به نام خدا

سامانه‌های یادگیری ماشین توزیع‌شده (پاییز ۱۴۰۳)

تمرین نوشتاری ۲

موعد تحویل: ۱۴۰۳/۱۰/۱۸

۱- می دانیم در چند دهه‌ی گذشته، نرخ افزایش ظرفیت دیسک‌ها بسیار بیشتر از سرعت آنها بوده است. توضیح دهید روش‌های ذخیره سازی توزیع شده چگونه مشکل کندی سرعت دیسک‌ها را برطرف میکنند؟

۲- فرض کنید توپولوژی حلقه را با n ماشین پیاده کرده اید. ماشین i ام به احتمال p_i ممکن است خراب شود. اگر بتوانید یک ماشین دیگر به عنوان پشتیبان تهیه کنید تا در صورت خرابی یکی از ماشین ها با آن جایگزین شود، احتمال خرابی این سیستم چقدر کاهش میابد؟ احتمال خرابی ماشین پشتیبان را صفر در نظر بگیرید.

۳- فرض کنید که قصد داریم عدد 6.342e-5 را که به صورت Float32 ذخیره شده است از یک سیستم با پردازنده Intel x86 به سیستم دیگری با پردازنده amd x86 ارسال کنیم. برای انجام این کار از پروتکل TCP/IP استفاده می‌کنیم. مراحل ارسال این پیام با توجه به فرآیند serialization چگونه است؟ روش آماده سازی و بازسازی را تشریح کنید.

۴- قطعه کد زیر، مربوط به تولید دو دیتافریم و اعمال برخی در تغییرات بر روی آن ها است. با توجه به قطعه کد زیر نحوه عملکرد و ترتیب اجرای این کد را توسط Spark با ذکر توضیحات کافی مشخص نمایید.

```
data = [("Alice", 25), ("Bob", 30), ("Cathy", 27)]
columns = ["Name", "Age"]
df = spark.createDataFrame(data, columns)
filtered_df = df.filter(df.Age > 26)
transformed_df = filtered_df.withColumn("AgeIn5Years", filtered_df.Age + 5)
renamed_df = transformed_df.withColumnRenamed("AgeIn5Years", "AgeAfterFiveYears")
########################################################################
data_2nd = [("James", 28), ("Josh", 32), ("Sarah", 21)]
columns_2nd = ["Name", "Age"]
df_2nd = spark.createDataFrame(data_2nd, columns_2nd)
filtered_df_2nd = df_2nd.filter(df_2nd.Age > 26)
transformed_df_2nd = filtered_df_2nd.withColumn("AgeIn15Years", filtered_df_2nd.Age + 15)
renamed_df_2nd = transformed_df_2nd.withColumnRenamed("AgeIn15Years", "AgeAfterFiftheenYears")
row_count_2nd = renamed_df_2nd.count()
########################################################################
result_2nd = renamed_df_2nd.collect()
result = renamed_df.collect()
```

۵- در مبحث serialization دیدیم که این مفهوم با وجود اهمیت زیاد، چالش‌هایی را ایجاد می‌کند. دو مورد از این چالش‌ها Object representation و Object references هستند. با ذکر مثال هر یک را توضیح دهید و برای هر کدام یک راه حل ارائه نمایید.

۶- دو مفهوم Static computation graph و Dynamic computation graph را با یکدیگر مقایسه کنید.

۷- مفهوم Model Sharding را توضیح دهید و استراتژی‌های مختلف آن در Pytorch را مقایسه کنید. این مفهوم در چه مواقعی استفاده می‌شود؟

۸- معماری زیر مربوط به مدل LeNet است. با فرض استفاده از Vannilla AdamW و SGD، با در نظر گرفتن batch size=8، میزان حافظه مصرفی برای آموزش این شبکه چقدر است؟ فرآیند استنتاج چطور (با batch_size=1)؟