BerkeleyX: CS120x Distributed Machine Learning with Apache Spark

Bookmarks

Week 1 - Course Overview, Software Setup, and Machine Learning Basics

Syllabus Quiz
Quiz due Aug 08, 2016 at 23:00 UTC

Setting up the Course Software Environment
Lab 0 due Aug 08, 2016 at 23:00 UTC

(Optional) Survey about your Machine and Setup Experience

Lecture 1: Course Overview and Introduction to Machine Learning
Quiz due Aug 08, 2016 at 23:00 UTC

Lab1a/Lab1b - Math and Python Review and RDD overview
Lab due Aug 08, 2016 at 23:00 UTC

Bookmark

Data Growth

(1/1 point)

Data is growing faster than Moore's law.

- (selected) True ✔
- False

EXPLANATION

New data is being generate at a rate that exceeds the growth in processing power (per cpu) or disk storage (per disk), which is why we need a solution that scales out (multiple nodes) versus scales up.

Supervised and Unsupervised

(1/1 point)

Select all of the true statements:

- [ ] Supervised models are trained on unlabeled data
- [x] Supervised models are trained on labeled data ✔
- [x] Unsupervised models are trained on unlabeled data ✔
- [ ] Unsupervised models are trained on labeled data

✔

Note: Make sure you select all of the correct options—there may be more than one!

EXPLANATION

Supervised models are trained using labeled data to build a model that minimizes an evaluation metric based on the label, while unsupervised models are trained with unlabeled data to find hidden structure in the data.

Spam Model

(1/1 point)

A model that predicts spam / not-spam is a:

- Regression Model
- Classification Model ✔
- Clustering Model

EXPLANATION

Regression models are used to predict continuous values, while classification models predict two or more categories. Clustering is an unsupervised learning method that finds latent structure in unlabeled data.

Gradient Descent

(1/1 point)

Gradient descent can be used to train various kinds of regression and classification models.

- True ✔
- False

EXPLANATION

Gradient descent is a general purpose iterative algorithm that scales efficiently and is used train a variety of machine learning models.

Probabilistic Interpretation

(1/1 point)
Which of the following models has the most direct probabilistic interpretation?

- (•) Logistic Regression ✔
- ( ) K-means
- ( ) Linear Regression

EXPLANATION

The prediction from a logistic regression model can be interpreted as the probability that the label is 1. In contrast K-means is an unsupervised clustering model that returns cluster membership while linear regression returns real numbers as predictions. Neither returns results that can directly be interpreted as probabilities.

Overfitting

(1/1 point)

If we train and evalute our model using our test data, it is likely that our evaluation of the resulting model will be:

- Overly Pessimistic
- Fair
- Overly Optimistic ✔

EXPLANATION

Since we have used the same data for training and testing, it is likely that we've overfit our model to the test data, so evaluating the model on the test data no longer provides a fair estimate of expected performance. It is likely that our model will perform worse than our estimate when used on unseen data.

Big O Notation

(1/1 point)

An algorithm with O(n^3) complexity will always run slower than an algorithm with O(n^2) complexity.

True

False ✔

EXPLANATION

Complexity is based off of how the algorithm scales for very large n. Constants and lower-order terms are dropped. For example, a model that takes 100*n^2 time has O(n^2) complexity, as does a model that takes 4n + n^2 time. A counterexample for this question is: algorithm A has running time n^3 and algorithm B has running time 10n^2. Until n >= 10, A is faster than B.