به نام خدا

دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)
دانشکده مهندسی کامپیوتر و فناوری اطلاعات

تمرین سری دوم داده کاوی – بخش پیاده سازی

توضیحات:

- پاسخ به تمرین‌ها باید به صورت انفرادی صورت گیرد و در صورت مشاهده هرگونه تقلب نمره صفر برای کل تمرین منظور خواهد شد.
- تمیزی و خوانایی گزارش تمرین از اهمیت بالایی برخوردار است.
- گزارش تمرین خود را در قالب یک فایل PDF با نام **«HW2_StudentNumber.pdf»** به همراه کد های بخش پیاده سازی (فایل های ipynb. یا py.) در فایلی به نام **«HW2_StudentNumber.zip»** قرار داده و در سایت درس در مهلت معین بارگذاری نمایید.
- توجه داشته باشید که به سوالات پیاده سازی بدون گزارش نمره ای تعلق نمی گیرد.
- در صورت داشتن اشکال می‌توانید از طریق ایمیل **datamining.fall2020@gmail.com** با تدریس‌یاران درس در ارتباط باشید.
- همچنین لازم بذکر است که اگر مواردی در کلاس تدریس نشده انتظار می رود که خود دانشجویان جستجو کنند و انجام دهند.

نیم‌سال اول ۰۰-۹۹

سوال ۶- درخت تصمیم

مجموعه داده ما در این بخش داده های کشتی تایتانیک است و هدف پیش بینی زنده ماندن یا نماندن سرنشینان کشتی می‌باشد. در این بخش میخواهیم با استفاده از درخت تصمیم مدلی را آموزش دهیم که بتواند با دقت بالایی این پیش‌بینی را انجام دهد (بالای ۷۵ درصد). برای این منظور ۲ مرحله را باید طی کرد:

۱- آماده سازی داده ها: همانطور که می‌دانید الگوریتم درخت تصمیم برای داده های عددی است، در نتیجه داده های عددی را باید به داده های عددی تبدیل کنید. (انتخاب روش تبدیل با خودتان است)

۲- آموزش (training) و دسته بندی (classification): در مرحله آخر باید مدل را با داده های train آموزش داده و دقت آن را توسط داده های test ارزیابی کنید. برای این منظور از کتاب خانه موجود scikit-learn استفاده کنید. همچنین در انتها برنامه شما باید **درخت تصمیم نهایی را رسم نماید**. با تغییر پارامتر هایی چون عمق درخت، تابع تقسیم (gini, entropy) **دقت های مختلف را اندازه گیری کرده و گزارش کنید (۴ حالت مختلف)**

توجه: داده ها به صورت پیش فرض به دو دسته train و test تقسیم شده اند (test.csv , train.csv) و در فایل زیپ مجموعه داده تایتانیک (titanic.zip) قابل دسترس هستند. همچنین توضیحات مربوط به ستون های این مجموعه داده را می‌توانید در این لینک مطالعه کنید.

سوال ۷- Naive Bayes and KNN

در این سوال هدف کار بر روی داده‌های مربوط به ۳۰۳ بیمار بر اساس ۱۳ ویژگی مشخص برای هر بیمار است. می‌خواهیم با استفاده از روش‌های naive bayes و knn مدل‌هایی بسازیم که با استفاده از این ویژگی‌ها بتواند پیش‌بینی کند که آیا یک بیمار با مشخصات جدید مشکل قلبی دارد یا خیر. در این بخش هم مانند سوال قبل کار اصلی به دو بخش قابل تقسیم است:

۱- تحلیل و آماده سازی داده: در این بخش هدف بررسی ویژگی‌های آماری هر ویژگی و در صورت امکان، ساختن ویژگی‌های جدید با استفاده از ویژگی‌های موجود است. همچنین لازم است مجموعه داده را به دو دسته train و test برای مرحله ساخت مدل تقسیم کنید. (می‌توانید ۲۰ درصد داده‌ها را برای test و باقی را برای train در نظر بگیرید)

۲- ساخت و آموزش مدل: در این بخش با مدل‌های naive bayes و knn را با داده‌های train آموزش دهید و نتیجه عملکرد را با معیار accuracy بسنجید. برای پیاده‌سازی مدل‌ها نیز می‌توانید از کتابخانه scikit-learn استفاده نمایید.

در نهایت از تحلیل‌ها و آماده‌سازی مرحله ۱ و ساخت و ارزیابی مدل در مرحله ۲ یک گزارش تهیه کنید.

توجه: برای توضیح بیشتر مجموعه داده می‌توانید به این لینک مراجعه کنید.