پردیس دانشکده‌های فنی دانشگاه تهران

دانشکده برق و کامپیوتر

تمرین شماره ۲

سیستم‌های هوشمند

نیمسال دوم-سال تحصیلی ۹۶-۹۷

مدرس: دکتر رشاد حسینی

زمستان ۹۶

موعد تحویل: ۱۶ فروردین ۹۷

سوال اول:

درخت تصمیم

در این سوال قصد داریم با پیاده سازی درخت تصمیم با استفاده از الگوریتم ID3 داده های letter recognition که در پیوست آمده است را طبقه بندی کنیم.

این داده ها حروف دست نویسی هستند که دارای ۱۶ ویژگی و ۲۶ کلاس هستند که در قالب پیکسل های سیاه و سفید نوشته شده اند. این ۱۶ ویژگی تشکیل شده از طول و عرض حروف نوشته شده، میانگین و واریانس پیکسل های سیاه در راستا های مختلف ، تعداد پیکسل های سیاه و ... می باشد. با توجه به این ویژگی ها داده ها به ۲۶ کلاس که همان حروف الفبای انگلیسی هستند تقسیم شده اند. تعداد نمونه های train برابر با ۱۶۰۰۰ و تعداد نمونه های test برابر با ۴۰۰۰ است.

۱) با استفاده از داده های train درخت را بر مبنای معیار ناخالصی افزوده اطلاعاتی (IG) با الگوریتم ID3 آموزش دهید و با استفاده از داده های test ، شکل درخت نهایی (سه لایه اولیه)، دقت طبقه بند و ماتریس Confusion را گزارش کنید.

۲) حال با استفاده از الگوریتم ID3 درخت را بر مبنای Gini Index با الگوریتم ID3 آموزش دهید و شکل درخت نهایی (سه لایه اولیه)، دقت طبقه بندی و ماتریس Confusion را گزارش کنید. کدام یک از معیار ها موفق تر عمل کرده است؟

۳) جای دو ویژگی با بیشترین IG را در قسمت الف عوض کنید و درخت را دوباره تشکیل دهید. دقت طبقه بندی را با حالت الف مقایسه کنید.

۴) با تعیین شرطی مناسب و یا هرس کردن، از over fit شدن مدل الف و ب جلوگیری کنید و نتایج را با حالت الف و ب مقایسه کنید. مدت زمان آموزش درخت را در تمامی حالات مقایسه کنید.

۵) حال قصد داریم برای استخراج مدل از داده ها از الگوریتم "Random Forest" استفاده کنیم. بدین منظور می توانید ویژگی ها را به گروه های K تایی تقسیم کنید و تعداد K درخت آموزش دهید. حال با استفاده از "Majority Voting" کلاس داده ی تست را تخمین بزنید. این عملیات را به ازای K های مختلف تکرار کرده و بهترین K را گزارش کنید.

سوال دوم:

KNN

در این سوال سعی داریم الگوریتم KNN را بر روی داده های iris پیاده کنیم. این داده ها دارای ۴ ویژگی طول کاسبرگ، عرض کاسبرگ، طول گلبرگ و عرض گلبرگ می باشد.

داده ها بر اساس این ویژگی ها به ۳ کلاس گل رز(کلاس ۱)، مریم (کلاس ۲) و نرگس (کلاس ۳) تقسیم شده اند.

۱) با استفاده از 6-folded cross validation داده ها را به دسته های آموزش و تست تقسیم کرده و به ازای K های ۳ و ۵ و ۷ و ۹ عملکرد طبقه بندی با معیار فاصله اقلیدسی را گزارش کنید؟ کدام K نتایج بهتری ساخته است؟

۲) با استفاده از بهترین K که در قسمت قبل یافتید ، عملکرد طبقه بندی 3-folded cross validation را بر اساس فاصله اقلیدسی و فاصله کسینوسی مقایسه کنید.

سوال سوم:

K-means Clustering

در این سوال به پیاده سازی الگوریتم خوشه بندی K-means می پردازیم. همان طور که در کلاس مطرح شد ، این الگوریتم که از دسته الگوریتم های بدون نظارت است و سعی دارد تابع هزینه ای که به صورت مجموع فاصله ی نمونه های متعلق به هر خوشه تا مرکز ان خوشه تعریف می شود را مینیموم کند. مجموعه داده ی مورد استفاده در این بخش ، همانند سوال قبل، مجموعه داده ی iris است.

۱) تعداد تکرار های الگوریتم K-means را برابر 150 در نظر گرفته و تعداد خوشه ها را {3,5,7,9} در نظر بگیرید. مناسب ترین عدد برای تعداد خوشه ها را با استفاده از معیار شباهت درونی و بیرونی خوشه ها به دست اورید. این عدد را با نتیجه ی سوال دوم مقایسه کرده و براساس ان دو الگوریتم KNN و K-means را مقایسه کنید.

۲) الگوریتم بالا را ۵ مرتبه با شرایط اولیه ی تصادفی مختلف پیاده سازی کرده و تابع هزینه و میانگین و واریانس انرا در ۵ تکرار (به ازای تعداد خوشه های {3,5,7,9}) رسم کنید.

به نکات زیر توجه فرمایید:

۱. برای آشنایی بیشتر با داده های letter recognition و iris می توانید به لینک های زیر مراجعه کنید؛

- https://archive.ics.uci.edu/ml/machine-learning-databases/letter-recognition /
- https://archive.ics.uci.edu/ml/machine-learning-databases/iris /

۲. فایل گزارش خود را با فرمت pdf ، به انضمام کدهای MATLAB خود در قالب یک فایل zip تا زمان تحویل در سایت درس با فرمت زیر بارگذاری کنید:

[Name]_[student number]_SI_Assignment[Assignment Number].zip

۳. اصلی ترین بخش هر تمرین کامپیوتری، **گزارش کار** آن است و بخش عمده نمره به آن تعلق می‌گیرد. لذا برای هر بخش، توضیحات کافی به همراه نتایج شبیه سازی خود را در گزارش کار خود بیاورید. گزارش کار لازم است فرمت یک گزارش علمی داشته باشد. از گرفتن عکس از نوشته‌های خود و الصاق آن در گزارش خود خودداری کنید. یک تمپلیت برای گزارش در سایت درس آپلود شده است. لازم است گزارش حاوی جوابهای شفاف و تحلیل کافی برای نتایج باشند.

۴. کدهای خود را تا حد امکان واضح، بی‌ابهام و ساده بنویسید و هر جایی که احساس می‌کنید فهم کد شما مشکل خواهد بود حتما از کامنت استفاده کنید.

۵. کد مربوط به هر سوال را در یک فایل جداگانه با اسم P? که علامت سوال نشان‌دهنده‌ی شماره سوال است ذخیره کنید. قسمت‌های مختلف یک سوال را با کمک ٪٪ از هم جدا کنید.

۶. می‌توانید پرسش‌های خود را از طریق ایمیل با دستیاران مربوطه مطرح کنید. khaghani.javad@gmail.com و mahdiar_nekoui@yahoo.com مطرح کنید.

۷. کپی کردن کار یکدیگر تخلف محسوب می‌شود و در صورت مشاهده کوچکترین تخلف، نمره‌ای به هیچ کدام از طرفین تعلق نمی‌گیرد.

۸. به ازای یک روز تاخیر ۵ درصد جریمه، به ازای یک هفته ۱۰ درصد و تا دوهفته تاخیر ۲۰ درصد جریمه در نظر گرفته خواهد شد. پس از آن هیچ نمره‌ای به تمرین تعلق نخواهد گرفت.

شاد باشید...