به نام خدا

دانشکده برق و کامپیوتر
دانشکدگان فنی دانشگاه تهران

مبانی علوم داده

امتحان پایان‌ترم

۱۰ تیر ۱۴۰۳

نام و نام خانوادگی:

شماره دانشجویی:

نکات

- استفاده از کتاب، جزوه و هر وسیله‌ای بجز خودکار ممنوع است.
- وجود هرگونه تلفن یا هر نوع وسیله‌ی الکترونیکی به‌عنوان تقلب محسوب می‌گردد.
- در طول امتحان به هیچ سوالی پاسخ داده نمی‌شود. در صورت اشتباه یا ابهام، دلیل را ذکر کنید و یا فرض خود را بنویسید.
- مدت امتحان ۱۲۰ دقیقه است.

سوال	نمره
۱	/۲۱
۲	/۱۲
۳	/۱۲
۴	/۲۴
۵	/ ۲۰
۶	/ ۸
۷	/ ۵
جمع	/۱۰۲

۱) برای هر یک از گزاره‌های 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.

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.

پ) 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 نظریه‌های جامعه‌شناسان غربی، سوادآموزی موجب از میان رفتن بیکاری نمی‌شود.

ت) 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 می‌شوند یک نفر را به تصادف انتخاب کرده و در مورد کیفیت غذا از او سوال کنیم.

ث) با توجه به این که ۴۹/۲ درصد مردم ایران مذکر هستند و ۶۴ درصد ایرانیان فوتبال تماشا می‌کنند، مردان ایرانی علاقه‌مند به فوتبال ۳۱/۵ درصد کل جامعه را تشکیل می‌دهند.

ج) میانگین درآمد خانواده‌های 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.

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

۲) برای هر یک از کارهای زیر، مشخص کنید آیا خوشه‌بندی به روش k-میانگین روش مناسبی است یا خیر.

الف) با داشتن اطلاعات تاریخی مربوط به آب و هوا، می‌خواهیم 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 شده‌اند را پیدا کنیم.

پ) با داشتن الگوهای زمانی استفاده کاربران از یک وبسایت، می‌خواهیم گروه‌های مختلف کاربران را شناسایی کنیم.

ت) با داشتن داده‌های فروش تعداد زیادی محصول در یک فروشگاه، میزان فروش هر محصول در آینده را پیش‌بینی کنیم.

۳) شما به عنوان یک دانشمند داده در تیم طراحی سامانه توصیه‌گر یک فروشگاه اینترنتی مانند دیجی‌کالا استخدام شده‌اید. 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 ویژگی‌ها عبارتند از:

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

مجموعه داده بیش از ۵۰۰ ویژگی مختلف دارد و وظیفه شما استفاده از الگوریتم‌های کاهش ابعاد جهت آماده‌سازی داده برای یک مدل یادگیری ماشین جهت پیش‌بینی محصول بعدی که مشتری احتمالا خریداری خواهد کرد، است. با توجه به خصوصیات این مجموعه داده و هدف مورد نظر، از کدام روش‌های 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 مقیاس‌پذیری مدل، به طور کامل شرح دهید.

۴) **درست یا غلط؟ (با توجیه مختصر) (۸ پرسش)**

الف) برای اینکه از overfitting یک مدل پیش‌بینی 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 پیچیده‌ی مورد نیاز را خودش یاد بگیرد و این باعث قدرتش میشود.

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

ث) اگر مدلی بین دو متغیر x و 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 پیش‌پردازش جهت استفاده از ویژگی‌های categorical ندارند.

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

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

۵) پرسش‌هایی با پاسخ کوتاه

الف) جدولی داریم از نمرات دانشجویان کلاس علوم داده که مشخصات دانشجو (نام، ش دانشجویی، مقطع، رشته) و نمره‌های امتحان و تمرینات آنها ستون‌هایش می‌باشد. برای متوسط نمره در هر رشته از چه SQL query استفاده کنیم؟

ب) ایده اصلی word embedding ها چیست؟ در مدل‌های زبانی از این‌ها چگونه برای درک متن استفاده می‌شود؟

پ) چرا از الگوریتم گرادیان کاهشی برای یافتن پارامترهای بهینه استفاده می‌شود؟ (نسبت به بقیه روش‌ها چه برتری دارد؟)

ت) دانستن توزیع مقادیر یک ویژگی در دادگان چگونه به انتخاب روش پیش‌پردازش کمک می‌کند؟

ث) برای تفسیر میزان اهمیت یک ویژگی در یک logistic regression چه می‌توان کرد؟

٦) یک انتخابات بین دو کاندیدا فرض کنید. صندوق‌های 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 ویژگی‌هایی استفاده می‌کنید.)

۷) مهمترین چالشی که در انجام پروژه پایانی درس با آن مواجه شدید چه بود و برای حل آن چه کردید؟