به نام خداوند بخشنده مهربان

تمرین و پروژه شماره ۱

درس داده کاوی

عنوان: استفاده از الگوریتم های درخت های تصمیم

استاد درس: حاجی محمدی

تاریخ تحویل: ۱۴۰۱/۰۱/۱۸

I. مقاله‌های زیر را مطالعه کرده و خلاصه آنهارا در دو صفحه بیان کنید. ایده های مطرح شده و خلاصه ای از جمع بندی مقاله حتما آورده شود. همچنین تحلیل خود را در مورد مقاله (نقاط قوت و ضعف) بیان کنید

- نیازی به آوردن فرمول ها، جداول و نمودار ها نمی باشد.
- فقط ترجمه ی بخش های چکیده ویا جمع بندی مقاله مورد قبول نیست.
- تحلیل های قوی تر، نمره اضافی خواهد داشت.

- **Student Academic Performance Prediction by using Decision Tree Algorithm**
- **Classification Based on Decision Tree Algorithm for Machine Learning**

II. در جدول زیر یک دسته داده‌های استاندارد از یک فروشگاه ***Mp3 Player*** جمع آوری شده است. هر نمونه توسط سه ویژگی نمایش داده شده است. عمل طبقه بندی بر اساس فیلد هدف ***"Customer Satisfaction"*** در دو دسته 'yes' و 'no' صورت می گیرد. درخت تصمیم گیری را جهت طبقه این دسته از داده ها به صورت دستی و بر اساس ***Information Gain*** و ***Gain Ratio*** ایجاد کنید. تشریح جزییات در هر مرحله الزامی است. قوانین به دست‌آمده از درخت تصمیم را تشریح نمایید.

Memory	Battery life	Price	Customer satisfaction
<=4	long	<=150	yes
>4	long	>150	yes
>4	long	<=150	yes
<=4	long	>150	yes
>4	long	>150	yes
>4	low	>150	yes
<=4	low	>150	no
<=4	low	>150	no
>4	low	<=150	yes
<=4	low	<=150	no
<=4	medium	<=150	no
>4	medium	<=150	no
<=4	medium	>150	yes
>4	medium	>150	yes
>4	medium	<=150	no

III. انجام این تمرین جهت آشنائی بیشتر با یادگیری درخت تصمیم از طریق استفاده از آن در یک مثال عملی است.

برای انجام این تمرین از دسته بندی کننده درخت تصمیم در نرم افزار **Weka** استفاده کنید.

اطلاعات بیشتر در مورد این برنامه (دریافت برنامه به همراه کد آن، راهنمای استفاده، نحوه کار با واسط گرافیکی و ...) در آدرس زیر موجود است :

- WEKA Machine Learning Project
- https://www.tutorialspoint.com/weka/weka_tutorial.pdf
- https://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf
-

داده مورد استفاده:

برای انجام این تمرین از مجموعه داده **Census Income Data Set** از دیتابیس UCI استفاده کنید. این مجموعه داده (شامل مشخصات کلی آن، مجموعه داده های آموزش، مجموعه داده های تست و ...) از طریق آدرس زیر قابل دسترسی است:

http://archive.ics.uci.edu/ml/datasets/Census+Income

در پاسخ به سؤالات، باید موارد زیر رعایت گردند:

1. گزارش به تفکیک شماره سؤال تنظیم شود.
2. در صورت وجود هرگونه فرض خاص در پاسخ به سؤالات، آن را در متن پاسخ سؤال ذکر نمایید.
3. در هر سؤال باید پارامترهای تنظیم شده در نرم افزار Weka برای اجرای الگوریتم ها ذکر گردد. (در صورت استفاده از پارامترهای تنظیم شده در یک سؤال برای سؤال بعدی، باید این مسئله در سؤال بعد صریحاً ذکر شود.)
4. کلیه فایل ها و داده های تولید شده در مراحل انجام هر سؤال (درصورت وجود) و خروجی های بدست آمده(شامل خروجی های خام، جداول ویا نمودارها) باید به تفکیک شماره سؤال(در folder جداگانه) درگزارش قرار داده شوند. باید در متن پاسخ سؤال به این فایل ها ارجاع داده شود.
5. برای سؤال 4 و 5، در صورت نیاز به تغییر کد در Weka ، فایل های تغییر یافته را در ضمیمه گزارش قرار داده و بخش تغییر یافته از کد را در صورتی که کوچک باشد(حداکثر چند خط کد)، در متن گزارش بیاورید(توجه : انجام پیاده سازی در سؤالاتی که نیاز به تغییر کد دارند، نمره ویژه خواهد داشت).

6. نیازی به تحلیل نتایج هر سؤال با جزئیات وجود ندارد اما باید برای پاسخ هر سؤال، نتیجه گیری مختصر (در حد یک پاراگراف کوتاه) ارائه گردد.

سوالات

1. نمونه ها را بصورت تصادفی به 10 مجموعه مساوی ولی غیر تکراری تقسیم کنید. بازای هر یک از 9مجموعه حاصل یک درخت تصمیم ایجاد کرده و از مجموعه دهم برای تست استفاده کنید. نموداری از خطاهای حاصل را رسم کرده و نتیجه را بحث کنید.
2. عملکرد درخت را بر اساس تغییر تعداد نمونه‌های آموزشی به صورت نمودار رسم کنید.(منحنی یادگیری)
3. نتیجه را برای درخت های هرس شده و هرس نشده مقایسه کنید. همچنین اختلاف اندازه های درخت هرس شده و نشده را مشخص کنید. (نمره اضافی)
4. برای انتخاب ویژگی مناسب برای هر گره علاوه بر روش استفاده از نسبت بهره (***Gain Ratio***) که به صورت پیش فرض در الگوریتم J48 در Weka استفاده می شود، روش‌های زیر را نیز پیاده-سازی کنید:
   - تصادفی ***Random*** که در آن یکی از ویژگی ها بصورت تصادفی انتخاب می‌شوند.
   - بهره اطلاعات (***Information Gain***)
5. با انتخاب تعداد ثابتی مثال یادگیری، برنامه را با استفاده از معیار تصادفی، بهره اطلاعات و نسبت بهره بر روی یک مجموعه داده اجرا کرده و نتیجه را مقایسه کنید. برای معیار تصادفی هیستوگرام سایز درخت تولید شده را برای 20 بار تکرار آزمایش رسم کنید.

IV. انجام این تمرین جهت آشنائی بیشتر با یادگیری درخت تصمیم از طریق استفاده از آن در یک مثال عملی است.

جهت انجام بخش پیاده سازی، از Jupyter استفاده کنید و فایل نهایی را با پسوند . ipynb آپلود کنید.

همچنین شماره دانشجویی خود را به عنوان نام فایل در نظر بگیرید.

داده مورد استفاده:

برای انجام این تمرین از مجموعه داده **Car Evaluation Data Set** از دیتابیس UCI استفاده کنید. این مجموعه داده (شامل مشخصات کلی آن، مجموعه داده های آموزش، مجموعه داده های تست و ...) از طریق آدرس زیر قابل دسترسی است:

https://archive.ics.uci.edu/ml/datasets/Car+Evaluation

سوالات

الف) الگوریتم درخت تصمیم را بر روی مجموعه داده ی ذکر شده اجرا کنید. دقت و خطای آموزش و تست را گزارش کنید .

ب) با استفاده از روش K-Fold Cross Validation ، عمق بهینه‌ی درخت را محاسبه کنید.

ج) الگوریتم درخت تصمیم را این بار با عمق بهینه‌ی حاصل از قسمت ب بر روی دیتاست پیاده کنید. خروجی را با قسمت الف مقایسه کرده ونتایج را تفسیر کنید.

آنچه که تحویل میدهید:

1. خلاصه مقاله‌های مطالعه شده
2. گزارشی به فرمت فایل PDF شامل پاسخهای داده شده هر یک از سوالات به همراه فایلها وسورس کدهای لازم
3. فایل pdf و کد را بصورت یکجا در قالب یک فایل zip در سامانه کورسز آپلود کنید (نام فایل = شماره دانشجویی2&شماره دانشجویی1#HW1).

دانشجویان عزیز توجه داشته باشند که برای بوجود آمدن امکان ارزیابی صحیح کار انجام شده، لازم است تا هر دانشجو آزمایشات را جداگانه و یا فقط در گروه‌های تعیین شده انجام داده و نتایج جداگانه ای را نیز ارائه نماید.

***توجه** : هرگونه شباهت در گزارش و پاسخ تشریحی به منزله تقلب است و کل نمره تمرین صفر است. (میتوانید از اینترنت به عنوان منبع کمکی هم در سوالا ت تشریحی و هم در سوالا ت پیاده سازی استفاده کنید با ذکر منبع و ارجاع به آن، اما کپی برداری ممنوع است و نمره ی صفر تعلق میگیرد.) به گزارشات یکسان یا مشابه، نمره ای تعلق نمی گیرد.*

در ضمن بصورت تصادفی از یک و یا چند دانشجو خواسته خواهد شد تا کار انجام شده را در کلاس درس توضیح بدهند.

امکان تمدید زمان ارسال پروژه وجود ندارد. لذا لطفا دوستان قبل از 59 : 23 تاریخهای مقرر، نسبت به بارگذاری پروژه اقدام فرمایند.

هر گونه سوال در مورد نحوه استفاده از نرم افزار Weka ، پایتون و یا سوالات را با تدریس یاران درس در میان بگذارید.

موفق باشید