ONNX Proposal for Steering Committee

Stefan Acin, VP Eng. coming from ML Eng.
April 5th, 2023

Proposal

- Float64 calculation (and output) in TreeEnsambleRegressor and TreeEnsambleClassifier ML Ops
  https://github.com/onnx/onnx/blob/main/docs/Operators-ml.md#ai.onnx.ml.TreeEnsembleRegressor

Why - Technical

Precision loss due to float32 conversion with ONNX
http://www.xavierdupre.fr/app/mlprodict/helpsphinx/notebooks/onnx_shaker.html#runtime-supporting-float64-for-decisiontreeregressor

Additionally
ONNX graph, single or double floats
http://www.xavierdupre.fr/app/mlprodict/helpsphinx/notebooks/onnx_float32_and_64.html

Tricky detail when converting a random forest from scikit-learn into ONNX
http://www.xavierdupre.fr/app/mlprodict/helpsphinx/notebooks/onnx_float_double_skl_decision_trees.html

Issues when switching to float
https://onnx.ai/sklearn-onnx/auto_tutorial/plot_ebegin_float_double.html

Why - Usecase

- Models for Regulated Industry: Life Sciences Manufacturing
- Model Portability, but more importantly Model Retrocompatibility
- Tabular and TS Data with strong preference for “traditional” ML over DL
- SciKit-Learn, XGBoost, TSLearn, PySpark … PyTorch
- Model Metrics and inference results stay consistent
- Precision >>> Speed
- Case of alternate conversions is simpler in other cases (e.g. LinReg)

Next Steps

- Suggestions or alternative directions? Right Track?
- SIG and contribution to ONNX / ORT?
- Contacts - Slack? (specific channel/people?)
- Points to more information on how ONNX (and possibly ORT) is structured / implemented?

aizon

Thank You