websites, etc. where wrong results could lead to customer churn and this could be harmful to the business.*

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

# Recall

Recall (Sensitivity)— explains how many of the actual positive cases we were able to predict correctly with our model. It is a useful metric in cases where False Negative is of higher concern than False Positive. It *is important in medical cases where it doesn't matter whether we raise a false alarm but the actual positive cases should not go undetected!*

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

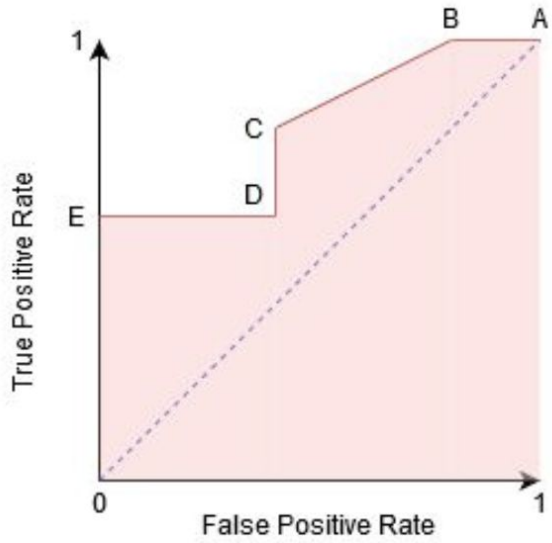
# AUC-ROC

AUC-ROC—The Receiver Operator Characteristic (ROC) is a probability curve that plots the TPR(True Positive Rate) against the FPR(False Positive Rate) at various threshold values and separates the 'signal' from the 'noise'.

The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes. From the graph, we simply say the area of the curve ABDE and the X and Y-axis.

From the graph shown below, the greater the AUC, the better is the performance of the model at different threshold points between positive and negative classes. This simply means that When AUC is equal to 1, the classifier is able to perfectly distinguish between all Positive and Negative class points. When AUC is equal to 0, the classifier would be predicting all Negatives as Positives and vice versa. When AUC is 0.5, the classifier is not able to distinguish between the Positive and Negative classes.

Metrics like accuracy, precision, recall are good ways to evaluate classification models for balanced datasets, but if the data is imbalanced then other methods like ROC/AUC perform better in evaluating the model performance.



# QUESTIONS?