علی آراسته ۹۶۱۰۱۱۶۵

گزارش سری چهارم تمرین کامپیوتری آمار و احتمال

سوال ۱:

۱) در ابتدا با استفاده از دستور table2array داده‌های مربوط به متغیرهای weeks و tgrams را به ترتیب در آرایه‌های weeks_array و tgrams_array ذخیره می‌کنیم و سپس با استفاده از دستورات mean، median، var، min و max مقادیر میانگین، میانه، واریانس، مینیمم و ماکزیمم هر یک از دو متغیر را بدست می‌آوریم.

۲) یکی از روش‌های آزمون فرض آماری در مورد میانگین جامعه، آزمون تی (T-test) یا آزمون تی استیودنت (T-student) است. معمولا در سه حالت از آزمون تی برای قضاوت در مورد میانگین استفاده می‌شود.

۱. آزمون تی تک نمونه‌ای :(One sample t test) آزمون در مورد برابری میانگین جامعه یا مقدار ثابت و معلوم.
۲. آزمون تی دو جامعه مستقل :(Two independent sample t test) آزمون در مورد برابری میانگین دو جامعه مستقل.
۳. آزمون تی برای زوج متغیرها :(Paired sample t test) آزمون برابر میانگین دو متغیر از یک جامعه.

در همه این آزمون‌ها فرض نرمال بودن جامعه یا جامعه‌ها وجود دارد. همچنین ثابت بودن واریانس نیز برای هر سه حالت آزمون از فرض‌های اولیه است.

آزمون تی تک نمونه‌ای (One Sample T Test)

فرض کنید در یک جامعه نرمال می‌خواهیم میانگین جامعه را با یک مقدار مشخص (که ممکن است تحلیل‌گر حدس زده است) مقایسه کنیم. اگر میانگین جامعه را μ و مقدار حدس زده شده را با μ_0 نشان دهیم، فرضیات یعنی «فرض صفر» (Null Hypothesis) و «فرض مقابل» یا «فرض مخالف» (Alternative Hypothesis) مربوط به آزمون تک نمونه‌ای به صورت زیر نوشته می‌شوند.

$$H0: \mu = \mu0, H1: \mu \neq \mu0$$

بنابراین فرض صفر را به صورت «میانگین جامعه با مقدار μ_0 برابر است» و فرض مقابل را به صورت «میانگین جامعه با مقدار μ_0 برابر نیست» می‌خوانیم.

آماره آزمون برای حالت تک نمونه‌ای به صورت زیر نوشته می‌شود:

$$T = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

در این جا، منظور از $\bar{x}$ میانگین نمونه، μ_0 مقدار حدسی برای میانگین و s انحراف استاندارد نمونه است. همچنین n نیز تعداد نمونه را نشان می‌دهد.

آماره آزمون در این حالت دارای توزیع t با $n-1$ درجه آزادی است. یعنی داریم:

$$T \sim t(n-1)$$

براساس مقدار احتمال (p-value) واحتمال خطای نوع اول (α) می‌توان در مورد رد فرض صفر تصمیم گرفت.

آزمون تی دو نمونه‌ای مستقل (Two Independent Sample T Test)

در این حالت با دو جامعه مستقل مواجه هستیم و می‌خواهیم میانگین آن دو را با یکدیگر مقایسه کنیم. آزمون تی با فرضیات زیر صورت بگیرد.

$$H0: \mu A = \mu B, H1: \mu A \neq \mu B$$

آماره آزمون در این حالت به صورت زیر نوشته خواهد شد.

$$T = \frac{(\overline{xA} - \overline{xB}) - (\mu A - \mu B)}{\sqrt{\frac{sA^2}{nA} + \frac{sB^2}{nB}}}$$

این آماره نیز دارای توزیع t با درجه آزادی زیر است:

$$\text{degrees of fredom} = \frac{(\frac{sA^2}{nA} + \frac{sB^2}{nB})^2}{\frac{sA^4}{nA^2(nA-1)} + \frac{sB^4}{nB^2(nB-1)}}$$

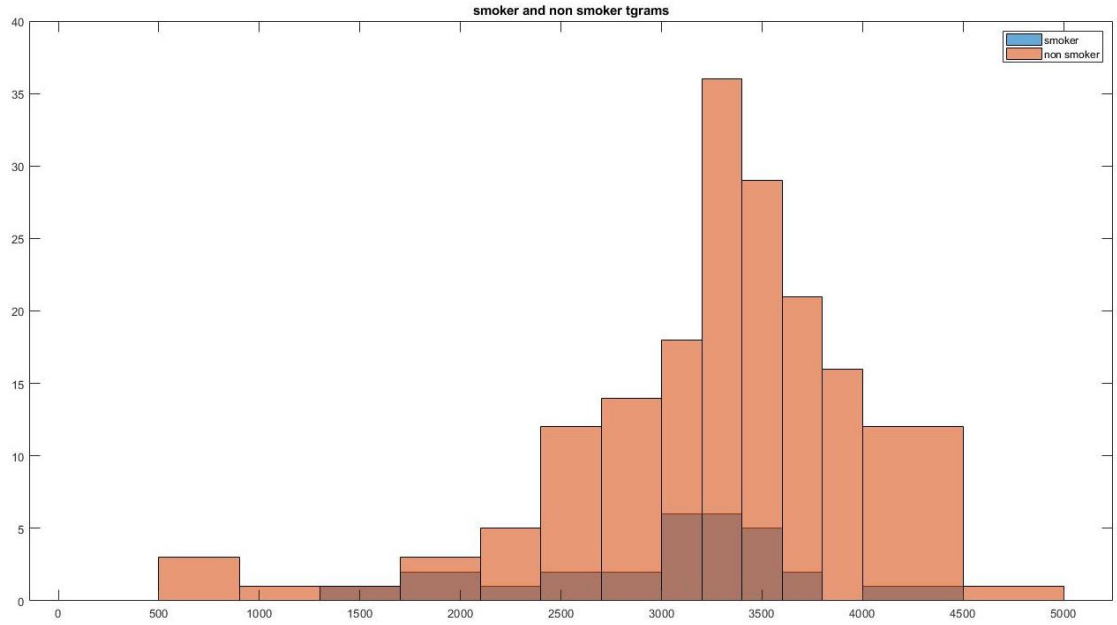
آزمون تی نمونه‌های وابسته(Paired Sample T Test)

آزمون تی دو نمونه‌ای یا آزمون تی زوجی از این جهت که برآورد واریانس در آماره آزمون آن متفاوت با حالت آزمون دو نمونه مستقل است مورد بحث قرار می‌گیرد. از این گونه آزمون بخصوص در زمانی که برای هر مشاهده دوبار اندازه‌گیری یک متغیر کمی صورت گرفته، استفاده می‌شود.

۳) الف) در این قسمت فرض صفر را $\mu = ۲۵$ و فرض مقابل را $\mu > ۲۵$ در نظر می‌گیریم. با استفاده از مطالب گفته شده در قسمت ب، مقدار T و مرز ناحیه بحرانی را بدست می‌آوریم. از آن جایی که مقدار T درون ناحیه بحرانی است یا به طور معادل مقدار p_value از α کوچکتر است، فرض صفر رد می‌شود و در نتیجه گزاره صحت دارد.

ب) در این قسمت فرض صفر را $\mu = ۳۹$ و فرض مقابل را $\mu < ۳۹$ در نظر می‌گیریم. با استفاده از مطالب گفته شده در قسمت ب، مقدار T و مرز ناحیه بحرانی را بدست می‌آوریم. از آن جایی که مقدار T درون ناحیه بحرانی است یا به طور معادل مقدار p_value از α کوچکتر است، فرض صفر رد می‌شود و در نتیجه گزاره صحت دارد.

۴) در این قسمت فرض صفر را $\mu_0 - \mu_1 = 0$ و فرض مقابل را $\mu_0 - \mu_1 < 0$ در نظر می‌گیریم. با استفاده از مطالب گفته شده در قسمت ب، مقدار T و مرز ناحیه بحرانی را بدست می‌آوریم. از آن جایی که مقدار T خارج از ناحیه بحرانی است یا به طور معادل مقدار p_value از α بزرگتر است، فرض صفر تایید می‌شود و در نتیجه گزاره صحت ندارد.



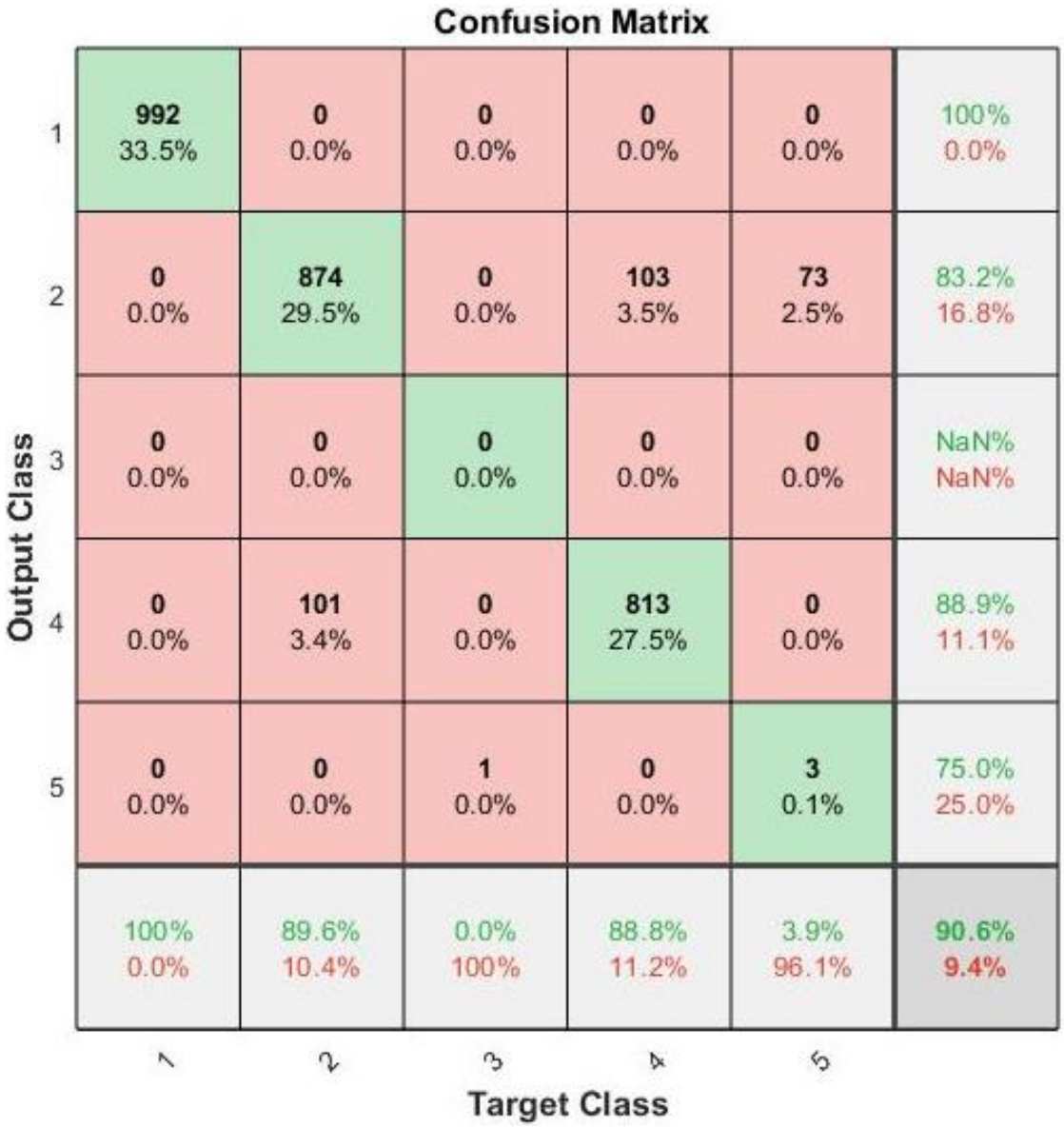
سوال ۲:

۱) ابتدا با استفاده از دستور load، data را در برنامه لود می‌کنیم. سپس با استفاده از دستور table2array داده‌های مربوط به متغیرهای مختلف را در آرایه‌هایی همنام با متغیرها ذخیره می‌کنیم. در متغیر children، مقدار NaN را با مقدار ۴ عوض می‌کنیم. در ادامه با استفاده از دستور unique، تعداد مقدارهای مختلف هر متغیر را می‌یابیم. حال با استفاده از حلقه‌های for، به هر یک از مقادیر مختلف هر یک از متغیرها یک عدد طبیعی نسبت می‌دهیم. در انتها با کنار هم قرار دادن آرایه‌های مربوط به متغیرهای مختلف، ماتریس عددی data را می‌سازیم.

۲) ابتدا با استفاده از دستور randperm برداری حاوی جایگاه هر نمونه می‌سازیم. سپس با استفاده از بردار بدست آمده، ماتریس random_order_data را از روی ماتریس data_numeric_array می‌سازیم. سپس train داده اول ماتریس بدست آمده را به بخش آموزش و بقیه را به بخش تست اختصاص می‌دهیم.

۳) ابتدا سطرهای مربوط به هر یک از مقادیر متغیر class و از روی آن احتمال هر یک از این مقادیر را بدست می‌آوریم. در ادامه برای هر یک از ویژگی‌ها یک ماتریس که تعداد سطر آن برابر تعداد مقادیر متغیر class و تعداد ستون آن برای تعداد مقادیر آن متغیر است، در نظر می‌گیریم و در خانه (i,j) آن، احتمال j امین مقدار آن متغیر به i امین مقدار متغیر class را قرار می‌دهیم.

۴) ابتدا با استفاده از فرمول بدست آمده، احتمال هر یک از مقادیر متغیر class به شرط رخ دادن هر یک از نمونه‌های تست را بدست می‌آوریم. سپس ماکزیمم این مقادیر و شماره مربوط به مقداری از متغیر class که به ازای آن مقدار ماکزیمم بدست آمده است، برای هر یک از نمونه‌ها ذخیره می‌کنیم. این شماره‌ها همان مقادیر پیش‌بینی شده برای مقدار متغیر class هستند. در انتها با استفاده از دستور plotconfusion، confusion matrix مربوط به دسته‌بندی انجام شده را رسم می‌کنیم.



سوال ۳:

آ) ابتدا با استفاده از دستور load، data را در برنامه لود می‌کنیم. سپس با استفاده از دستور table2array داده‌های مربوط به متغیرهای مختلف را در آرایه‌هایی همنام با متغیرها ذخیره می‌کنیم. در ادامه با استفاده از دستور unique، تعداد مقدارهای مختلف هر متغیر را می‌یابیم. حال با استفاده از حلقه‌های for، به هر یک از مقادیر مختلف هر یک از متغیرها غیر عددی یک عدد طبیعی نسبت می‌دهیم. در انتها با کنار هم قرار دادن آرایه‌های مربوط به متغیرهای مختلف، ماتریس عددی data را می‌سازیم.

ب) ابتدا با استفاده از دستور randperm برداری حاوی جایگاه هر نمونه می‌سازیم. سپس با استفاده از بردار بدست آمده، ماتریس random_order_data را از روی ماتریس data_numeric_array می‌سازیم. سپس train داده اول ماتریس بدست آمده را به بخش آموزش و بقیه را به بخش تست اختصاص می‌دهیم.

ج) با استفاده از دستور fitlm روش داده‌های آموزش (train) مدل خطی فیت می‌کنیم تا متغیر area را برحسب سایر متغیرها تخمین بزنیم. در اطلاعات خروجی این تابع در ستون Estimate، ضرایب فیت برای هر یک از ترم‌های مدل خطی، در ستون SE تخمین خطای آن ضریب، در ستون t_stat آماره t برای تست کردن این فرضیه که آن ضریب می‌توانست صفر باشد و در ستون p_value احتمال این که به طور تصادفی آن ضریب مقدار کنونی خود را داشته باشد، آمده است. بنابراین هر چه p_value بیشتر باشد، آن ویژگی کمتر در پیش‌بینی مساحت سوخته شده، تاثیر دارد. لذا در این دیتاست، به ترتیب داده‌های DMC، X و month بیشترین تاثیر را دارند.

د) ضریب تعیین (R_squared) مقدار متناسب تغییرات در متغیر جواب y را نشان می‌دهد که در مدل رگرسیون خطی به وسیله متغیر مستقل X بیان می‌شود. هر چه مقدار R_squared بیشتر باشد تنوع بیشتری با مدل رگرسیون خطی بیان می‌شود.

۵) با استفاده از mdl.Fitted و دیتا بخش آموزش دیتاست، با یک حلقه for، MSE مورد نظر را محاسبه می‌کنیم.

۶)با استفاده از mdl.predict و دیتا بخش تست دیتاست، با یک حلقه for، MSE مورد نظر را محاسبه می‌کنیم.