به نام خدا

تمرین هشتم مقدمه‌ای بر یادگیری ماشین

حسین ابراهیمی ـ ۹۵۱۰۵۳۰۲

مدرس : دکتر جمال‌الدین گلستانی

سوال C۵

روش اول.

مقدار دقت در روش SVM با کرنل خطی برابر با 77.34% شد و ماتریس confusion آن برابر است با:

$$\begin{pmatrix} 336 & 8 & 16 & 35 & 6 & 1 & 63 & 1 & 8 & 0 \\ 6 & 454 & 5 & 18 & 0 & 0 & 5 & 3 & 1 & 0 \\ 10 & 0 & 324 & 9 & 73 & 2 & 85 & 1 & 9 & 0 \\ 26 & 9 & 9 & 391 & 19 & 0 & 26 & 1 & 1 & 0 \\ 6 & 2 & 64 & 33 & 328 & 0 & 79 & 1 & 6 & 0 \\ 3 & 0 & 7 & 2 & 1 & 355 & 4 & 42 & 11 & 12 \\ 75 & 3 & 53 & 21 & 66 & 0 & 282 & 0 & 11 & 0 \\ 0 & 0 & 0 & 0 & 0 & 25 & 2 & 507 & 1 & 20 \\ 11 & 1 & 15 & 5 & 9 & 7 & 17 & 11 & 433 & 1 \\ 0 & 0 & 1 & 1 & 0 & 14 & 1 & 29 & 4 & 457 \end{pmatrix}$$

روش دوم.

ابتدا مقدار γ که عکس مقدار σای است که در صورت سوال ذکر شده است را برابر با مقادیر $10^{-10}, 10^{-9}, \cdots, 10^{5}, 10^{6}$ قرار دارم. بهترین دقت را اعداد 10^{-7} و 10^{-8} داشتند. سپس مقدار γ در بازه‌ی اعداد بین این دو تغییر دادم و بهترین نتیجه روی داده validation برای $\gamma = 38 \times 10^{-8}$ بدست آمد که برابر با 86.42% درصد دقت بود و ماتریس confusion آن برابر شد با:

$$\begin{pmatrix} 401 & 1 & 4 & 20 & 1 & 1 & 52 & 0 & 7 & 0 \\ 5 & 494 & 11 & 15 & 1 & 0 & 2 & 0 & 0 & 0 \\ 6 & 0 & 405 & 10 & 53 & 0 & 42 & 0 & 5 & 0 \\ 22 & 0 & 1 & 469 & 6 & 0 & 7 & 0 & 1 & 0 \\ 0 & 1 & 35 & 19 & 392 & 0 & 38 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 411 & 0 & 21 & 0 & 10 \\ 78 & 0 & 50 & 14 & 42 & 0 & 306 & 0 & 7 & 0 \\ 0 & 0 & 0 & 0 & 0 & 13 & 0 & 462 & 1 & 28 \\ 1 & 0 & 5 & 3 & 2 & 3 & 6 & 1 & 521 & 1 \\ 0 & 0 & 0 & 0 & 0 & 5 & 0 & 20 & 1 & 460 \end{pmatrix}$$

روش سوم.

مقادیر k را برای مدل KNN از ۱ تا ۲۰ تغییر دادم و بهترین دقت بر روی داده‌ی validation برای $k = 4$ اتفاق افتاد که برابر بود با 81.08%.

$$\begin{pmatrix} 411 & 1 & 9 & 16 & 4 & 0 & 40 & 0 & 5 & 0 \\ 4 & 475 & 7 & 12 & 1 & 0 & 2 & 0 & 0 & 0 \\ 10 & 0 & 375 & 5 & 64 & 0 & 43 & 1 & 2 & 0 \\ 43 & 11 & 8 & 429 & 14 & 0 & 15 & 0 & 1 & 0 \\ 2 & 0 & 76 & 27 & 333 & 0 & 45 & 0 & 0 & 0 \\ 3 & 0 & 1 & 1 & 0 & 382 & 2 & 55 & 3 & 44 \\ 123 & 1 & 86 & 8 & 38 & 0 & 253 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 4 & 0 & 467 & 2 & 26 \\ 3 & 0 & 22 & 6 & 4 & 1 & 8 & 8 & 463 & 2 \\ 0 & 0 & 1 & 0 & 0 & 2 & 0 & 23 & 0 & 466 \end{pmatrix}$$

روش چهارم.

با یادگیری بر روی داده آموزشی، مقدار خطای درخت یادگیری بروی داده validation برابر با 73.92% شد.

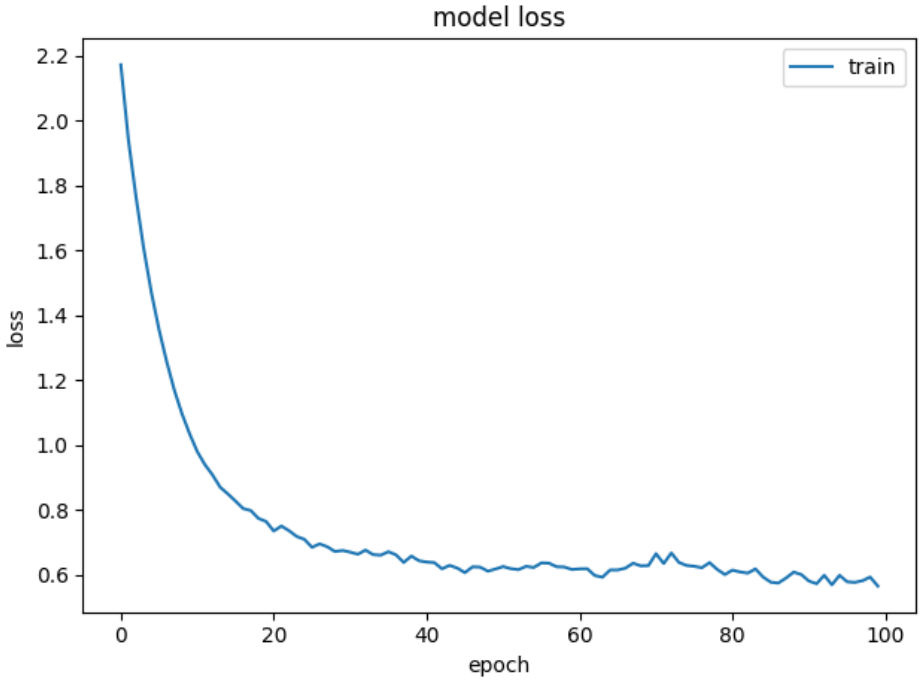
$$\begin{pmatrix} 319 & 7 & 17 & 32 & 9 & 3 & 84 & 2 & 11 & 0 \\ 4 & 462 & 3 & 26 & 6 & 0 & 3 & 0 & 3 & 0 \\ 8 & 3 & 314 & 13 & 65 & 1 & 78 & 1 & 12 & 0 \\ 39 & 28 & 9 & 347 & 22 & 3 & 36 & 0 & 7 & 1 \\ 9 & 9 & 88 & 24 & 278 & 2 & 55 & 1 & 14 & 0 \\ 0 & 2 & 0 & 1 & 0 & 385 & 2 & 46 & 19 & 30 \\ 78 & 4 & 58 & 22 & 66 & 1 & 256 & 2 & 13 & 1 \\ 0 & 0 & 0 & 0 & 0 & 42 & 0 & 439 & 5 & 44 \\ 14 & 0 & 14 & 8 & 3 & 15 & 10 & 9 & 444 & 1 \\ 1 & 0 & 2 & 0 & 3 & 12 & 3 & 31 & 4 & 452 \end{pmatrix}$$

روش پنجم

با استفاده از کتابخانه Keras ابتدا لایه ورودی با اندازه تعداد هاfeature می‌سازیم سپس لایه‌های پنهان را با اندازه ۱۰۰ با تابع Dense به آن اضافه می‌کنیم و در انتها لایه آخر را با ۱۰ نورون و با تابع فعال‌سازی softmax به آن اضافه می‌کنیم. در انتها بزرگترین مقدار این ۱۰ نورون که بیان‌کننده آن است که از بقیه حالات احتمال بیشتری دارد تا آن label (شماره نورون) را داشته باشد خروجی می‌دهیم.

با مقایسه مقدار دقت مدل برای تابع‌های فعال‌سازی کتابخانه ،Keras بهترین نتیجه بروی داده validation با ۱۰۰ epoch را تابع sigmoid با 79.08% دقت داشت. ماتریس confusion و نمودار تابع هزینه برحسب زمان یادگیری به شکل زیر است:

$$\begin{pmatrix} 382 & 3 & 11 & 48 & 6 & 0 & 44 & 0 & 5 & 0 \\ 0 & 475 & 9 & 13 & 2 & 0 & 0 & 0 & 1 & 0 \\ 10 & 0 & 361 & 7 & 107 & 0 & 37 & 0 & 1 & 0 \\ 26 & 5 & 2 & 448 & 14 & 0 & 13 & 0 & 2 & 0 \\ 4 & 2 & 38 & 31 & 375 & 0 & 24 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 382 & 0 & 34 & 4 & 56 \\ 90 & 2 & 74 & 36 & 123 & 0 & 183 & 0 & 7 & 0 \\ 0 & 0 & 0 & 0 & 0 & 15 & 0 & 442 & 2 & 61 \\ 1 & 0 & 7 & 6 & 5 & 5 & 15 & 6 & 440 & 1 \\ 0 & 0 & 0 & 0 & 0 & 2 & 0 & 25 & 1 & 466 \end{pmatrix}$$



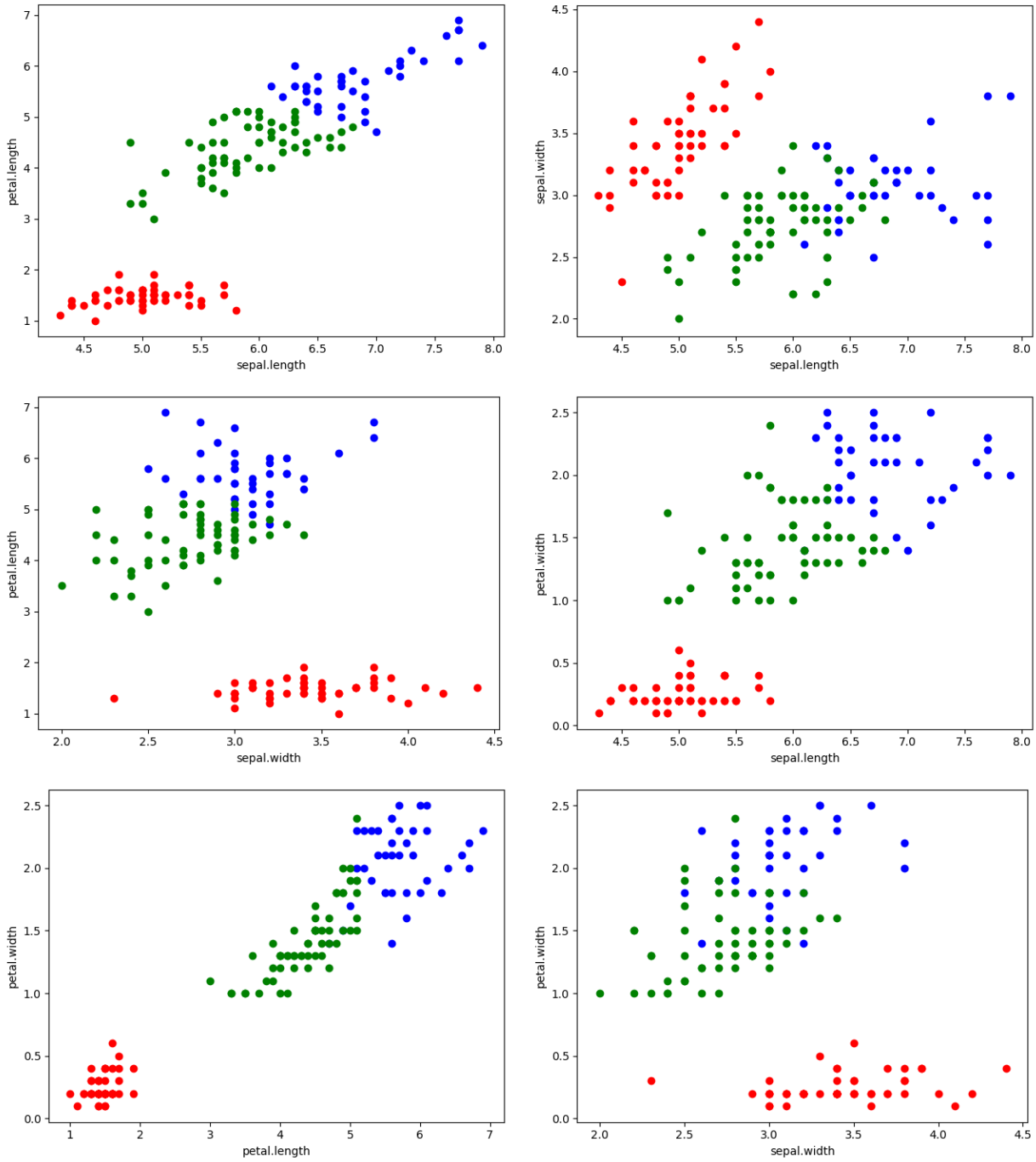
شکل ۱: تابع هزینه بر حسب epoch

مقایسه

همانطور که در قسمت‌های بالا نشان دادیم بهترین دقت از بین روش‌های گفته شده بروی داده fashion-Mnist، روش SVM با استفاده از کرنل گوسی داشت که مقدار دقت آن بروی داده‌ی validation که به صورت تصادفی از بین داده‌ها انتخاب شده بود برابر با 86.42% بود.

سوال C۶

ب.



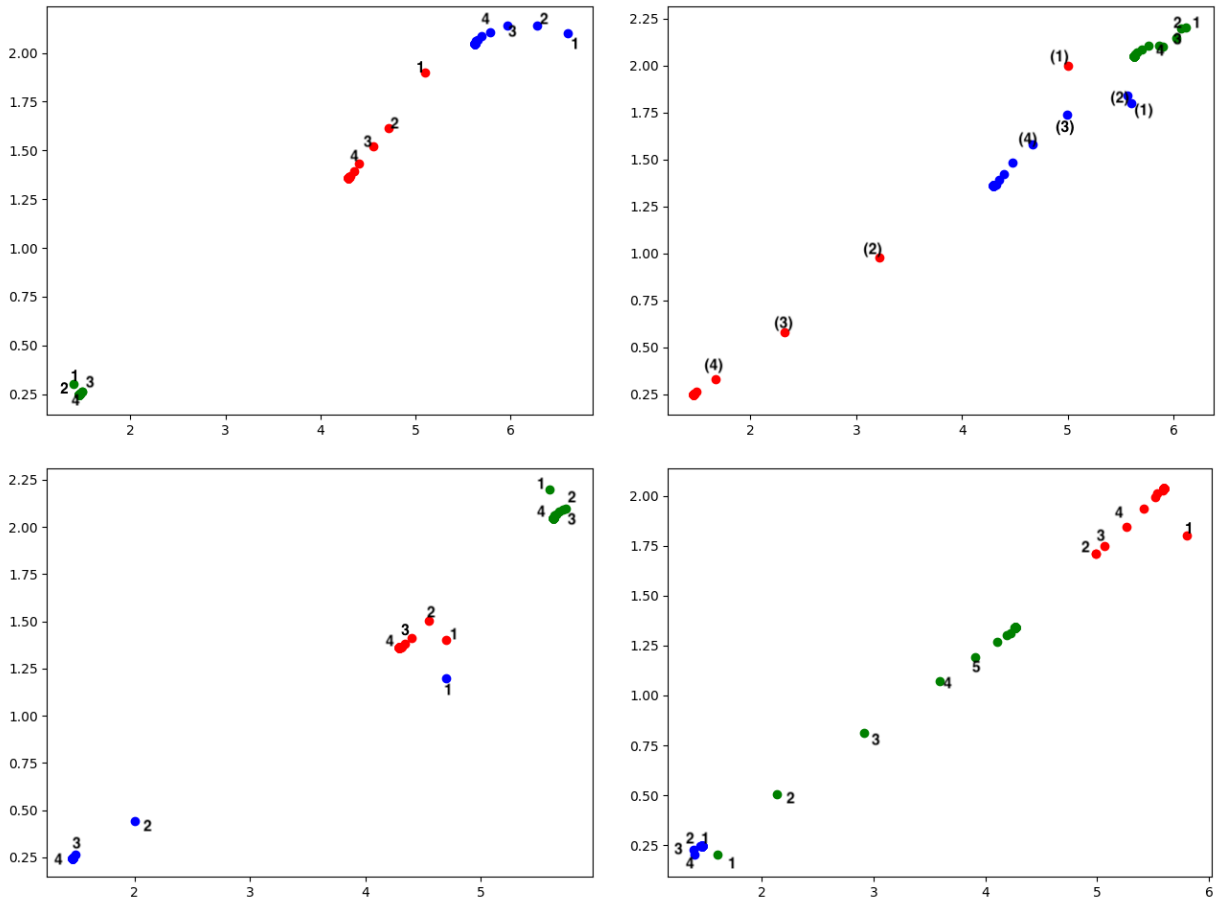
شکل ۲: نمودار داده‌ها بر حسب ویژگی‌های مختلف

ج.

همانطور که در نمودارهای بالا می‌توان دید، برای شکل‌هایی که یک محور آن sepal.width است، جدایی و کلاسترشدن داده‌ها بر اساس ویژگی دیگر آن نمودارها انجام شده است و این ویژگی نقشی ندارد . در واقع برای گل‌هایی با category متفاوت با توجه به نمودارهای بالا این ویژگی مقادیر بسیار شبیه به هم دارد و نمی‌تواند اطلاعات اضافه‌ای برای تفاوت گل‌ها بدهد پس می‌توان با

دلایل ذکر شده این ویژگی را از بین هاfeature حذف کرد بدون آنکه خطا در clustring تغییر چندانی بکند.

د.



شکل ۳: نمودارهای همگرایی centroid ها با شروع‌های رندوم متفاوت

نقاط با رنگ‌های متفاوت در هر یک از نمودارها نشان دهنده centroid یکی از cluster ها می‌باشد و اعداد نمایش داده شده شماره گام مربوط به آن نقطه برای آن centroid است. همانطور که از شکل‌ها پیداست به ازای شروع‌های رندوم و متفاوت مقدار نهایی‌ای که centroid ها به آن همگرا می‌شوند یکسان است.