

A类： Spark系统级开发

SPARK-29185

- Add new SaveMode types for Spark SQL jdbc datasource
- It is necessary to add new SaveMode for Delete, Update, and Upsert, such as:

SaveMode.Delete

SaveMode.Update

SaveMode.Upsert

So that Spark SQL could support legacy RDBMS much better, e.g. Oracle, DB2, MySQL etc.

- SPARK-20602
- Adding LBFGS optimizer and Squared_hinge loss for LinearSVC
- Currently LinearSVC in Spark only supports OWLQN as the optimizer (check <https://issues.apache.org/jira/browse/SPARK-14709>). I made comparison between LBFGS and OWLQN on several public dataset and found LBFGS converges much faster for LinearSVC in most cases.
- The following table presents the number of training iterations and f1 score of both optimizers until convergence

Dataset	LBFGS with hinge	OWLQN with hinge	LBFGS with squared_hinge
news20.binary	31 (0.99)	413(0.99)	185 (0.99)
mushroom	28(1.0)	170(1.0)	24(1.0)
madelon	143(0.75)	8129(0.70)	823(0.74)
breast-cancer-scale	15(1.0)	16(1.0)	15 (1.0)
phishing	329(0.94)	231(0.94)	67 (0.94)
a1a(adult)	466 (0.87)	282 (0.87)	77 (0.86)
a7a	237 (0.84)	372(0.84)	69(0.84)

- data source: <https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html>
- training code: `new LinearSVC().setMaxIter(10000).setTol(1e-6)`
- LBFGS requires less iterations in most cases (except for a1a) and probably is a better default optimizer.