Question: Propose a model monitoring pipeline and describe how you would track model drift in 500 words.

Real-world data is dynamic, and AI models that rely on it must ensure that their performance remains accurate over time. Input data distributions might evolve, and relationships between features and outcomes may shift. Known as model drift, this can cause significant drops in prediction accuracy. To combat this, I propose a model monitoring pipeline that conducts **r**eal-time monitoring and responds to different types of drift.

In the first step of the pipeline, the model is monitored in real time. We continuously track its performance and behaviour in production using sliding windows. These windows periodically capture metrics such as accuracy, F1-score, AUC, prediction confidence (e.g., entropy), and class output distributions. A drop in accuracy or spikes in entropy can serve as early warning signs. If significant deviations are detected, the system triggers a drift check, else, it continues to run as usual.

Once triggered, the data is split into old and new windows. A fixed-size reference window (e.g., data from the past 30 days) is used as the old baseline. A comparison window of the same size is created from the most recent data. For example, if *t* is the detection time, then $data_{t-30} - data_t$ is the old window, and $data_{t+1} - data_{t+30}$ is the new. As more data arrives, both windows slide forward together.

To detect drift, we run statistical tests depending on feature type. For continuous variables, we use the Kolmogorov–Smirnov (KS) test, since it's non-parametric and works across any distributions. For categorical features, we apply the Chi-squared test or Jensen-Shannon Divergence. These tests assess whether the distributions have shifted significantly. High statistical scores mean that there is significant difference between old and new data distributions. Following that, we can implement a drift attribution layer identifies which features contribute most to the drift. We can use values like SHAP (SHapley Additive exPlanations) to measure the importance of each feature to the final output and can be used on any type of model.

Next, we calculate a drift severity score for each feature using the test statistics (e.g., KS statistic or Chi-squared value). These can be aggregated into an overall drift impact index, for example, a weighted average of top drifting features. For e.g. for *n* features and for each feature *i*, a drift score *d* is calculated (using the previous statistical tests) and a weight *w* (from SHAP values) is assigned. We can possibly use the following as a drift impact index. This weighted index gives more influence to features that are important.

$$Drift\ impact\ Index: \frac{\sum_{i=1}^{n} w_i\ d_i}{\sum_{i=1}^{n} w_i}$$

As more checks are performed over time, we define thresholds. Low severity may need monitoring, while medium or high severity triggers alerts and retraining. This scoring helps reduce false positives and prioritise serious drift cases.

Finally, the model is retrained using updated data. This includes refreshing learned parameters and re-running hyperparameter tuning. Once validated, the new model is deployed, and the real-time monitoring loop resumes, closing the pipeline.

This framework not only tracks model drift but also offers explainability and actionable thresholds for intervention, ensuring good model performance in dynamic environments.

(499 words)

References

- *Kolmogorov Smirnov test for AI: When and where to use it*. Arize AI. (2024, April 8). https://arize.com/blog-course/kolmogorov-smirnov-test/
- *Model Drift: Identifying and Monitoring for Model Drift in ML Engineering and Production*. Medium (2024, Feb 17). https://medium.com/@anicomanesh/model-drift-identifying-and-monitoring-for-model-drift-in-machine-learning-engineering-and-0f74b2aa2fb0
- *Tackling data and model drift in AI: Strategies for maintaining accuracy during ML model inference*. International Journal of Science and Research Archive. (2023, September 19). https://ijsra.net/sites/default/files/IJSRA-2023-0855.pdf
- *Chi-Square Goodness of Fit Test*. Jmp Statistical Discovery. (n.d.) https://www.jmp.com/en/statistics-knowledge-portal/chi-square-test/chi-square-goodness-of-fit-test
- *An Introduction to SHAP Values and Machine Learning Interpretability*. Datacamp. (2023, Jun 28). https://www.datacamp.com/tutorial/introduction-to-shap-values-machine-learning-interpretability