Final Exam Sample Questions

DataScience Spring 1404
Mohammad Amanlou

From CS188, FA18, UC Berkeley: http://ai.berkeley.edu.

برای هر یک از گزاره های زیر مشخص کنید آیا مغالطه آماری وجود دارد یا خیر. پاسخ خود را به طور کامل توضیح داده و در صورت وجود مغالطه آماری نحوه تصحیح آن را شرح دهید.

- الف) به گفته معاون امور جوانان و ساماندهی وزارت ورزش و جوانان آمار طلاق در سال ۱۳۹۶ نسبت به سال قبل از آن از ۱۸۱۰۴۹ مورد به ۱۷۴۵۹۰ مورد کاهش یافته است که این کاهش ۳/۵۷ درصدی نشان از عملکرد موفق این سازمان دارد.

برای هر یک از گزاره های زیر مشخص کنید آیا مغالطه آماری وجود دارد یا خیر. پاسخ خود را به طور کامل توضیح داده و در صورت وجود مغالطه آماری نحوه تصحیح آن را شرح دهید.

- ب) طبق اعلام مرکز ملی آمار افزایش متوسط درآمد سالانه خانوار شهری و روستایی در سال ۱۳۹۹ نسبت به سال ۱۳۹۸ به ترتیب ۲۴/۴ و ۲۷/۴ درصد بوده است در حالی که افزایش متوسط هزینه سالانه خانوار شهری و روستایی در مدت زمان مشابه به ترتیب ۲۰/۶ و ۲۱/۷ درصد اعلام شده، به عبارت دیگر با وجود تورم اوضاع معیشتی مردم رو به بهبود است.

برای هر یک از گزاره های زیر مشخص کنید آیا مغالطه آماری وجود دارد یا خیر. پاسخ خود را به طور کامل توضیح داده و در صورت وجود مغالطه آماری نحوه تصحیح آن را شرح دهید.

- ج) افزایش میزان باسوادی در کشور از ۸۴/۶ درصد در سال ۱۳۸۵ به ۸۷/۶ درصد در سال ۱۳۹۵، با افزایش نرخ بیکاری از ۱۰/۳ درصد به ۱۳/۵ درصد همراه بوده است، بنابراین بر خلاف نظریه های جامعه شناسان غربی، سوادآموزی موجب از میان رفتن بیکاری نمی‌شود.

برای هر یک از گزاره های زیر مشخص کنید آیا مغالطه آماری وجود دارد یا خیر. پاسخ خود را به طور کامل توضیح داده و در صورت وجود مغالطه آماری نحوه تصحیح آن را شرح دهید.

- د) برای بررسی نظر دانشجویان درباره کیفیت غذای سلف، کافی است در جلوی درب خروجی سلف بایستیم و از هر ۵ نفری که از سلف خارج می شوند یک نفر را به تصادف انتخاب کرده و در مورد کیفیت غذا از او سوال کنیم.

برای هر یک از گزاره های زیر مشخص کنید آیا مغالطه آماری وجود دارد یا خیر. پاسخ خود را به طور کامل توضیح داده و در صورت وجود مغالطه آماری نحوه تصحیح آن را شرح دهید.

- ه) با توجه به این که ۴۹/۲ درصد مردم ایران مذکر هستند و ۶۴ درصد ایرانیان فوتبال تماشا می کنند، مردان ایرانی علاقه مند به فوتبال ۳۱/۵ درصد کل جامعه را تشکیل میدهند.

برای هر یک از گزاره های زیر مشخص کنید آیا مغالطه آماری وجود دارد یا خیر. پاسخ خود را به طور کامل توضیح داده و در صورت وجود مغالطه آماری نحوه تصحیح آن را شرح دهید.

- و) میانگین درآمد خانواده های ایرانی با در نظر گرفتن تورم، در یک دهه گذشته ۶ درصد افزایش یافته است که این نشان از بهبود وضعیت اقتصادی اکثریت افراد جامعه دارد.

برای هر یک از گزاره های زیر مشخص کنید آیا مغالطه آماری وجود دارد یا خیر. پاسخ خود را به طور کامل توضیح داده و در صورت وجود مغالطه آماری نحوه تصحیح آن را شرح دهید.

- ز) در آزمون فرضی value-p برابر با ۰/۰۱ شده است، بنابراین ۹۹٪ اطمینان داریم فرض صفر اشتباه است.

برای هر یک از کارهای زیر، مشخص کنید آیا خوشه بندی به روش Kmeans روش مناسبی است یا خیر.

- الف) با داشتن اطلاعات تاریخی مربوط به آب و هوا، می خواهیم پیشبینی کنیم آیا فردا بارانی است یا خیر.
- ب) با داشتن مجموعه بزرگی از مقالات از وبسایتهای مختلف، میخواهیم موضوعات اصلی که توسط این مقالات پوشش داده شده اند را پیدا کنیم.
- پ) با داشتن الگوهای زمانی استفاده کاربران از یک وبسایت، می خواهیم گروه های مختلف کاربران را شناسایی کنیم.
- ت) با داشتن داده های فروش تعداد زیادی محصول در یک فروشگاه، میزان فروش هر محصول در آینده را پیش بینی کنیم.

شما به عنوان یک دانشمند داده در تیم طراحی سامانه توصیه گر یک فروشگاه اینترنتی مانند دیجی‌کالا استخدام شده اید. مجموعه داده بزرگی در اختیار شما قرار گرفته است که ویژگی‌های مختلفی درباره رفتار کاربران، جزئیات محصولات، و تاریخچه تراکنشها را داراست. برخی از این ویژگی ها عبارتند از:

- مشخصات مشتریان (سن، جنسیت، مکان)
- رفتار کاربران (صفحاتی که بازدید کرده‌اند، مدت زمانی که در هر صفحه گذرانده اند)
- تاریخچه خرید کاربران (محصولات خریداری شده، تعداد دفعات خرید هر محصول، مبلغ خرج شده)
- ویژگیهای محصول (دسته بندی، قیمت، امتیاز کاربران)

مجموعه داده بیش از ۵۰۰ ویژگی مختلف دارد و وظیفه شما استفاده از الگوریتمهای کاهش ابعاد جهت آماده سازی داده برای یک مدل یادگیری ماشین جهت پیش بینی محصول بعدی که مشتری احتمالا خریداری خواهد کرد، است. با توجه به خصوصیات این مجموعه داده و هدف مورد نظر، از کدام روش های کاهش ابعاد استفاده خواهید کرد و چرا؟ پاسخ خود را با در نظر گرفتن طبیعت این مجموعه داده، اهمیت ارتباط بین ویژگی‌ها، و نیاز به تفسیرپذیری و مقیاس پذیری مدل، به طور کامل شرح دهید.

درست یا غلط؟ (با توجیه مختصر)

الف) برای اینکه از overfittingیک مدل پیش بینی پیشگیری کنیم، تکرار مثال‌های موجود در داده آموزش به ما کمک میکند.

ب) به منظور تخمین دقت یک دسته‌بند، بهترین راه استفاده از داده آموزش میباشد.

پ) یک شبکه عصبی عمیق میتواند ویژگیهای پیچیده ی مورد نیاز را خودش یاد بگیرد و این باعث قدرتش میشود.

ت) مدلهایی مانند چت جی پیتی توانایی پیشبینی احتمال رخداد یک جمله در زبان انگلیسی را دارند.

ث) اگر مدلی بین دو متغیر xو y به صورت $y = \Theta_0 + \Theta_1 x^2$ تعریف شود، دیگر نمیتوان از روشهای simple linear regression برای تخمین پارامترهای Θ استفاده کرد.

ج) درخت تصمیم یا جنگل تصادفی لزوما نیازی به پیش پردازش جهت استفاده از ویژگی های categorical ندارند.

چ) نقش تابع سیگموید در regression logistic این است که خروجی را به احتمال تبدیل کند.

ح) الگوریتم گرادیان کاهشی همیشه کمینه یک تابع را پیدا میکند.

پرسش هایی با پاسخ کوتاه

- الف) جدولی داریم از نمرات دانشجویان کلاس علوم داده که مشخصات دانشجو (نام، شماره دانشجویی، مقطع، رشته) و نمره های امتحان و تمرینات آنها ستون هایش می‌باشد. برای متوسط نمره در هر رشته از چه query SQL استفاده کنیم؟
- ب) ایده اصلی embedding wordها چیست؟ در مدل های زبانی از اینها چگونه برای درک متن استفاده میشود؟
- پ) چرا از الگوریتم گرادیان کاهشی برای یافتن پارامترهای بهینه استفاده میشود؟ (نسبت به بقیه روش ها چه برتری دارد؟)
- ت) دانستن توزیع مقادیر یک ویژگی در دادگان چگونه به انتخاب روش پیش پردازش کمک میکند؟
- ث) برای تفسیر میزان اهمیت یک ویژگی در یک regression logisticچه می توان کرد؟

پرسش هایی با پاسخ تشریحی

- یک انتخابات بین دو کاندیدا فرض کنید. صندوق های یک روستا به دلایلی گم میشوند . از شما به عنوان مهندس علم داده درخواست کمک میشود تا رای دو کاندیدا را برای این روستا تخمین بزنید. توضیح دهید که چگونه این کار را انجام میدهید. (مثلا از چه روشی و چه ویژگی هایی استفاده می کنید.)

پرسش‌هایی با پاسخ چندگزینه‌ای

- چه عواملی باعث می‌شود که داده‌ها به عنوان Big Data شناخته شوند؟
- الف) حجم، تنوع، سرعت
- ب) حجم، ارزشیابی، صحت
- ج) صحت، کمیت، سرعت
- د) همه موارد بالا

پرسش‌هایی با پاسخ چندگزینه‌ای

- چه تکنیک‌هایی برای مدیریت داده‌های بزرگ پیشنهاد می‌شود؟
- الف) نمونه‌برداری
- ب) توزیع داده‌ها
- ج) پخش و استریمینگ
- د) همه موارد بالا

پرسش‌هایی با پاسخ چندگزینه‌ای

- کدام یک از ویژگی‌های Hadoop برای کار با داده‌های بزرگ بهینه است؟
- الف) پردازش موازی
- ب) ذخیره‌سازی داده‌ها در فضای ابری
- ج) خواندن و نوشتن در دیسک‌های توزیع‌شده
- د) همه موارد بالا

پرسش‌هایی با پاسخ چندگزینه‌ای

- در الگوریتم MapReduce، چرا برای برخی محاسبات چندمرحله‌ای کندی مشاهده می‌شود؟
- الف) به دلیل نیاز به استفاده از حافظه زیاد
- ب) به دلیل تکرار و ذخیره داده‌ها در سیستم‌های توزیع‌شده
- ج) به دلیل محدودیت سرعت در پردازش‌های مبتنی بر دیسک
- د) هیچکدام

پرسش‌هایی با پاسخ چندگزینه‌ای

- در فناوری Apache Spark چه نوع داده‌ای به عنوان RDD معرفی می‌شود و چرا مهم است؟
- الف) داده‌های تصادفی که به طور موازی پردازش می‌شوند
- ب) داده‌هایی که به صورت ایمن و مقاوم در برابر خرابی‌ها ذخیره می‌شوند
- ج) داده‌های خام که برای پردازش نیاز به اعمال الگوریتم‌های خاص دارند
- د) همه موارد بالا

پرسش‌هایی با پاسخ چندگزینه‌ای

- در رابطه با RDD (Resilient Distributed Datasets)، چه مفهومی به Fault Toleranceمربوط می‌شود؟
- الف) بازسازی داده‌های گمشده از طریق ردیابی تحولات
- ب) پردازش داده‌ها بدون توجه به خرابی‌ها
- ج) ذخیره‌سازی داده‌ها به شکل ایمن برای بازیابی آسان
- د) هیچکدام

پرسش‌هایی با پاسخ چندگزینه‌ای

- چه ابزارهایی در MLlib برای تحلیل داده‌های بزرگ موجود است؟
- الف) رگرسیون لجستیک، درخت تصمیم
- ب) خوشه‌بندی K-means، تجزیه PCA
- ج) فیلتر کردن غیر منفی ماتریس NMF
- د) همه موارد بالا

پرسش‌هایی با پاسخ چندگزینه‌ای

- چرا در Big Data برای سرعت پردازش و ذخیره‌سازی به فناوری‌هایی مانند Spark نیاز است؟
- الف) به دلیل نیاز به پردازش داده‌ها به صورت همزمان و با سرعت بالا
- ب) برای انجام تحلیل‌های پیچیده و زمان‌بر به صورت موازی
- ج) برای تسهیل پردازش‌های موازی و کاهش زمان پردازش
- د) همه موارد بالا

تفاوت بین سیستم‌های مبتنی بر قواعد، زنجیره‌ها و workflowها با Language Agentها چیست؟

- سیستم‌های مبتنی بر قواعد معمولاً بر اساس مجموعه‌ای از قواعد از پیش تعریف شده کار می‌کنند که باید به‌طور دستی تنظیم شوند.زنجیره‌ها معمولاً برای پردازش وظایف به صورت پشت سر هم و طبق یک تسلسل خاص طراحی می‌شوند.Workflowها مجموعه‌ای از مراحل پردازشی هستند که به ترتیب اجرا می‌شوند و ممکن است با محیط‌های مختلف تعامل کنند.

چگونه معماری CoALA به بهبود عملکرد Language Agentها کمک می‌کند؟

- CoALA یک معماری طراحی‌شده برای Language Agentها است که از یک سیستم حافظه مدولار استفاده می‌کند. این معماری شامل حافظه تجربی (که تجربیات گذشته را ذخیره می‌کند)، حافظه معنایی (که دانش عمومی را ذخیره می‌کند) و حافظه رویه‌ای (که اطلاعات در مورد نحوه اجرای وظایف را ذخیره می‌کند) است. این ساختار به Language Agentها این امکان را می‌دهد که اطلاعات مختلف را به‌طور مؤثر پردازش کرده و تصمیمات بهتر و هوشمندتری بگیرند.