به نام خدا

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر

درس سیستم‌های هوشمند

تمرین شماره ۴

آذر ۱۴۰۰

فهرست سوالات

سوال ۱: تحلیلی

در این سوال هدف پیاده‌سازی دستی الگوریتم های خوشه‌بندی است.

الف : خوشه‌بندی با روش کا-میانگین

با استفاده از روش خوشه‌بندی کا-میانگین[۱]، داده های زیر را به ۲ خوشه تقسیم کنید.

نقاط ابتدایی را به صورت تصادفی در نظر بگیرید و مراحل را تا پایدار شدن خوشه‌بندی ادامه دهید.

برای دست‌یابی به دید بهتری از مساله، نقاط را رسم کنید و مرکز خوشه ها را مشخص کنید.

جدول ۱-۱: داده‌های مربوط به سوال الف کا-میانگین

i	x_1	x_2
A	1	1
B	1	0
C	0	2
D	2	4
E	3	5

ب : خوشه‌بندی سلسله‌مراتبی

با استفاده از روش پیوند واحد[۲] و محاسبه‌ی فاصله با نرم اقلیدسی[۳]، داده های زیر را خوشه بندی کرده و نمودار درختی[۴] آن را رسم کنید.

جدول ۲-۱: داده‌های مربوط به سوال ب پیوند واحد

i	x	y
P_1	0.22	0.38
P_2	0.35	0.32
P_3	0.26	0.19
P_4	0.08	0.41
P_5	0.45	0.30

[۱] K-Means Clustering
[۲] Single Link
[۳] L2-Norm
[۴] Dendrogram

سوال ۲: پیاده‌سازی الگوریتم خوشه‌بندی

در این سوال به پیاده سازی الگوریتم خوشه بندی کا-میانگین می‌پردازیم. همانطور که در کلاس مطرح شد، این الگوریتم از دسته الگوریتم‌های بدون نظارت است و سعی دارد تابع هزینه‌ای که به صورت مجموع فاصله‌ی نمونه‌های متعلق به هر خوشه تا مرکز آن خوشه تعریف می‌شود را کمینه کند. در این سوال با در نظر گرفتن نرم اقلیدسی و مجموعه داده iris الگوریتم را پیاده‌سازی می‌نماییم.

الف: تاثیر تعداد خوشه‌ها

تعداد تکرارهای الگوریتم را برابر با ۱۵۰ و تعداد خوشه‌ها را $\{5, 10, 20\}$ در نظر بگیرید. سپس تحلیل کنید که مناسب‌ترین تعداد خوشه‌ها کدام است (می توانید از نسبت شباهت درونی به شباهت بیرونی استفاده نمایید). همچنین نمودار مقدار هزینه را در حین اجرای الگوریتم رسم نمایید.

ب: تاثیر تکرار آزمایش

برای هر تعداد خوشه ذکر شده در بند پیشین، الگوریتم را به تعداد کافی تکرار کنید و بر اساس نتایج بدست آمده، میانگین و واریانس الگوریتم را برای هرکدام از تعداد خوشه‌ها محاسبه نمایید.

سوال ۳: یادگیری نیمه‌نظارت شده (امتیازی)

در این سوال بر روی دادهٔ Surgical کار می‌کنید. قصد داریم بر روی این داده در حالتی که برچسب بخش زیادی از داده را نداریم عمل طبقه بندی باینری را انجام دهیم. به این صورت که ستون عوارض[۵] را بعنوان هدف در نظر گرفته (که این ستون نشان از وخامت بیماری دارد) و باقی را بعنوان فضای ویژگی در نظر بگیرید.

الف: رگرسیون لجستیک

ابتدا مختصرا در رابطه با رگرسیون لجستیک[۶] بعنوان یک طبقه بند توضیح دهید. سپس داده را به این صورت تقسیم کنید که بصورت تصادفی یک درصد از داده برای آموزش (دارای برچسب)، ۷۴ درصد بدون برچسب و ۲۵ درصد برای آزمون باشد. با بررسی دادهٔ آموزش هیستوگرام مرتبط با طبقه‌ها را رسم نمایید و نامتعادل[۷] بودن آن را ارزیابی کنید.

ب: ارزیابی طبقه‌بند

طبقه‌بند خود را آموزش داده و نتایج (دقت و معیار F1[۸] و ماتریس آشفتگی) را برای آن بیان نمایید.

ج: یادگیری نیمه‌نظارت شده:

در این قسمت می خواهیم به کمک الگوریتم خود تعلیم[۹] که یک الگوریتم یادگیری نیمه‌نظارت شده می‌باشد از داده ی بدون برچسب نیز در آموزش طبقه بند خود کمک بگیریم. ابتدا این الگوریتم را توضیح داده و سپس آن را بدون استفاده از کتابخانه آماده پیاده‌سازی کنید. در ادامه نتایج بدست آمده برای دادهٔ آزمون را بدست آورید. سپس با مقایسه با بخش قبلی تحلیل کنید این روش چه کمکی به طبقه‌بند ما کرد؟

د: شرایط استفاده:

در رابطه با شرایط توقف الگوریتم خود تعلیم توضیح داده و با تغییر حد آستانهٔ رگرسیون لجستیک توضیح دهید که به نظر شما چه آستانه‌ای برای این داده مناسب است. هم چنین در هر تکرار از الگوریتم مشخص کنید که چه تعداد دادهٔ جدید به مجموعه آموزش اضافه می‌شود و معیار F1 را در طول تکرار مشخص کرده و نمودارهای مرتبط با این دو را رسم نمایید.

---

[۵] Complication
[۶] Logistic Regression
[۷] Imbalanced
[۸] F1 Score
[۹] Self-Training

نکات سوال سوم:

****نکته**: برای پیاده سازی الگوریتم خودتعلیم مجاز به استفاده از کتابخانه های آماده نمی‌باشید. برای باقی قسمت ها (رگرسیون لجستیک) و بررسی نتایج و ... در استفاده از کتابخانه های آماده یادگیری ماشین در پایتون (Sklearn) و toolkit ها و function های متلب آزاد هستید.**

****نکته**: برای این سوال از روش های کاهش بعد یا پیش‌پردازش های معمول استفاده نکنید و از تمامی داده‌ها به صورت خام استفاده کنید.**

سوال ۴: مقدمات احتمال

الف: سوالات تحلیلی

الف-۱:

می خواهیم بین دو نفر، یکی را به صورت تصادفی (با احتمال یکنواخت) انتخاب کنیم. الگوریتمی ارائه دهید که بتوانیم با استفاده از یک سکه‌ی خراب که با احتمال p شیر می‌آید، این کار را انجام دهیم.

الف-۲:

دو دانشجو تصمیم گرفته‌اند از وسایل نقلیه عمومی استفاده کنند. اولی از مترو و دومی از اتوبوس استفاده میکند. مدت زمانی که اولی منتظر مترو و دومی منتظر اتوبوس می‌ماند را به ترتیب با متغیرهای تصادفی X و Y که از توزیع نمایی با میانگین‌های متفاوت پیروی می کند، نشان میدهیم.

الف-۲-۱: مقدار کواریانس زیر را به دست آورید.

$$cov(min(X,Y), \max(X,Y))$$

الف-۲-۲: این دو دانشجو به دانشگاه می‌رسند و سر دو جلسه امتحان مختلف می‌نشینند. اگر نمراتشان را با U و V که از توزیع نرمال پیروی می‌کنند، نشان دهیم، دو مقدار کواریانس زیر را به دست آورید.

$$cov(V, \max(U,V))$$
$$cov(V, min(U,V))$$

الف-۳:

دو کارمند یک شرکت قرار ا ست یک برنامه‌ی تحلیل داده بنویسند. نحوه‌ی کدزنی آنها به این صورت است که به صورت نوبتی این کار را انجام می‌دهند و هرگاه یک نفر در حال کد زدن باشد، دیگری استراحت می‌کند. فرض کنیم X, Y به ترتیب متغیر تصادفی زمان‌های صرف شده روی این برنامه توسط کارمند اول و دوم باشد. همچنین، فرض کنید تابع pdf مشترک این دو متغیر تصادفی به صورت زیر می‌باشد:

$$f(x,y) = c\,(1 - x - y); \quad x > 0, \quad y > 0, \quad x + y < 1$$

مقادیر زیر را به دست آورید و توضیح دهید که این مقادیر چه چیزی را نشان می‌دهند

$$\Pr(X < 0.5)$$
$$E[X + Y]$$

ب: سوالات شبیه‌سازی

ب-۱:

در این سوال قصد داریم، با شبیه‌سازی مسئله تولد[۱۰]، نتایج آن را بررسی کنید. در این مسئله، احتمال اینکه در یک جمع n نفره، حداقل یک زوج تولد یکسانی داشته باشند، محاسبه می‌شود. برای محاسبه این احتمال، فرض کنید هر فرد، با احتمال یکسانی ممکن است در هر روزی از سال به دنیا آمده باشد و تعداد روزهای سال را ۳۶۵ در نظر بگیرید. به ازای یک n خاص، ۱۰۰۰۰ آزمایش انجام داده و تعداد آزمایش‌هایی که در آنها حداقل دو نفر متولد یک روز هستند را شمرده و احتمال این رخداد از روی آن به دست آورید. در واقع در هر آزمایش، n فرد با روز تولد تصادفی تولید خواهید کرد و داشتن روز تولد یکسان میان آنها را بررسی خواهید کرد. این کار را برای مقادیر مختلف n از ۱۰۰ انجام داده و نمودار احتمال بر حسب n را رسم کنید. می‌توانید از لینک زیر برای فهم بهتر مسئله استفاده کنید.

ب-۲:

در این مسئله به بررسی صحت قضیه حد مرکزی خواهید پرداختم. برای اینکار، ابتدا فرض کنید جامعه مورد بررسی ما از متغیر تصادفی نمایی با نرخ ۲ پیروی می‌کند. در هر مرحله، یک نمونه تصادفی با اندازه n 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. بار دیگر، همان کارهای قبلی را انجام دهید ولی این بار فرض کنید متغیر تصادفی جامعه، از نوع binomial(20, 0.8) random variable باشد. نتیجه‌خود را از این شبیه‌سازی بیان کنید.

[۱۰] Birthday Problem

نکات:

- مهلت تحویل این تمرین، سه‌شنبه ۷ دی است.
- انجام این تمرین به صورت یک نفره است.
- برای انجام تمرین‌ها فقط مجاز به استفاده از زبان های برنامه نویسی Python و MATLAB خواهید بود. در سوالاتی که از شما خواسته شده است یک الگوریتم را پیاده سازی کنید مجاز به استفاده از توابع آماده نمی‌باشید مگر اینکه در صورت سوال مجاز بودن استفاده از این توابع یا کتابخانه‌ها صریح ذکر شده باشد.
- کدهای مربوط به هر تمرین می‌بایست در پوشه‌ای با نام Codes در کنار گزارش کار شما موجود باشد. این کدها باید خوانا و به صورت مرتبط نام گذاری شده باشند، لذا توضیحات لازم را به صورت یادداشت[۱۱] در کدهای خود قرار دهید.
- گزارش شما در فرآیند تصحیح از اهمیت ویژه‌ای برخوردار است. لطفاً تمامی نکات و مفروضاتی که برای پیاده‌سازی‌ها و محاسبات خود در نظر می‌گیرید را در گزارش ذکر کنید.
- الزامی به ارائه توضیح جزئیات کد در گزارش نیست، ولیکن تحلیل و تفسیر نتایج بدست آمده الزامی است.
- گزارش‌ها تنها در قالب تهیه شده که روی صفحه درس در سامانه Elearn بارگذاری شده، تصحیح خواهند شد و به قالب‌های دیگر نمره‌ای تعلق نخواهد گرفت.
- در گزارش استفاده از زیرنویس برای تصاویر و بالانویس برای جداول الزامی است.
- در صورت مشاهدهٔ تقلب نمرات تمامی افراد شرکت‌کننده در آن ۱۰۰- لحاظ می‌شود.
- لطفا گزارش ، فایل کدها و سایر ضمائم مورد نیاز را با ترتیب نام‌گذاری زیر در صفحه درس در سامانه یادگیری الکترونیکی بارگذاری نمائید.

HW[HW Number]_[LastName]_[StudentNumber].zip

- در صورت وجود هرگونه ابهام یا مشکل می‌توانید از طریق رایانامه‌های زیر با دستیاران آموزشی مربوطه در تماس باشید:

سرکار خانم نورزاد – سوال ۱ – رایانامه { njnoorzad@gmail.com }

جناب آقای رکنی – سوال ۲ – رایانامه { a.rokni@ut.ac.ir }

جناب آقای طلاکوب – سوال ۳ – رایانامه { rezatalakoob@yahoo.com }

جناب آقای طلاکوب – سوال ۴ – رایانامه { salar.nouri@ut.ac.ir }

[۱۱] comment