دانشکده‌ی مهندسی کامپیوتر

یادگیری ماشین

نیم‌سال نخست ۹۷ـ۹۶

مدرس: حمید بیگی

پروژه‌ی پایانی

زمان تحویل: ۱۰ بهمن‌ماه

نکات زیر را رعایت کنید:

فایل گزارش را به همراه تمامی کدها در یک فایل فشرده و با عنوان $MLProject\_studentNumber$ ایمیل کرده و در صورت داشتن سوال از طریق $s.aqamiri@gmail.com$ به آقای "سعید آقامیری" ایمیل بزنید. پروژه در گروه‌های دو نفره انجام می‌گیرد. نمره‌دهی به این صورت است که مقدار نمره‌ی هر قسمت بیان شده است و مجموع نمره از۱۰۰ است. اما به گروه‌اول ۲۰ نمره اضافه، دوم ۱۵ و سوم ۱۰ نمره‌ی اضافی تعلق می‌گیرد. رتبه‌بندی گروه‌ها هم براساس دقت به دست آمده در دو قسمت ذکر شده است. توجه کنید که حتما باید گزارشی از نحوه‌ی انجام کار و تحلیل نتایج، به همراه کد ارسال شود. لازم به ذکر است که نحوه‌ی تقسیم نمرات میان اعضای گروه هم، متعاقبا اعلام خواهد شد
فایل فشرده‌ی پاسخ‌ها را به ایمیل درس machinelearning.ce717@gmail.com ارسال کنید.

مسئله‌ی ۱. تشخیص اعداد دست‌نویس MNIST

در طول ترم با قسمت‌های مختلف یک سیستم جهت دسته‌بندی و خوشه‌بندی آشنا شدید. تقریبا این دو کار، اصلی‌ترین وظیفه‌ها در یادگیری ماشین هستند. در این پروژه می‌خواهیم یک بار تمام این قسمت‌ها در کنار هم برای شناخت اعداد دست‌نویس انجام دهیم. مجموعه دادگان مورد استفاده، همان اعداد دست‌نویس MNIST هست. چون قبلا با این پایگاه داده آشنایی دارید، توضیح بیشتری درباره‌ی دانلود و راه‌اندازی و ... نمی‌دهیم.

۱. قسمت اول، دسته‌بندی (۶۰ نمره):

هر سیستم دسته‌بندی سه قسمت متداول دارد. اول استخراج و انتخاب ویژگی‌ها از داده‌های خام اولیه (۲۵ نمره)، کاهش بعد(۱۰نمره)، سپس انتخاب یک روش‌دسته بندی مناسب (۲۵ نمره) و آموزش روی دادگان آموزش و در آخر ارزیابی دقت بر روی دادگان آزمون. توجه کنید که باید علاوه بر کد، گزارشی هم از تحلیل نتایج و علت انتخاب روش مناسبتان در هر قسمت نیز بدهید. مثلا توضیح دهید که چرا درخت تصمیم بهتر از ماشین بردار پشتیبان است (یا برعکس) و چرا ویژگی خاصی را انتخاب کردید و ... معیار ارزیابی هم بر روی دادگان آزمون، معیار معروف دقت هست. یعنی تعداد پاسخ‌های صحیح تقسیم بر کل تعداد پرسش‌های آزمون

۲. قسمت دوم، خوشه‌بندی (۴۰ نمره):
هر سیستم خوشه بندی هم مانند دسته‌بندی، سه قسمت عمده دارد. انتخاب و استخراج ویژگی(۱۵ نمره)، کاهش بعد(۵ نمره) و نهایتا انتخاب روش مناسب برای خوشه‌بندی(۱۵ نمره) و علاوه بر این ها به دست آوردن معیار(۵ نمره) نیز مورد ارزیابی قرار می گیرد. باز هم در انتخاب روش‌های مختلف برای این سه قسمت آزادی کامل دارید. فقط صرفا باید گزارشی تهیه کنید و دلایل انتخاب روش‌تان را توضیح دهید.

توجه کنید که در این قسمت نیازی به دادگان تست نیست و آموزش و ارزیابی هر دو بر روی دادگان آموزش انجام می‌شود. معیار ارزیابی مورد استفاده در این قسمت، استفاده از رابطه‌ی خالص‌سازی می‌باشد که تعریف آن در زیر آمده است:

Purity = میانگین وزن‌دار (بر روی دسته‌های مختلف) نسبت داده‌های درست به تعداد کل داده‌های یک دسته

۳. دقت کنید که در بخش خوشه‌بندی، اجازه‌ی استفاده از برچسب داده‌ها را در زمان آموزش ندارید. فقط برای حساب کردن معیار خالص سازی باید از برچسب‌ها استفاده کنید.

برای مقایسه‌ی امتیاز هر تیم و مشخص شدن تیم‌های برتر به این صورت عمل می‌شود که دقت برای قسمت اول ۶۰ درصد و معیار ارزیابی خوشه‌بندی هم ۴۰ درصد تاثیر دارد. تیمی که در مجموع این دو معیار عملکرد بهتری داشته باشد، به عنوان تیم اول انتخاب می‌شود.