گزارش تمرین چهارم یادگیری ماشین

Fall 2020 - Machine Learning

چکیده

در این گزارش به بررسی سوالات تمرین ۴ خواهیم پرداخت.

۱

سوال ۱

در این سوال برای نمایش میانگین y از علامت $\hat{\mu}$ استفاده می‌نماییم. می‌دانیم به صورت کلی می‌توان نوشت:

$$\begin{aligned} &\hat{\mu} = \mu^T w \rightarrow (\hat{\mu_1} - \hat{\mu_2})^2 = w^T(\hat{\mu_1} - \hat{\mu_2})(\hat{\mu_1} - \hat{\mu_2})^T w = w^T S_B w \\ &\sigma^T = w^T \Sigma w \rightarrow \sigma_1^2 + \sigma_2^2 = w^T(\Sigma_1 + \Sigma_2) w = w^T S_W w \end{aligned} \tag{۱.۱}$$

حال برای بهینه سازی نسبت به w مشتق گرفته و مساوی با صفر قرار می‌دهیم:

$$\begin{aligned} &J = \frac{f}{g} \rightarrow j' = \frac{f'g - fg'}{g^2} = 0 \equiv f'g - fg' = 0 \\ &\rightarrow S_B w (w^T s_W w) - S_W w w^T S_B w = 0 \rightarrow S_B w = \lambda S_W w; \qquad \lambda = \frac{w^T S_B w}{w^T S_w w} \end{aligned} \tag{۲.۱}$$

با توجه به آنکه $S_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$ است، عملا می‌توان گفت این ماتریس یک ماتریس با رنک یک است و فضای تصویر آن یک فضای یک بعدی است که توسط بردار $(\mu_1 - \mu_2)$ ساخته می‌شود. بنابران می‌توان گفت که $S_B w$ در راستای بردار $(\hat{\mu_1} - \hat{\mu_2})$ واقع می‌شود:

$$\begin{aligned} &S_B w = k(\mu_1 - \mu_2) \rightarrow s_W w = \frac{k}{\lambda}(\mu_1 - \mu_2) \\ &\rightarrow w = \frac{k}{\lambda} s_W^{-1}(\mu_1 - \mu_2) = \frac{k}{\lambda}(\Sigma_1 + \Sigma_2)^{-1}(\mu_1 - \mu_2) \end{aligned} \tag{۳.۱}$$

از آنجا که ضریب w در تابع هزینه J بی‌اثر است($J(w) = J(\alpha w)$) می‌توان معادله ۳.۱ را به صورت زیر بازنویسی نمود:

$$\boxed{w^* = (\Sigma_1 + \Sigma_2)^{-1}(\mu_1 - \mu_2)}$$

۲

سوال ۲

۱.۲ الف

داریم:

$$N_c\mu_c = N\mu - N_1\mu_1 - N_2\mu_2 .... - N_{c-1}\mu_{c-1} = N_1(\mu - \mu_1) + N_2(\mu - \mu_2) + ...+$$
$$N_{c-1}(\mu - \mu_{c-1}) + N_c\mu$$
$$\to N_c(\mu_c - \mu) = N_1(\mu_1 - \mu) + N_2(\mu_2 - \mu) + ... + N_{c-1}(\mu_{c-1} - \mu) \tag{۱.۲}$$
$$if\ \ \zeta_i = N_i(\mu_i - \mu);\ \forall i \in \{1, ..., c\} \to \boxed{\zeta_c = \zeta_1 + ... + \zeta_{c-1}}$$

حال بر اساس ۱.۲ ماتریس S_B را مجددا بازنویسی می‌نماییم:

$$S_B = \zeta_1\zeta_1^T + \zeta_2\zeta_2^T + ... + \zeta_c\zeta_c^T \tag{۲.۲}$$

حال برای بررسی رنک ماتریس S_B کافی است بعد فضای تصویر را بررسی نماییم:

$$\forall y \in IM(S_B),\ \exists x : y = S_Bx \xrightarrow{2.2} y =< x, \zeta_1 > \zeta_1+ < x, \zeta_2 > \zeta_2 + ...+ < x, \zeta_c > \zeta_c \tag{۳.۲}$$
$$\to y = \alpha_1\zeta_1 + \alpha_2\zeta_2 + ... + \alpha_c\zeta_c \xrightarrow{2.1} = \gamma_1\zeta_1 + \gamma_2\zeta_2 + ... + \gamma_{c-1}\zeta_{c-1}$$
$$\to \boxed{\forall y \in IM(S_B)\ \exists\gamma_i,\ i \in \{1, .., c-1\} : y = \gamma_1\zeta_1 + ... + \gamma_{c-1}\zeta_{c-1}}$$

بنابر ۳.۲ به وضوح می‌توان مشاهده کرد که فضای تصویر ماتریس S_B توسط مجموعه بردارهای $B = \{\zeta_1, ...\zeta_{c-1}\}$ ساخته می‌گردد. در صورتی که این بردارها نسبت به هم مستقل خطی باشند، آنگاه می‌توان

گفت که B یک پایه برای فضای تصویر این ماتریس است و بنابراین این فضا از بعد $C-1$ می‌باشد. **بنابراین می‌توان شرط آنکه ماتریس S_B دارای رنک $C-1$ باشد را به طور معادل نوشت:**

$$\begin{gathered} \mu_1-\mu \perp \mu_2-\mu \perp \ldots \perp \mu_c-\mu \\ \boxed{\equiv \mu_1 \perp \mu_2 \perp \mu_3 \ldots \perp \mu_c} \end{gathered} \tag{۴.۲}$$

۲.۲ ب

اگر فرآیند نمونه برداری به درستی انجام شده باشد در این صورت داریم:

$$rank(S_w)=d \rightarrow rank(S_w^{-1})$$

که مقصود از d همان بعد فضاست. از طرفی در قسمت الف اثبات شد:

$$rank(S_B) \leq C-1$$

حال با ترکیب این دو شرط و با استفاده از گزاره‌های جبرخطی پیرامون رنک ترکیب دو تبدیل می‌توان گفت:

$$\begin{gathered} rank(S_w^{-1}S_B) \leq \min rank(S_w^{-1}), rank(S_B) \leq C-1 \\ \boxed{rank(S_w^{-1}S_B) = \text{ Number of non zero eigenvalues} \leq C-1)} \end{gathered} \tag{۵.۲}$$

۳.۲ ت

$$\begin{aligned} & S_T=\Sigma_x(x-m)(x-m)^T=\Sigma_{i=1}^{C}\Sigma_{x\in D_i}(x-m-m_i+m_i)(x-m-m_i+m_i)^T \\ & =\Sigma_{i=1}^{C}\Sigma_{x\in D_i}(x-m_i)(x-m_i)^T+\Sigma_{i=1}^{C}\Sigma_{x\in D_i}(m-m_i)(m-m_i)^T \\ & =S_W+\Sigma_{i=1}^{C}n_i(m-m_i)(m-m_i)^T=S_W+S_B \\ & \rightarrow \boxed{S_T=S_W+S_B} \end{aligned} \tag{۶.۲}$$

۳

سوال ۴

به صورت کلی و باتوجه به iid بودن نمونه‌ها می‌توان نوشت:

$$\hat{p}_n(x) = E[p_n(x)] = \int \frac{1}{V_n}\phi(\frac{x-y}{h_n})p(y)dy$$
$$= \begin{cases} 0 \qquad x \le 0 \\ \frac{1}{V_n}\int_0^x \frac{1}{a}e^{-\frac{x-y}{h_n}}dy = \frac{1}{a}\frac{h_n}{V_n}(1-e^{\frac{-x}{h_n}}) \qquad 0 \le x \le a \\ \frac{1}{V_n}\int_0^a \frac{1}{a}e^{-\frac{x-y}{h_n}}dy = \frac{1}{a}\frac{h_n}{V_n}(e^{\frac{-x}{h_n}}-1)e^{\frac{-x}{h_n}} \qquad a \le x \end{cases} \tag{۱.۳}$$

از طرفی مطابق خاصیت نرمالیزه بودن پنجره پارزن داریم:

$$\int \frac{1}{V_n}\phi(\frac{x-x_i}{h_n})dx = \frac{V_n}{h_n} = 1 \to V_n = h_n \tag{۲.۳}$$

با ترکیب ۱.۳ ، ۲.۳ داریم:

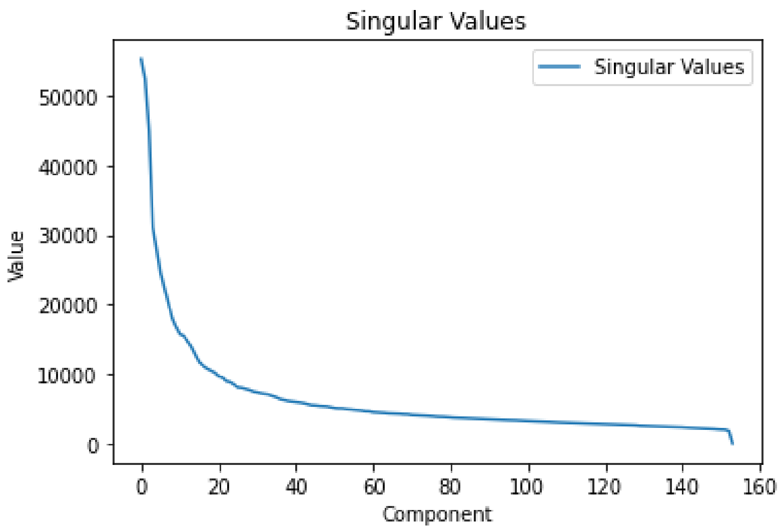
$$\to \boxed{\hat{p}_n(x) = \begin{cases} 0 \qquad x \le 0 \\ \frac{1}{a}(1-e^{\frac{-x}{h_n}}) \qquad 0 \le x \le a \\ \frac{1}{a}(e^{\frac{-x}{h_n}}-1)e^{\frac{-x}{h_n}} \qquad a \le x \end{cases}} \tag{۳.۳}$$

۴

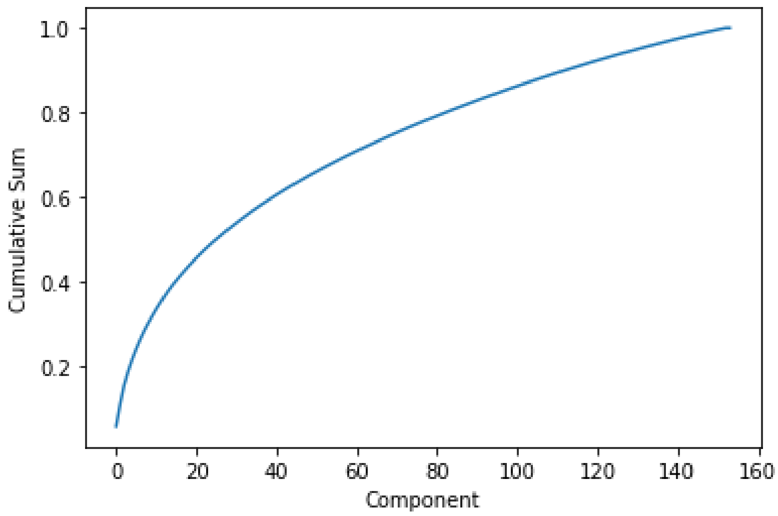
سوال۸

۱.۴ الف

شکل ۱.۴ بیانگر اجزای PCA می‌باشد.همانطور که در شکل مشاهده می‌شود، این مجموعه شامل مقادیر با اندازه‌های بسیار بزرگ تا مقادیر بسیار کوچک می‌باشد. به طور کلی می‌توان در نظر داشت که صرف نظر کردن از مقادیر کوچک در مقایسه با مقادیر بزرگ می‌تواند معیار خوبی برای بازسازی باشد. در واقع انتخاب مولفه‌هایی که سازنده ۹۰ درصد از واریانس باشند می‌تواند مناسب باشد. برای مثال در شکل ۲.۴ میزان واریانس پوشش داده شده توسط n نولفه بزرگ اول به صورت نسبی نمایش داده شده است. همانطور که مشاهده می‌شود در این مسئله استفاده از ۱۱۳ مولفه منجر به ساخت ۹۰ درصدی می‌گردد و **۴ مولفه اول تنها ۲۰ درصد از واریانس کل را در برخواهد داشت.**



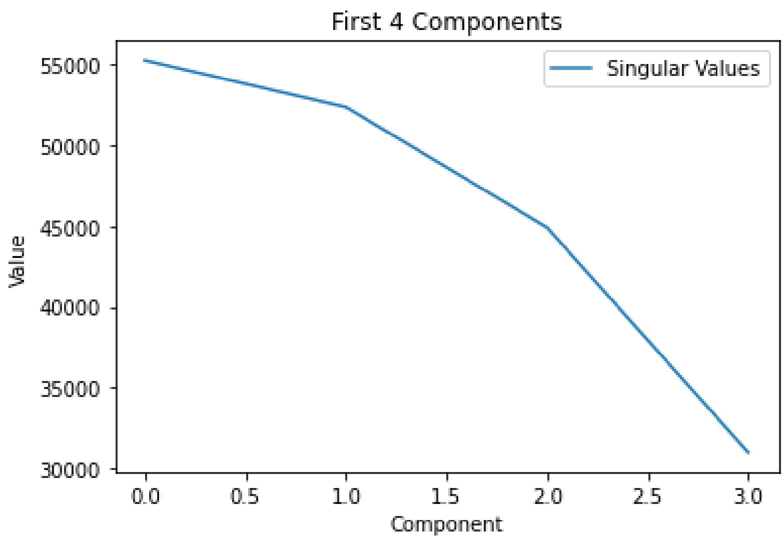
شکل ۱.۴: نمودار مولفه‌های PCA



شکل ۲.۴: نمودار میزان اثرگذاری n مولفه اول بزرگ

۲.۴ ب

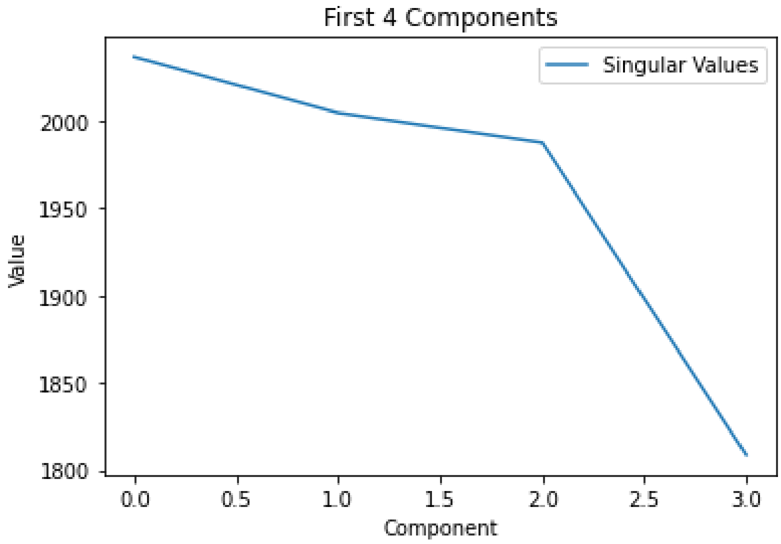
۴ مقدار ویژه بزرگ اول در شکل ۳.۴ رسم شده است. همانطور که در شکل ۲.۴ مشخص است این ۴ مولفه بیانگر 20% از کل واریانس هستند.



شکل ۳.۴: نمودار ۴ مولفه بزرگ اول

همچنین نمودار ۴ مولفه آخر در شکل ۴.۴ آورده شده است. همانطور که مشاهده می‌گردد این ۴ مولفه

نسبت به مولفه‌های ابتدایی اثر بسیار ناچیزی دارند.



شکل ۴.۴: تمودار ۴ مولفه آخر

۳.۴ پ

حال بر اساس شکل ۲.۴ متوجه می‌شویم با انتخاب ۱۱۰ مولفه اول به واریانسی برابر با 0.9 واریانس اولیه دست پیدا خواهیم کرد. سپس طبقه بند k نزدیکترین همسایه را پیاده می‌نماییم. مشاهده می‌گردد نتایج به شرح شکل ۵.۴ می‌باشد.

	With PCA Preprocess	No Preprocess
1NN	84.7458	84.7458
2NN	62.7119	62.7119

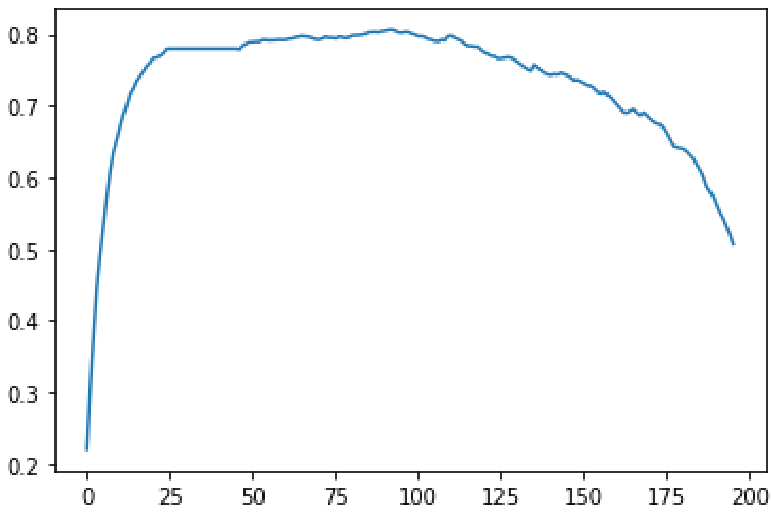
شکل ۵.۴: مقایسه دقت طبقه‌بندی‌های انجام شده در حضور طبقه‌بند و PCA

۵

سوال۹

۱.۵ الف

ابتدا نمودار CCR را برحسب ویژگی‌ها در یک نمودار رسم می‌نماییم. این نمودار در شکل ۱.۵ آورده شده است. در هر مرحله از الگوریتم مطابق الگوریتم Forward سعی شده است بهترین و موثرترین ویژگی از بین ویژگی‌های باقی‌مانده انتخاب شود



شکل ۱.۵: نمودار CCR برحسب تعداد ویژگی‌های انتخاب شده در الگوریتم Forward

۲.۵ ب

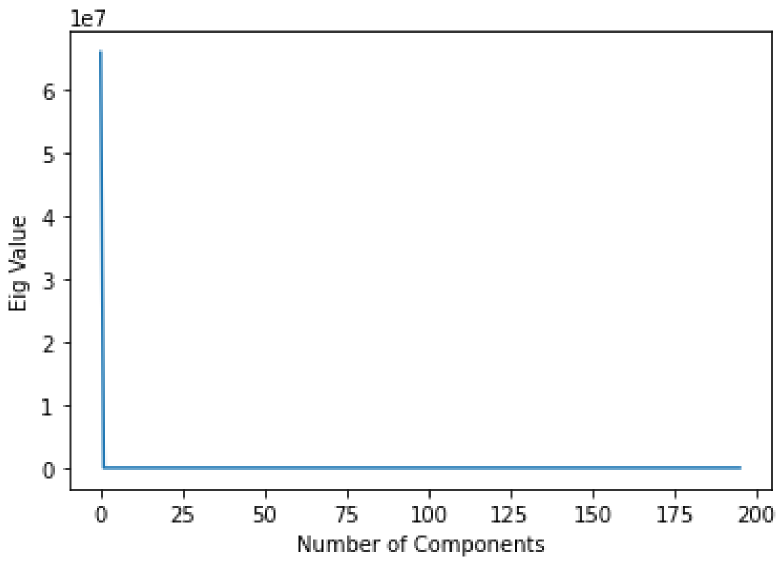
مطابق نمودار رسم شده در شکل ۱.۵ مشاهده می‌شود که طبقه‌بند در حضور ۹۰ ویژگی به دقت مطلوب دست پیدا کرده و در ادامه دقت ثابت و آهسته‌آهسته به سمت کاهش پیش‌میرود. بنابراین می‌توان مدعی شد که تعداد ۹۰ ویژگی گزینه مطلوب است.

۶

سوال ۱۰

۱.۶ الف

مطابق روابط درس به محاسبه‌ی ماتریس جداپذیری و مقادیر ویژه متناظر آن می‌پردازیم. نمودار این مقادیر در شکل ؟؟ ثبت شده است. همانطور که مشاهده می‌گردد به نظر می‌رسد عمده اطلاعات بر روی مولفه اول می‌باشد و سایر مولفه‌ها دارای مقدار ویژه نزدیک به صفر می‌باشند.

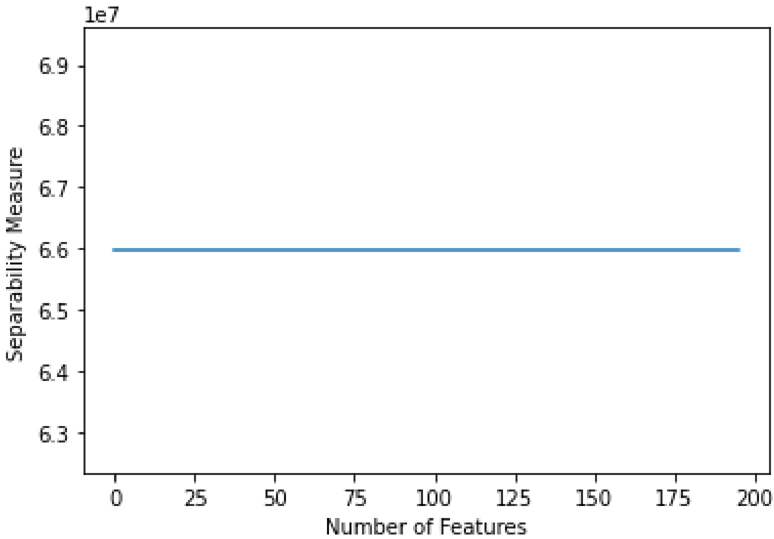


شکل ۱.۶: نمودار مقادیر ویژه ماتریس جدایی پذیری برحسب تعداد ویژگی‌ها

۲.۶ ب

معیار مورد نظر هم‌ارز با جمع k مولفه بزرگ اول مقدار ویژه‌هاست. نمودار معیار مورد نظر برحسب تعداد ویژگی‌های منتخب در شکل ۲.۶

رسم شده است. همانطور که در بند 'الف' نیز ذکر شد، به نظر می‌رسد که عموم اطلاعات ماتریس در



شکل ۲.۶: نمودار معیار جدایی پذیری برحسب تعداد ویژگی‌ها

مولفه‌ی اول قرار دارد و از این رو باقی مولفه‌ها تاثیر چندانی برروی این معیار نخواهند داشت. به نظر می‌آید انتخاب ۹ همان ۹ مولفه اول برای مسئله ۱۰ کلاسه($dim = C - 1$)کافی است.

۳.۶ ت

حال به طبقه بندی با استفاده از طبقه‌بند مذکور می‌پردازیم مشاهده می‌شود دقت نهایی به شرج زیر است:

$$CCR = 0.305$$

همانطور که مشاهده می‌شود این دقت پایین است. این امر می‌تواند به دلیل قدرت طبقه‌بند نیز باشد و اصولا ضعف ساختاری طبقه‌بند مانع از دقت مناسب گردد. به همین دلیل یک بار در حضور تمامی ویژگی‌ها نیز به آموزش طبقه‌بند می‌پردازیم. مشاهده می‌گردد در این حالت دقت طبقه بند برابر با 0.5 می‌شود. این موضوع اصلا منتظره نبوده. زیرا، مطابق دو بند قبل عمده‌ی اطلاعات ماتریس تنها در مولفه اول بردار ویژه‌ها بوده است و باقی مولفه‌ها فاقد اطلاعات چندانی هستند(تقریبا اطلاغاتی ندارند)به همین خاطر قاعدتا با انتخاب ویژگی‌های مذکور باید به جوابی حداقل هم‌ارز جواب اصلی رسید. اما دلیل این ناهمگونی احتمالا به خاطر تعداد کم دیتا می‌باشد. زیرا مسئله 196 بعد است و باید به صورت نمایی در این اردر دیتا داشته باشیم تا بتوان مدعی شد که تخمین ما از ماتریس کواریانس تخمین دقیقی است. در واقع به نظر می‌رسد که ماتریس

کواریانس فعلی باید متضمن اطلاعات بیشتری بر روی مولفه‌های دیگر باشد و بدین‌ترتیب است که توجیه انتخاب ویژگی‌های بیشتر از مولفه اول قابل قبول می‌باشد.

بنابراین در مسئله فعلی باید از روش‌هایی مانند FS به تعیین تعداد ویژگی‌ها پرداخت. مشاهده می‌گردد با انتخاب حدود ۱۰۰ ــ ۱۵۰ ویژگی محددا به دقت 50% دست پیدا می‌کنیم.