به نام خدا

یادگیری ماشین | نیمسال اول ۹۸-۹۷

پروژه نهایی

موعد تحویل: ۹ بهمن | نمره: ۱۰۰

معرفی پروژه

در این پروژه قصد داریم بر اساس داده‌ی سنسورهای تلفن همراه هوشمند، فعالیت فرد را تشخیص دهیم.

مجموعه دادگان

این مجموعه داده شامل بیش از ده هزار نمونه از شش فعالیت انسان می‌باشد. این فعالیت‌ها عبارت‌اند از سه فعالیت حرکتی شامل راه رفتن، بالا رفتن از پله و پایین آمدن از پله و سه فعالیت غیر حرکتی شامل ایستادن، نشستن و دراز کشیدن. همچنین به حالت گذار میان هر یک از فعالیت‌های غیرحرکتی نیز برچسب جداگانه اختصاص یافته است. در طی انجام هر فعالیت، یک گوشی هوشمند به کمر نمونه‌ها متصل شده است و داده‌ی مربوط به اندازه‌ی شتاب و سرعت زاویه‌ای گوشی در سه بعد، ذخیره می‌شود. بعد از انجام مراحل پیش‌پردازش، ۵۶۱ ویژگی از داده‌ی سنسورهای شتاب‌سنج و سرعت‌سنج زاویه‌ای در حوزه‌ی زمان و فرکانس استخراج شده است. ویژگی‌های استخراج‌شده برای ۷۰ ٪ داده‌ها به‌عنوان داده‌ی آموزش در پوشه‌ی Train قرار دارد. ۳۰ ٪ دیگر داده‌ها به‌عنوان داده‌ی تست در پوشه‌ی Test قرار خواهد گرفت که برای ارزیابی نتایج شما استفاده می‌شود و طی فرایند آموزش در اختیار شما نخواهد بود.

هدف پروژه

در این پروژه قصد داریم با استفاده از روش‌های یادگیری ماشین که در طول‌ترم بیان شده است فعالیت‌های مختلف افراد را شناسایی و تحلیل کنیم.

دسته‌بندی

این مجموعه داده شامل ۱۲ فعالیت گوناگون انسانی می‌باشد.

۱. با استفاده از روش‌های Logistic Regression، Random Forest، SVM، Adaboost و شبکه‌ی عصبی Fully Connected دو لایه، دقت دسته‌بندی را گزارش کنید. برای تنظیم هایپر پارامترها از داده‌های validation استفاده کنید.

۲. حال برای روش‌های Linear SVM و Logistic Regression و شبکه‌ی عصبی، پنالتی L1 و L2 را به Loss Function اضافه کنید. ضریب منظم‌سازی برای دو پنالتی بالا را از میان مقادیر $\{۱۰^{-۴}, ۲*۱۰^{-۴}, ۵*۱۰^{-۴}, ۱۰^{-۳}, ۲*۱۰^{-۳}, ۵*۱۰^{-۳}, ۱۰^{-۲}, ۱۰^{-۱}, ۱, ۱۰, ۱۰۰\}$ انتخاب کنید. خطا برای دو روش پنالتی گفته‌شده را بررسی نمایید. تعداد پارامترهای ناصفر (بزرگتر از $۱۰^{-۶}$) برای هر ضریب منظم‌سازی برای دو پنالتی را محاسبه کرده، نمایش دهید و آن را تحلیل کنید.

۳. تعداد داده‌های ورودی یکی از عوامل تاثیرگذار بر دقت مدل است. k درصد از داده‌های آموزش را به‌صورت تصادفی انتخاب کرده و برای kهای { ۵، ۱۰، ۲۰، ۵۰، ۱۰۰ } اثر تعداد داده‌های ورودی بر مدل‌های قسمت اول را بررسی کنید.

۴. برای هر یک از روش‌های قسمت اول به جز شبکه‌ی عصبی، تعداد l ویژگی از مجموع ۵۶۱ ویژگی داده‌شده را انتخاب کرده و دقت خروجی را گزارش کنید. (l از میان مقادیر { ۵، ۱۰، ۵۰، ۱۰۰، ۵۶۱ } انتخاب می‌شود.) یافتن روش مناسب برای انتخاب ویژگی‌ها بر عهده‌ی شما می‌باشد. توضیح دهید که چگونه می‌توان روشی ارائه داد که با کمک نتایج قسمت دوم، ویژگی‌های مناسبی انتخاب کرد. نتیجه‌ی روش پیشنهادی خود را گزارش دهید و با دیگر روش‌ها مقایسه کنید.

برای آشنایی بیشتر با برخی روش‌ها می‌توانید اینجا را نگاه کنید. توجه کنید که ممکن است یک روش انتخاب ویژگی، برای تمامی دسته‌بندها بهترین انتخاب نباشد. به‌عنوان‌مثال اثر روش انتخاب ویژگی Tree-based feature selection بر دسته‌بند Random Forest را گزارش دهید. پس از یافتن بهترین انتخاب برای هر دسته‌بند، علت این انتخاب را تحلیل کنید.

۵. در این قسمت هدف یافتن بهترین نتیجه است. با انتخاب تنها ۱۰ ویژگی، بهترین دقت خروجی را به دست آورید. انتخاب روش Feature Selection یا Feature Extraction، روش دسته‌بندی و پارامترهای مدل بر عهده‌ی شما می‌باشد. استفاده از هر ایده و روش برای بهبود نتیجه در این قسمت آزاد می‌باشد. به بالاترین درصدها نمره‌ی اضافه تعلق می‌گیرد.

خوشه‌بندی

۱. در این قسمت قصد داریم بدون استفاده از برچسب‌های نمونه‌ها، آن‌ها را به ۱۲ گروه مختلف تقسیم کنیم. برای روش‌های K-means و GMM[1]، دقت خروجی را گزارش دهید. هر روش را چندین بار اجرا کرده و بهترین نتیجه با توجه به تابع هزینه را اعلام کنید. پس از انتخاب مدل با بهترین پارامترها، معیار Rand Index را محاسبه نمایید.

۲. در این قسمت اثر کاهش ابعاد بر دقت خوشه‌بندی را بررسی می‌کنیم. با استفاده از PCA ابعاد داده‌ی ورودی را به ۲۰ کاهش داده و مجدد الگوریتم خوشه‌بندی را اجرا کنید. نتیجه‌ی این قسمت را با قسمت قبل مقایسه کرده و تحلیل کنید. داده‌ها را در فضای دو بعدی، با رنگ‌های متفاوت برای هر خوشه نمایش داده و با قسمت قبل مقایسه کنید.

[1] Gausian Mixture Model

تنظیمات و قالب‌بندی

نحوه‌ی پیاده‌سازی و اجرای کد باید به قالب زیر باشد.

۱. انتخاب توابع کرنل، یافتن تنظیمات بهینه، انتخاب روش انتخاب ویژگی و تعیین پارامترهای مناسب مدل بر عهده‌ی شما می‌باشد. نتایج نهایی افراد مختلف با یکدیگر مقایسه خواهد شد.

۲. کد باید به نحوی باشد که داده‌های آموزش در فایل به آدرس '..\data\Train\' و داده‌های تست در فایل به آدرس '..\data\Test\' قرار گرفته باشند. با قرار دادن داده‌های آموزش و تست در آدرس‌های گفته شد، کد باید بدون مشکل اجرا گردد.

۳. توابع موردنیاز باید به‌صورت ماژولار پیاده‌سازی شود.

پیاده‌سازی

برای پیاده‌سازی بخش‌های مختلف می‌توانید از هر کتابخانه‌ای استفاده کنید. قسمتی از نمره‌ی شما به انتخاب محیط پیاده‌سازی مربوط می‌شود؛ بنابراین پیش از این‌که اجرای پروژه را شروع کنید باید یک تحقیق نسبتاً جامع از کتابخانه‌های معروف داشته باشید، خوبی‌ها و بدی‌های کتابخانه‌هایی که دیده‌اید را در گزارشتان ذکر کرده و یکی را برای اجرای این پروژه انتخاب کنید. استفاده از کتابخانه‌های موجود در پایتون و متلب در پروژه بلامانع است.
موارد تحویلی شما شامل کدهایی که توسط شما پیاده‌سازی شده‌اند و مستندات پروژه (معرفی کوتاه کتابخانه‌ها و دلیل انتخاب شما، پاسخ به سؤالات مطرح شده، توضیحات پیاده‌سازی و خروجی‌ها) می‌باشد.