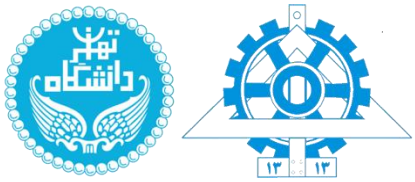
به نام خدا



سامانه‌های یادگیری ماشین توزیع‌شده

تمرین کامپیوتری ۳- بخش اول

موعد تحویل: ۵ دی ۱۴۰۱، ساعت ۲۳:۵۵

مسئول تمرین: احمدرضا سزاوار
ahmadreza.sezavar@gmail.com

لطفا پیش از شروع کار برروی تمرین، به نکات زیر توجه کنید:

- برای دسترسی به UI ماشین Spark Master، میتوانید به آدرس http://raspberrypi-dml0:8080 رفته و از نام کاربری admin و گذرواژه dmlsAdmin استفاده کنید.
- برای دسترسی به UI مربوط به HDFS میتوانید به آدرس http://raspberrypi-dml0:9870/explorer.html بروید. لطفا فایلهای دانشجویان دیگر را تغییر ندهید.

- برای دسترسی به HDFS در کد خود، میتوانید از آدرس hdfs://raspberrypi-dml0:9000 استفاده کنید. مثلا اگر بخواهید خروجی را در فایل tmp واقع در دایرکتوری dousti بنویسید، باید به کمک تابع مربوطه در pyspark، آن را در آدرس hdfs://raspberrypi-dml0:9000/dousti/tmp بنویسید.

- لطفا با کمک راهنمای دستورات HDFS که در سایت elearn آپلود شده یک دایرکتوری برای خود بسازید و فایل هایتان را فقط در آن قرار دهید.
- تمرین باید روی کلاستر درس (کلاستر ایجاد شده از برد های رزبری پای) انجام شود و آدرس IP های جدید در زیر نوشته شده است:

raspberrypi-dml0 (172.18.35.204)

raspberrypi-dml1 (172.18.35.205)

raspberrypi-dml2 (172.18.35.206)

raspberrypi-dml3 (172.18.35.203)

- برای پیاده سازی از زبان Python و کتابخانه pyspark استفاده کنید. دقت کنید که به جای mllib از کتابخانه ml استفاده کنید.
- دقت کنید که منابع کلاستر بین همه ی دانشجویان مشترک است. به همین دلیل سعی کنید بیشتر از مقدار مورد نیاز از آن استفاده نکنید.
- قبل از شروع تمرین بهتر است ویدیو آماده شده مرتبط که در سامانه درس آپلود شده را مشاهده نمایید.
- سوالات خود را می توانید از طریق ایمیل، گروه تلگرام درس یا تلگرام مسئول تمرین مطرح کنید. کد یا پاسخ سوالات را در گروه یا با سایر دانشجویان به اشتراک نگذارید.

سوال ۱

در این سوال، با استفاده از فایل City.txt که اسامی شهرهای ایران در آن تکرار شده است، قرار است با استفاده از **spark rdd** به موارد زیر پاسخ دهیم:

الف) تعداد تکرار شهر ها را در کل فایل داده شده را در خروجی نشان دهید، برای مثال اگر شهر تهران 3 بار در کل فایل تکرار شده است، در خروجی باید Tehran: 3 چاپ شده باشد.

ب) در هر خط از فایل داده شده، اسامی تعدادی شهر نوشته شده است، هر خط را به صورت جداگانه براساس حرف ابتدایی شهر مرتب کنید و خروجی را نشان داده و فایل خروجی را نیز ذخیره کنید.

سوال ۲

در این سوال با استفاده از دیتاست داده شده (فایل data.csv) می خواهیم یک برنامه یادگیری ماشین از نوع رگرسیون انجام دهیم، دیتاست شامل چهار ویژگی مختلف است که قرار است با کمک آنها مقادیر ستون label را پیشبینی کنیم. مدل های مورد نظر نیز شامل چهار مدل مختلف زیر می باشند:

- Linear Regression
- Decision Tree
- Random Forest
- Gradient-Boosted Trees

الف) در ابتدا داده ها را خوانده و بررسی های زیر را روی داده ها انجام دهید:

- پنج رکورد اول دیتاست را نمایش دهید.
- دیتاست را از نوع مولفه های آماری مانند میانگین، واریانس، کمینه و بیشینه تحلیل کنید.
- برای هر یک از ستون ها، تعداد مقدار های Null را مشخص کنید.

ب) داده ها را به دو بخش آموزش و تست به نسبت 80 به 20 و به صورت رندوم تقسیم کنید.

ج) برای هر یک از مدل های فوق:

- پیش پردازش های لازم و آموزش را به کمک Pipeline انجام دهید.
- برای حداقل 5 نمونه مقدار پیشبینی شده با مقدار واقعی را نمایش دهید.

د) برای ارزیابی و مقایسه مدل ها با هم از دو ابزار Root Mean square Error و R^2 استفاده می کنیم، مدل ها را به کمک این دو متریک ارزیابی و تحلیل کرده و با هم مقایسه کنید.

سوال ۳

مدلی که بهترین عملکرد را در سوال 2 داشته است، در نظر بگیرید، آموزش مدل را با توجه به موارد زیر یکبار دیگر انجام دهید و با هم مقایسه کنید. به ازای هریک موارد زیر، دستور spark-submit استفاده شده را نیز ذکر کنید.

- یک ماشین و یک هسته
- یک ماشین و دو هسته
- دو ماشین و هر ماشین یک هسته
- دو ماشین و هر ماشین دو هسته

نحوه تحویل تمرین

فایل‌هایتان را به صورت زیر نام‌گذاری کنید:

۱. گزارش: `report.pdf` (دقت کنید که فقط فرمت PDF برای گزارش پذیرفته می‌شود.)

۲. نام‌گذاری کدها بر اساس جدول زیر باشد.

شماره سوال	بخش	نوع کد	نام فایل‌ها
۱	الف	Python, text	Spark_rdd_1.py Spark_rdd_1.txt
	ب	Python, text	Spark_rdd_2.py Spark_rdd_2.txt
2	الف تا د	Python, text	LR_Model.py, LR_Model.txt, DT_Model.py, DT_Model.txt, RF_Model.py, RF_Model.txt, GBT_Model.py, GBT_Model.txt
3	الف	Python, text	Best_Model.py Best_Model.txt

در انتها، همه‌ی فایل‌ها را در پوشه‌ای قرارداده و نام آن را شماره‌ی دانشجویی خود قرار دهید. پوشه را zip کرده و آن‌را در سایت elearn آپلود کنید.