Item 116
git_comments:

git_commits:

1. **summary:** [SPARK-9642] [ML] LinearRegression should supported weighted data
   message: [SPARK-9642] [ML] LinearRegression should supported weighted data In many modeling application, data points are not necessarily sampled with equal probabilities. Linear regression should support weighting which account the over or under sampling. work in progress. Author: Meihua Wu <meihuawu@umich.edu> Closes #8631 from rotationsymmetry/SPARK-9642.

github_issues:

github_issues_comments:

github_pulls:

1. **title:** [SPARK-9642] [ML] LinearRegression should supported weighted data
   body: In many modeling application, data points are not necessarily sampled with equal probabilities. Linear regression should support weighting which account the over or under sampling. work in progress.

github_pulls_comments:

1. Jenkins, add to whitelist
2. ok to test
3. [Test build #42318 has finished]
   (https://amplab.cs.berkeley.edu/jenkins/job/SparkPullRequestBuilder/42318/console) for PR 8631 at commit [`e9093cb`]
   (https://github.com/apache/spark/commit/e9093cbea2554fbc124899a58e3cbfdade5ea795). - This patch **fails Scala style tests**. - This patch merges cleanly. - This patch adds the following public classes _(experimental)_: - `case class WeightedLabeledPoint(label: Double, features: Vector, weight: Double)`
4. @dbtsai Thank you for OKing the test. My patch depends on the `MultivariateOnlineSummarizer` in your PR for applying weights to logistics regressions ([link](https://github.com/apache/spark/pull/7884)). My patch should be ready to test after your PR is merged.
5. Hello, weighted `MultivariateOnlineSummarizer` is merged which unblocks you. Thanks.
6. @dbtsai Thank you for your comments. I have revised the patch. Please test.
7. [Test build #42579 has finished]
   (https://amplab.cs.berkeley.edu/jenkins/job/SparkPullRequestBuilder/42579/console) for PR 8631 at commit [`3f98247`]
   (https://github.com/apache/spark/commit/3f98247801368a86aaffabd78b3755bf36fab330). - This patch **fails Spark unit tests**. - This patch merges cleanly. - This patch adds no public classes.
8. "org.apache.spark.HeartbeatReceiverSuite.reregister if heartbeat from removed executor" failed, which should be unrelated to this patch. @dbtsai Will you please initial a retest?
9. Jenkins, retest this please
10. [Test build #42611 has finished]
    (https://amplab.cs.berkeley.edu/jenkins/job/SparkPullRequestBuilder/42611/console) for PR 8631 at commit [`3f98247`]
    (https://github.com/apache/spark/commit/3f98247801368a86aaffabd78b3755bf36fab330). - This patch **fails Spark unit tests**. - This patch merges cleanly. - This patch adds no public classes.
11. [Test build #42621 has finished]
    (https://amplab.cs.berkeley.edu/jenkins/job/SparkPullRequestBuilder/42621/console) for PR 8631 at commit [`2afa2a1`]
    (https://github.com/apache/spark/commit/2afa2a190368adb99ec398c64744fc7dafc98bed). - This patch **passes all tests**. - This patch merges cleanly. - This patch adds the following public classes _(experimental)_: - `class Interaction(override val uid: String) extends Transformer`
12. @dbtsai Thanks for the comment on indentation. I have fixed it in the patch.
13. [Test build #42640 has finished]
    (https://amplab.cs.berkeley.edu/jenkins/job/SparkPullRequestBuilder/42640/console) for PR 8631 at commit [`1f731c2`]

(https://github.com/apache/spark/commit/1f731c28ad8a59f3bf432435253dc7b0984f46b4). - This patch **fails Spark unit tests**. - This patch merges cleanly. - This patch adds the following public classes _(experimental)_: - `class AFTSurvivalRegression @Since("1.6.0") (@Since("1.6.0") override val uid: String)` - `require(censor == 1.0 || censor == 0.0, "censor of class AFTPoint must be 1.0 or 0.0")`
14. [Test build #42670 has finished] (https://amplab.cs.berkeley.edu/jenkins/job/SparkPullRequestBuilder/42670/console) for PR 8631 at commit [`854d0bb`] (https://github.com/apache/spark/commit/854d0bb58d0a6b43135ce9e750e4f9df36a65003). - This patch **passes all tests**. - This patch merges cleanly. - This patch adds no public classes.
15. Can you merge the master to resolve the conflicts? Also, add warning in training summary that it ignores the training weights currently (except for the objective trace). Other than those small items, LGTM. You may remove WIP.
16. [Test build #42746 has finished] (https://amplab.cs.berkeley.edu/jenkins/job/SparkPullRequestBuilder/42746/console) for PR 8631 at commit [`57c57f1`] (https://github.com/apache/spark/commit/57c57f102ae3d55149c8d3fc3cd7d4c95531f9b3). - This patch **fails Scala style tests**. - This patch merges cleanly. - This patch adds the following public classes _(experimental)_: - `case class Sort(`
17. [Test build #42747 has finished] (https://amplab.cs.berkeley.edu/jenkins/job/SparkPullRequestBuilder/42747/console) for PR 8631 at commit [`b0144ce`] (https://github.com/apache/spark/commit/b0144cef37986c97329d7416d53ff9da75d94350). - This patch **fails PySpark unit tests**. - This patch merges cleanly. - This patch adds no public classes.
18. [Test build #42757 has finished] (https://amplab.cs.berkeley.edu/jenkins/job/SparkPullRequestBuilder/42757/console) for PR 8631 at commit [`b3fae99`] (https://github.com/apache/spark/commit/b3fae9954d24d9d88b5bbd016e8f285cae1825fe). - This patch **passes all tests**. - This patch merges cleanly. - This patch adds the following public classes _(experimental)_: - `case class Sort(`
19. Thanks. Merged into master.

github_pulls_reviews:

1. We decided to keep `count` as it, and add `weightSum`.
2. Refactor the `Instance` case class out from LoR, and use it for code readability.
3. use `lit` and `col` for simplifying the code. See example in LoR.
4. The doc is changed in LoR. Please sync with that.
5. indentation. see LoR for example.
6. ditto
7. Please add `if (weight == 0) return this`.
8. make `data` as `instance`
9. `private var weightSum: Double = 0.0`
10. remove extra line
11. Move `)` to the end of line `combOp`
12. Please use `activeData.map`
13. ditto
14. ditto
15. Make `case LabeledPoint(label, features) => Instance(label, weight = 0.0, features)` for easier readability.
16. remove this extra line.
17. Since you already move `case Instance(label, weight, features) =>` to new line, let's do ``` scala def add(instance: Instance): this.type = { instance match { case Instance(label, weight, features) => ... } } ```
18. Good point. I will revise it. Thanks!
19. Could you add a block of `{}` here. Thanks. After this, it's good to go.

jira_issues:

jira_issues_comments: