تمرین ششم درس یادگیری ماشین
دکتر باباعلی

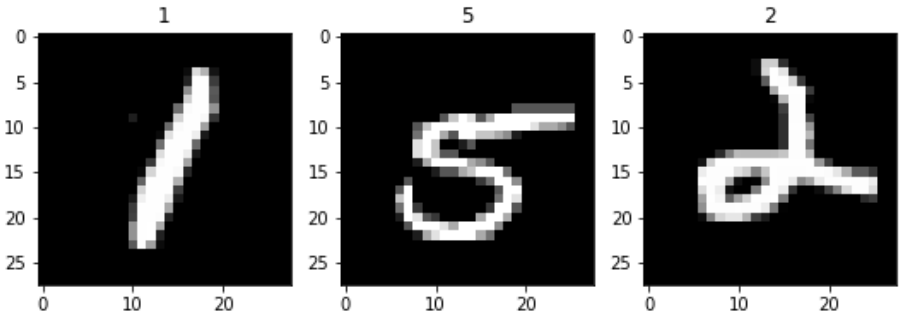
سیدعلیرضا مولوی

فهرست مطالب

۱ داده ها

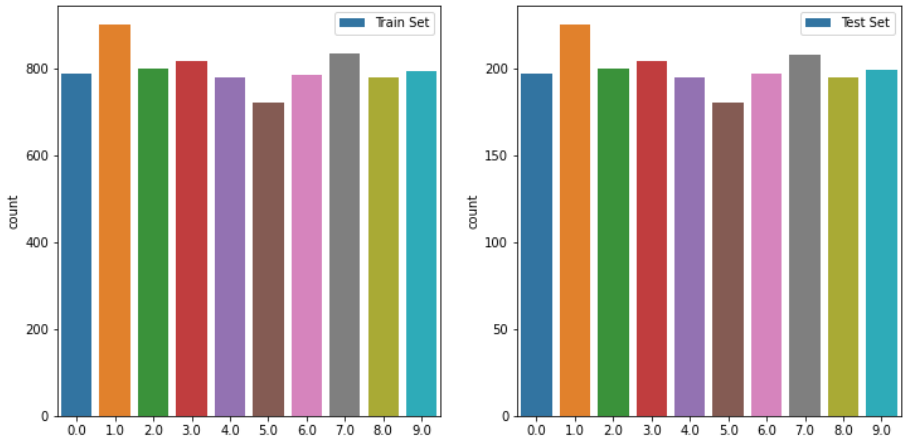
داده های mnist را از سایت openml تهیه کردم، در واقع فرآیند دانلود را به عهده کتابخانه scikit-learn گذاشتم.
تعداد داده ها ۷۰۰۰۰ نمونه است که ۸۰۰۰ نمونه را به عنوان داده های آموزشی و ۲۰۰۰ نمونه را به عنوان داده های تست انتخاب کرده ایم.
داده ها تصاویر سیاه و سفید از اعداد انگلیسی با رزولوشن ۲۸ × ۲۸ اند. در شکل ۱ نمونه از داده ها نمایش داده شده اند.

شکل ۱: نمونه از داده های آموزشی



اعداد از شماره ۰ تا ۱۰ است، بنابراین مسئله دسته بندی ۱۰ کلاسه داریم. در شکل ۲ تعداد کلاس های موجود در داده های آموزشی و تست نمایش داده شده است، با توجه به شکل ۲ داده ها متعادل اند.

شکل ۲: تعداد کلاس ها در داده های آموزشی و تست



۱.۱ پیش پردازش

ویژگی داده ها ماتریسی از بعد ۲۸ × ۲۸ اند، بنابراین ویژگی داده ها رو به بردار یک بعدی ۷۸۴ تغییر شکل می دهیم.
به دلیل اینکه از مدل K-Nearest Neighbor استفاده می کنیم، این مدل به **شدت** به مقیاس اندازه های ویژگی ها وابسته است، بنابراین پیش از استفاده از مدل داده ها نرمالیز می کنیم.

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (۱)$$

که در داده mnist به صورت زیر در می آید:

$$x_{min} = ۰, \quad x_{max} = ۲۵۵ \implies x' = \frac{x}{۲۵۵} \tag{۲}$$

۲ مدل K-Nearest Neighbor

۱.۲ مدل بدون وزن

ما k همسایه ورودی را انتخاب می کنیم و کلاس ورودی را برابر با مد[۱] کلاس k همسایه (کلاس متناظر با اکثریت k همسایه) قرار می دهیم.

برای یافتن k بهینه از بین ۳۲،۱۶،۵،۴،۲،۱ از 5-Fold Cross validation استفاده کرده ایم.

نتیجه در جدول ۱آ و شکل ۳آ نمایش داده شده است؛ بنابراین مقدار بهینه $k = ۵$ است و دقت داده آموزشی و تست برابر است با:

$$Train\ Acc = ۰/۹۶۱۴ \qquad Test\ Acc = ۰/۹۴۱۵$$

۲.۲ مدل وزن دار

در این حالت ما به هر یک از k همسایگی وزنی اختصاص می دهیم که این وزن رابطه معکوس با فاصله دارد، یعنی هر چه همسایگی دور تر باشد وزن کمتری و هر چه نزدیک تر باشد وزن بیشتری دارد. در نهایت کلاسی را انتخاب می کنیم که مجموع وزن هایش بیشینه باشد.

$$w_i = \frac{۱}{d(x^{(i)}) + \epsilon} \tag{۳}$$

که در معادله ۳ $d(x^{(i)})$ فاصله نمونه i ام با ورودی است و ϵ مقدار بسیار کوچکی برای جلوگیری از ۰ شدن مخرج است.

رابطه زیر برای مدل وزن دار است (رابطه در جزوه است).

$$\hat{y}_{x-new} = argmax_j W(j) \tag{۴}$$

$$W(j) = \sum_{i \in N_k} w_i I(y_i = j) \tag{۵}$$

توجه: حالت بدون وزن در واقع حالت خاصی از حالت وزن دار است، به طوری که تمام وزن ها مستقل از فاصله، مقداری مساوی (مثلا ۱) داشته باشند.

برای یافتن k بهینه از بین ۳۲،۱۶،۵،۴،۲،۱ از 5-Fold Cross validation استفاده کرده ایم.

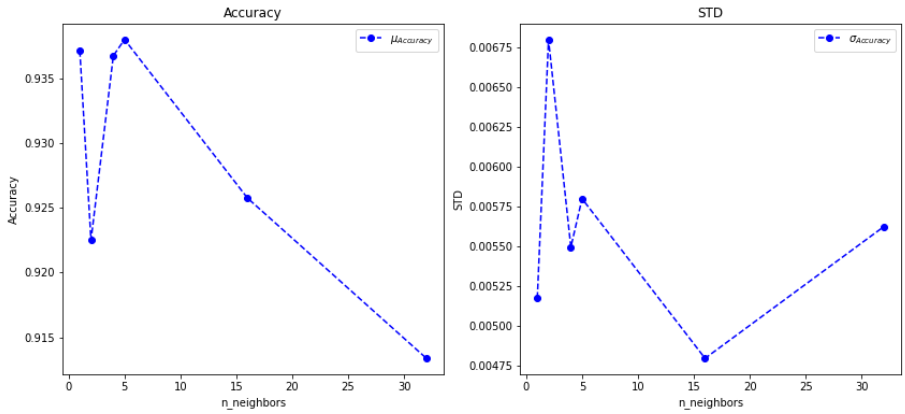
نتیجه در جدول ۱ب و شکل ۳ب نمایش داده شده است؛ بنابراین مقدار بهینه $k = ۴$ است و دقت داده آموزشی و تست برابر است با:

$$Train\ Acc = ۱/۰۰۰۰ \qquad Test\ Acc = ۰/۹۴۴۰$$

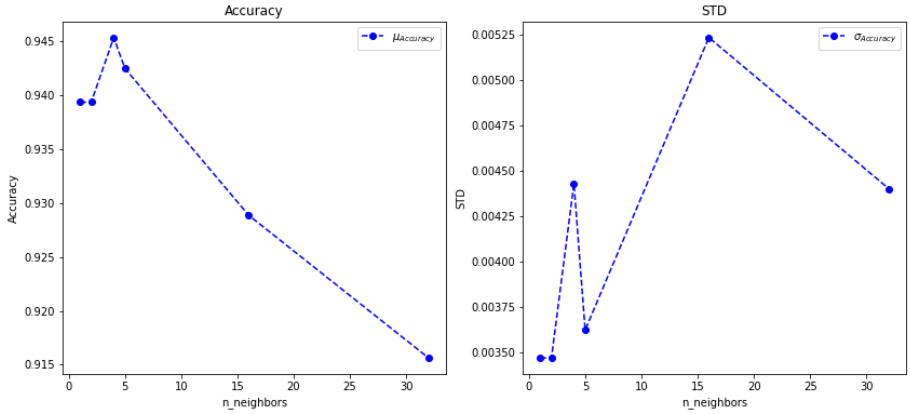
[۱] Mode

شکل ۳: دقت و انحراف از معیار مدل آموزش داده شده با cv=۵. شکل سمت چپ دقت - شکل سمت راست انحراف از معیار

(آ) مدل بدون وزن



(ب) مدل وزن دار



جدول ۱: دقت و انحراف از معیار مدل آموزش داده شده با $cv = ۵$.

(آ) مدل بدون وزن

تعداد همسایه K	دقت آموزش	انحراف معیار آموزش
۱	۰/۹۳۷۱	۰/۰۰۵۲
۲	۰/۹۲۲۵	۰/۰۰۶۸
۴	۰/۹۳۶۸	۰/۰۰۵۵
۵	۰/۹۳۸۰	۰/۰۰۵۸
۱۶	۰/۹۲۵۸	۰/۰۰۵۸
۳۲	۰/۹۱۳۴	۰/۰۰۵۶

(ب) مدل وزن دار

تعداد همسایه K	دقت آموزش	انحراف معیار آموزش
۱	۰/۹۳۹۴	۰/۰۰۳۵
۲	۰/۹۳۹۴	۰/۰۰۳۵
۴	۰/۹۴۵۴	۰/۰۰۴۴
۵	۰/۹۴۲۶	۰/۰۰۳۶
۱۶	۰/۹۲۸۹	۰/۰۰۵۲
۳۲	۰/۹۱۵۶	۰/۰۰۴۴