بنام خدا

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر

درس سیستم‌های هوشمند

تمرین شماره پنج

دی 1401

فهرست سوالات

سوال 1 – سوالات آماری (پیاده سازی)

بخش 1

در این سوال می خواهیم به بررسی problem Secretary بپردازیم. در این مسئله فرض می شود که یک مدیر داریم که می خواهد از بین n کاندید، بهترین آنها را انتخاب کند. کاندید ها یکی پس از دیگری و به صورت تصادفی مصاحبه می شوند. تصمیم قبول یا رد شدن هر کاندید بلافاصله بعد از مصاحبه گرفته می شود. هنگامی که یک کاندید رد می شود، امکان قبولی دوباره آن شخص وجود ندارد. در حین مصاحبه، مدیر اطلاعات لازم برای طبقه بندی کاندیدهایی که تاکنون مصاحبه شده اند را به دست می آورد اما از کیفیت کاندید های بعدی ناآگاه است. هدف در این مسئله، یافتن بهرین استراتژی برای بیشینه کردن احتمال گزینش بهترین کاندید در بین n کاندید است. یک استراتژی، رد کردن k کاندید اول و سپس انتخاب اولین کاندیدی است که از این k کاندید رد شده بهتر باشد. اما باید توجه کرد که اگر k عددی کوچک باشد، اطلاعات کافی برای انتخاب بهترین کاندید را نداریم و اگر k بزرگ باشد، امکان اینکه بهترین کاندید در بین k کاندید رد شده باشد افزایش مییابد

1. به ازای $n = 100$ یک لیست از اعداد ۰ تا $n - 1$ تولید کنید. این لیست نشان دهنده n کاندید است که کیفیت این کاندیدها در این لیست آورده شده اند. حال به ازای $k = range(5, 101, 5)$ و برای ۱۰۰۰۰ آزمایش، احتمال انتخاب بهترین کاندید را به ازای رد کردن k کاندید اول به دست آورید و در یک scatter plot رسم کنید، سپس بهترین عدد به دست آمده برای k را گزارش کنید .
2. به ازای $k = \frac{n}{e}$ که e عدد نپر است و به ازای ۳ تا ۱۰۰ کاندید، احتمال انتخاب بهترین کاندید را بعد از رد کردن k کاندید اول در ۱۰۰۰۰ آزمایش به دست آورید و scatter plot این احتمالات را رسم کنید.
3. (امتیازی) به صورت دستی، احتمال انتخاب بهترین کاندید را به دست آورید

بخش 2

در این بخش به بررسی قضیه حد مرکزی می‌پردازیم. در ابتدا تابع Distribution از فایل Dists.py را به برنامه خود وارد کنید و با وارد نمودن سه رقم آخر شماره دانشجویی خود به عنوان ورودی این تابع، یک جمعیت با اندازه 100،000 به عنوان خروجی به دست آورید. حال از این جمعیت به صورت تصادفی n نمونه به دست آورید و میانگین این n نمونه را محاسبه کنید. سپس برای s بار این کار را انجام داده و نمودار میانگین این نمونه‌ها را رسم کنید. برای قسمت بعد، دادگان wine.csv که در اختیار شما قرار داده شده است را در نظر بگیرید. ابتدا نمودار تابع توزیع ستون 12 این دادگان را رسم کنید. سپس برای این جمعیت نیز در هر مرحله، یک نمونه تصادفی با اندازه n های مختلف از این متغیرهای تصادفی بردارید و این کار را s بار تکرار کنید. سپس، نمودار میانگین این نمونه‌ها را رسم کنید و با نمودار نرمال مقایسه کنید. همچنین، میانگین و انحراف معیار این نمونه‌ها را با مقداری که از قضیه حد مرکزی به دست می‌آید، مقایسه کنید.

سوال 2 – سوالات آماری (تحلیلی)

بخش 1

احمد یک دانشجو یادگیری ماشین است، او می‌داند که در بیزساده انگارانه[1] یک فرض مهم، فرض استقلال ویژگی‌ها است که البته در بسیاری از موارد واقعی، این فرض برقرار نیست. حال احمد در یک مسئله دنیای واقعی برای حل این مشکل، با استفاده از یک روش آمار و احتمالی، داده ها را به ابعاد جدیدی منتقل می‌کند (داده ها با ویژگی های جدیدی توصیف می‌شوند) و سپس از بیز ساده انگارانه استفاده می‌کند. او استدلال میکند که در فضای جدید باتوجه به عمود بودن و غیرهمبسته بودن ویژگی‌ها، اکنون فرض بیزساده انگارانه صحیح است. نظر شما در مورد استدلال احمد چیست؟ آیا استدلال او درست است؟ لطفا تحلیل کنید.

[1] Naïve Bayes

بخش 2

فرض کنید داده های $\{x_1, x_2, \dots, x_n\}$ را در اختیار داریم. این داده ها نمونه هایی i.i.d[1] از توزیع نرمال تک متغیره هستند. با استفاده از Maximum Likelihood Estimation ، پارامتر های توزیع (میانگین و واریانس) را بدست آورید.

سوال 3 – بیز ساده‌انگارانه[2]

دسته ای از دادگان در فایل های پیوست شده موجود می‌باشد که در آن‌ها، اعداد 0 تا 9 به صورت دستی نگاشته شده‌اند. هدف آن است که با استفاده از الگوریتم بیز ساده‌انگارانه، درصد دقت[3] تشخیص اعداد در دادگان آزمون[4] "validationimages" بدست آید. به نکات زیر توجه نمایید.

- دادگان آموزش در "trainingimages" گردآوری شده‌اند. برچسب داده‌های آزمایش و داده‌های آزمون را می توانید به ترتیب در "traininglabels" و "validationlabels" بیابید.
- فایلی با عنوان "NB.ipynb" نیز، به عنوان راهنمایی در نظر گرفته شده است. استفاده از این فایل الزامی نمی‌باشد.
- لینک‌ها و مراجع خود را حتما ذکر نمایید.

1) حداقل سه ویژگی[5] برای طراحی طبقه‌بند[6] خود معرفی نمایید. در صورت وجود داده های کمّی، نحوه بدست آوردن احتمالات لازم برای الگوریتم را توضیح دهید.

2) درصد دقت تشخیص اعداد در دادگان آزمون را بدست آورید. در صورتی که در الگوریتم از ابرمتغیری[7] استفاده می‌کنید، مقدار آن را گزارش دهید.

[1] independent and identically distributed
[2] Naïve Bayes
[3] Accuracy
[4] Test dataset
[5] Feature
[6] Classifier
[7] Hyperparameter

نکات تحویل:

- مهلت تحویل این تمرین تا **20 دی** میباشد.
- انجام این تمرین به صورت یک نفره است.
- برای انجام این تمرین تنها مجاز به استفاده از زبان برنامه نویسی پایتون هستید.
- در صورت وجود تقلب نمره تمامی افراد شرکت کننده در آن 100- لحاظ میشود.
- لطفا پاسخ تمرین خود را (به همراه کد/گزارش سوال کامپیوتری) به صورت زیر در صفحه درس آپلود نمایید:

HW [HW number] _ [Last name] _ [Student number].zip

- در صورت وجود هر گونه ابهام یا مشکل میتوانید از طریق ایمیل با مسئولان حل تمرین در تماس باشید.

سوال 1: روزبه نهاوندی (roozbeh.nahavandi@ut.ac.ir)

سوال 2: محمدرضا تیموریان فرد (mr.teymoorian@ut.ac.ir)

سوال 3: دریا افضلی (darya.afzali@ut.ac.ir)