**Cheat Sheet for SEAS 8500 – Fundamentals of AI Enabled Systems – Spring 2024 – Michael Wacey**

Chapter 1 - Overview of Machine Learning Systems

Machine learning is an approach to (1) learn (2) complex patterns from (3) existing data and use these patterns to make (4) predictions on (5) unseen data.

And: 6. It’s repetitive 7. The cost of wrong predictions is cheap 8. It’s at scale 9. The patterns are constantly changing

Inputs and patterns go into Traditional software and out comes output.

Inputs and Outputs go into Machine Learning software and out comes patterns.

Supervised v unsupervised learning vs no training systems (Zero shot learning)

Chart, bubble chart

Description automatically generated

Table 1-1. Key differences between ML in research and ML in production

|  |  |  |
| --- | --- | --- |
|  | Research | Production |
| Requirements | State-of-the-art model performance on benchmark datasets | Different stakeholders have different requirements |
| Computational priority | Fast training, high throughput | Fast inference, low latency |
| Data | Statica | Constantly shifting |
| Fairness | Often not a focus | Must be considered |
| Interpretability | Often not a focus | Must be considered |

Research usually prioritizes fast training, whereas production usually prioritizes fast inference.

Chapter 2 - Introduction to Machine Learning Systems Design

However, most systems should have these four characteristics: **reliability, scalability, maintainability,** and **adaptability**.

Diagram

Description automatically generated

Diagram

Description automatically generated

most ML engineers just use common loss functions like RMSE or MAE (mean absolute error) for regression, logistic loss (also log loss) for binary classification, and cross entropy for multiclass classification.

One approach is to combine these two losses into one loss and train one model to minimize that loss:

loss = ɑ quality\_loss + β engagement\_loss

But: ɑ quality\_score + β engagement\_score is better

Diagram

Description automatically generated

Chapter 3 - Data Engineering Fundamentals

ACID (atomicity, consistency, isolation, durability)

Chapter 4 - Training Data

Convenience sampling, Snowball sampling, Judgment sampling, Quota sampling

|  |  |  |
| --- | --- | --- |
| Method | How | Ground truths required? |
| Weak supervision | Leverages (often noisy) heuristics to generate labels | No, but a small number of labels are recommended to guide the development of heuristics |
| Semi- supervision | Leverages structural assumptions to generate labels | Yes, a small number of initial labels as seeds to generate more labels |
| Transfer learning | Leverages models pretrained on another task for your new task | No for zero-shot learning  Yes for fine-tuning, though the number of ground truths required is often much smaller than what would be needed if you train the model from scratch |
| Active learning | Labels data samples that are most useful to your model | Yes |

Precision = True Positive / (True Positive + False Positive)

Recall = True Positive / (True Positive + False Negative)

F1 = 2 × Precision × Recall / (Precision + Recall)

We can plot the true positive rate against the false positive rate for different thresholds. This plot is known as the ROC curve (receiver operating characteristics).

Chapter 5 - Feature Engineering

Missing not at random (MNAR)

This is when the reason a value is missing is because of the true value itself. In this example, we might notice that some respondents didn’t disclose their income. Upon investigation it may turn out that the income of respondents who failed to report tends to be higher than that of those who did disclose. The income values are missing for reasons related to the values themselves.

Missing at random (MAR)

This is when the reason a value is missing is not due to the value itself, but due to another observed variable. In this example, we might notice that age values are often missing for respondents of the gender “A,” which might be because the people of gender A in this survey don’t like disclosing their age.

Missing completely at random (MCAR)

This is when there’s no pattern in when the value is missing. In this example, we might think that the missing values for the column “Job” might be completely random, not because of the job itself and not because of any other variable. People just forget to fill in that value sometimes for no particular reason. However, this type of missing is very rare. There are usually reasons why certain values are missing, and you should investigate.