

تمرین سری دوم (مهلت تحویل تا قبل از شروع امتحانات)

برای این تمرین مجموعه داده letter-recognition از لینک زیر را در نظر می‌گیریم.

<https://archive.ics.uci.edu/ml/datasets/letter+recognition>

هر رکورد در این مجموعه داده مربوط به تصویر یک حرف انگلیسی (A-Z) است. هر رکورد شامل ۱۶ ویژگی عددی است که از تصویر حروف استخراج شده است (برای مثال یک ویژگی می‌تواند متوسط تعداد پیکسل‌های تیره در تصویر مربوطه باشد). جزئیات مربوط به این شانزده ویژگی در لینک بالا موجود است. علاوه بر ۱۶ ویژگی مورد بحث، هر رکورد شامل برچسپ تصویر مورد نظر است که یکی از حروف A تا Z می‌باشد (اولین آیتم هر رکورد، در واقع برچسپ مربوطه است). در این تمرین می‌خواهیم با استفاده از متدهای دسته‌بندی که در کلاس یادگرفته‌ایم، نتایج مربوط به الگوریتم‌های دسته‌بندی مختلف را روی این مجموعه داده آزمایش کرده و آنها را با هم مقایسه کنیم. برای سادگی کار، ما فقط داده مربوط به حروف X و K را در نظر می‌گیریم. پس می‌خواهیم نتایج مربوط به متدهای دسته‌بندی مختلف را برای تشخیص حرف X از K را بسنجیم. داده مربوط به حروف X و K در فایل XK-recognition.data جدا شده و در صفحه وب درس قرار داده شده است. داده‌ها را بصورت تصادفی به دو بخش train و test تقسیم کرده‌ایم. داده‌های یادگیری در فایل train.data و داده‌های تست در فایل XK-test.data قرار داده شده است. کد پایتون استخراج داده‌ها recognize.py در صفحه وب درس قرار داده شده. دقت کنید، شما فعلاً نیازی به اجرای این کد ندارید. همچنین برای هماهنگی بیشتر با تمرین قبلی، در داده‌های فایل train و test برچسپ‌ها به انتهای خط منتقل شده و بصورت 1 و -1 در آمده است. حال شما بایستی الگوریتم‌های یادگیری که در زیر ذکر شده است را روی داده‌های فایل train آزموده و نتیجه را با داده‌های فایل test بسنجید.

- دسته‌بندی پرسپترون با الگوریتم پکت pocket algorithm
 - دسته‌بندی با روش SVM با کرنل‌های خطی، چندجمله‌ای و RBF. مقادیر مختلف را برای پارامتر C در نظر بگیرید. برای مثال 0.01, 0.05, 0.1, 0.5, 1, 2
 - دسته‌بندی با روش KNN (نزدیکترین همسایه). در این روش پارامتر K را در مقادیر 1, 3, 5 تست کنید. فاصله نقاط را با معیار فاصله اقلیدسی در نظر بگیرید.
- میزان خطای دسته‌بندی را برای روش‌های مختلف و پارامترهای متفاوت در جدولی ثبت کنید. بهترین روش کدام است؟ جای فایل train و test را عوض کنید. نتایج آیا بهتر خواهد شد؟