۱) Vladition و training برای hyper parameter و test فقط یک بار بعد از train برای سنجش

۲) الف) y۱: b↑, V↓ ← underfit   ب) b↑ V↑ ← (a)

y۲: b↓, V↓ ← good fit   b↓ V↓ ← (g)

y۳: b↓, V↑ ← over fit   b↓ V↑ ← (o)

۳)

الف) E | بعد از مدتی تغییری نمی کنند دلیل رفتار KNN

الف) B | ابتدا over fit و سپس under fit

ب) D | شدیدا اورفیت، فیت شده

ب) B | به دلیل over fit

پ) B | converge و مشکل local minima

ت) B | تقریبا مانند training

ت) A | زیادتر overfit

ث) B | بعد از مدتی overfit

باسمه تعالی

دانشگاه صنعتی شریف

دانشکده مهندسی کامپیوتر

مقدمه‌ای بر یادگیری ماشین

استاد درس: دکتر زارچی

آزمون میانترم - سوال دوم

خوشه خوشه

در این مسئله می‌خواهیم به بررسی الگوریتم خوشه‌بندی k-means بپردازیم. فرض کنید:

$$X = x_1, x_2, ..., x_n$$

داده‌های ما باشد و γ یک ماتریس Indicator باشد به این صورت که $\gamma_{ij} = 1$ اگر x_i متعلق به خوشه j باشد و در غیر این صورت برابر 0 است. فرض کنید $\mu_1, \mu_2, ..., \mu_k$ میانگین خوشه‌ها باشند. اعوجاج J برای داده‌ها به صورت زیر محاسبه می‌شود:

$$J(\gamma, \mu_1, ..., \mu_k) = n \sum_{j=1}^{k} \sum_{i=1}^{n} \gamma_{ij} ||x_i - \mu_j||^2$$

همچنین k را به عنوان مجموعه خوشه‌ها در نظر بگیرید $C = 1, ..., k$.

۱

آیا k-means نسبت به انتخاب نقاط اولیه حساس است، یعنی پاسخ آن بر اساس مجموعه‌ی نقاط اولیه تغییر می‌کند؟ اگر بله یک مثال ارائه کنید و اگر خیر، اثبات کنید. (۱۰ نمره)

پاسخ: بله (۳ نمره) مثال (۷ نمره)

۲

نشان دهید که الگوریتم k-means در زمان متناهی قدم به پایان می‌رسد. (راهنمایی: نشان دهید J تعداد محدودی حالت دارد.) (۲۰ نمره)

پاسخ:

نشان دادن اینکه مقدار اعوجاج بعد از هرگام یا ثابت میماند یا کاهش پیدا میکند (۸ نمره) و اگر ثابت ماند الگوریتم به پایان می‌رسد (در گام های بعد هم ثابت میماند) (۳ نمره) اشاره به اینکه فضای حالات محدود است و به دست آوردن این مقدار (۵ نمره) و نتیجه‌گیری نهایی (۴ نمره)

۳

اگر ابعاد داده نسبت به تعداد نمونه‌ها خیلی زیاد باشد و عملاً نمونه‌ها در یک فضای بزرگ پراکنده باشند، برای بهبود خوشه‌بندی از چه روشی استفاده می‌کنید؟

پاسخ:

اشاره به الگوریتم های مناسب مانند PCA و DBSCAN و ... (۱۰ نمره)

۴

نشان دهید که کمینه J یک تابع غیر افزایشی بر حسب k یا همان تعداد خوشه‌ها است. در این صورت آیا انتخاب مقدار هایپرپارامتر K بر اساس کمینه کردن مقدار J ایده‌ی خوبی است؟ اگر نه، چه ایده‌ی بهتری دارید؟ ۲۵ نمره

پاسخ:

با استفاده از استقرا، فرض کنید $L(k)$ کمینه تابع اعوجاج J برای k خوشه است. هدف نشان دادن این است که با افزایش تعداد خوشه‌ها از k به $k+1$، مقدار $L(k)$ کاهش یا ثابت می‌ماند: $L(k+1) \leq L(k)$.

خوشه‌بندی بهینه برای k خوشه، کمینه $L(k)$ است:

$$L(k) = \min_{\gamma, \mu_1, \ldots, \mu_k} J(\gamma, \mu_1, \ldots, \mu_k)$$

با افزایش تعداد خوشه‌ها به $k+1$، می‌توان یکی از خوشه‌های قبلی را به دو خوشه جداگانه تقسیم کرد. این تقسیم‌بندی جدید می‌تواند باعث کاهش مقدار تابع اعوجاج J شود یا حداقل مقدار آن ثابت بماند، زیرا تقسیم یک خوشه به دو خوشه می‌تواند انحراف درون‌خوشه‌ای را کاهش دهد. راه

دیگر اثبات: استفاده از انحراف درون خوشه‌ای و بین خوشه‌ای (۱۵ نمره)
شرح اینکه چرا انتخاب هایپرپارامتر K بر اساس کمینه کردن J ایده‌ی خوبی نیست (۴ نمره) و ایده‌ی بهتر (۶ نمره)

۵

فرض کنید $\hat{x}$ میانگین داده‌های نمونه باشد. مقادیر زیر را در نظر بگیرید:

$$T(X) = \frac{\sum_{i=1}^{n} ||x_i - \hat{x}||^2}{n}$$
$$W_j(X) = \frac{\sum_{i=1}^{n} \gamma_{ij} ||x_i - \mu_j||}{\sum_{i=1}^{n} \gamma_{ij}}$$
$$B(X) = \sum_{j=1}^{k} \frac{\sum_{i=1}^{n} \gamma_{ij}}{n} ||\mu_j - \hat{x}||$$

در اینجا $T(X)$ نشان دهنده انحراف کلی، $W_j(X)$ انحراف درون خوشه‌ای و $B(X)$ انحراف بین خوشه‌ای است. رابطه‌ی بین این ۳ مقدار به چه صورت است؟ نشان دهید که k-means می‌تواند به عنوان کمینه‌کننده میانگین وزن‌دار مقادیر درون خوشه‌ای و به طور تقریبی بیشینه کردن انحراف بین خوشه‌ای دیده شود. ۳۵ نمره

پاسخ:

رابطه‌ی بین این ۳ مقدار به این صورت است:

$$\begin{aligned} nT(X) &= \sum_{i=1}^{n} ||x_i - \hat{x}||^2 \\ &= \sum_{i=1}^{n} \|x_i - \hat{x}\|^2 = \sum_{i=1}^{n} \|x_i - \mu_j + \mu_j - \hat{x}\|^2 \\ &= \sum_{i=1}^{n} \|x_i - \mu_j\|^2 + \sum_{i=1}^{n} \|\mu_j - \hat{x}\|^2 + \sum_{i=1}^{n} (x_i - \mu_j)^T (\mu_j - \hat{x}) \\ &= \sum_{j=1}^{k} \sum_{i=1}^{n} \gamma_{ij} \|x_i - \mu_j\|^2 + \sum_{j=1}^{k} \sum_{i=1}^{n} \gamma_{ij} \|\mu_j - \hat{x}\|^2 + \sum_{j=1}^{k} \sum_{i=1}^{n} \gamma_{ij} (x_i - \mu_j)^T (\mu_j - \hat{x}) \end{aligned}$$

۸ نمره

حال داریم:

۱.

$$\sum_{j=1}^{k} \sum_{i=1}^{n} \gamma_{ij} \|x_i - \mu_j\|^2 = \sum_{j=1}^{k} n_j W_j(X) \quad \text{Where } : n_j = \sum_{i=1}^{n} \gamma_{ij}$$

۵ نمره

۲.

$$\sum_{j=1}^{k} \sum_{i=1}^{n} \gamma_{ij} \|\mu_j - \hat{x}\|^2 = \sum_{j=1}^{k} n_j ||\mu_j - \hat{x}||^2 = nB(X)$$

۵ نمره

۳.

$$\mu_j = \frac{\sum_{i=1}^{n} \gamma_{ij} x_i}{\sum_{i=1}^{n} \gamma_{ij}} \Rightarrow \sum_{i=1}^{n} \gamma_{ij} (x_i - \mu_j)^T (\mu_j - \hat{x}) = 0$$

۵ نمره

پس داریم:

$$nT(X) = \sum_{j=1}^{k} n_j W_j(X) + nB(X)$$

۴ نمره

از آنجا که $T(X)$ ثابت است، الگوریتم k-Means $W(X)$ را کمینه و در نتیجه $B(X)$ را بیشینه می‌کند. ۸ نمره

پاسخنامه و ریزنمرات بخش خلاف شیب

سوال ۱:

$$w^{t+1} = w^t - \eta \nabla_{w_i} F(w)$$ (۵ نمره)

$$F(w) = l(x^{(j)}, y^{(j)}, w) = 0 \rightarrow \nabla_w F(w) = \nabla_w \left[y^{(j)} (\sum_{i=1}^{d} w_i x_i^{(j)}) - ln(1 + exp(\sum_{i=1}^{d} w_i x_i^{(j)})) \right]$$

$$\nabla_{w_j} F(w) = y^{(j)} x_i^{(j)} - \frac{x_i^{(j)} exp(\sum_{i=1}^{d} w_i x_i^{(j)})}{1 + exp(\sum_{i=1}^{d} w_i x_i^{(j)})}$$

(۱۵ نمره)

سوال ۲:

در صورت استفاده از داده ساختارهای متراکم میانگین پیچیدگی زمانی از حاصلضرب ابعاد پیروی می‌کند $O(NX)$ (۱۰ نمره)

در صورت استفاده از داده ساختارهای تنک میانگین پیچیدگی زمانی از تعداد داده‌های غیرصفر پیروی می‌کند. $O(d)$ (۱۰ نمره)

سوال ۳:

$$w_i^{(t+1)} = w_i^t - \eta \nabla_{w_i} F(w)$$ (۵ نمره)

$$\frac{\partial F(w)}{\partial w_i} = y^{(j)} x_i^{(j)} - \frac{x_i^{(j)} . exp(\sum_{i=1}^{d} w_i x_i^{(j)})}{1 + exp(\sum_{i=1}^{d} w_i x_i^{(j)})} - \lambda w_i$$

(۱۵ نمره)

سوال ۴:

تنها تفاوت این است که از یک ضریب ثابت جدید استفاده کردیم که تاثیری در پیچیدگی ندارد. (۱۰ نمره)
بنابراین میانگین پیچیدگی زمانی برابر است با $O(NX)$.
(۱۰ نمره)

سوال ۵:

اشاره به صفر بودن جمله دوم معادله آپدیت وزن‌ها (۵ نمره)

$$w_i^{(t+1)} = w_i^t - \eta\lambda w_i^t = (1 - \eta\lambda)w_i^t$$

$$w_i^{(t+k)} = (1 - \eta\lambda)w_i^{(t+k-1)} = (1 - \eta\lambda)^k w_i^t$$

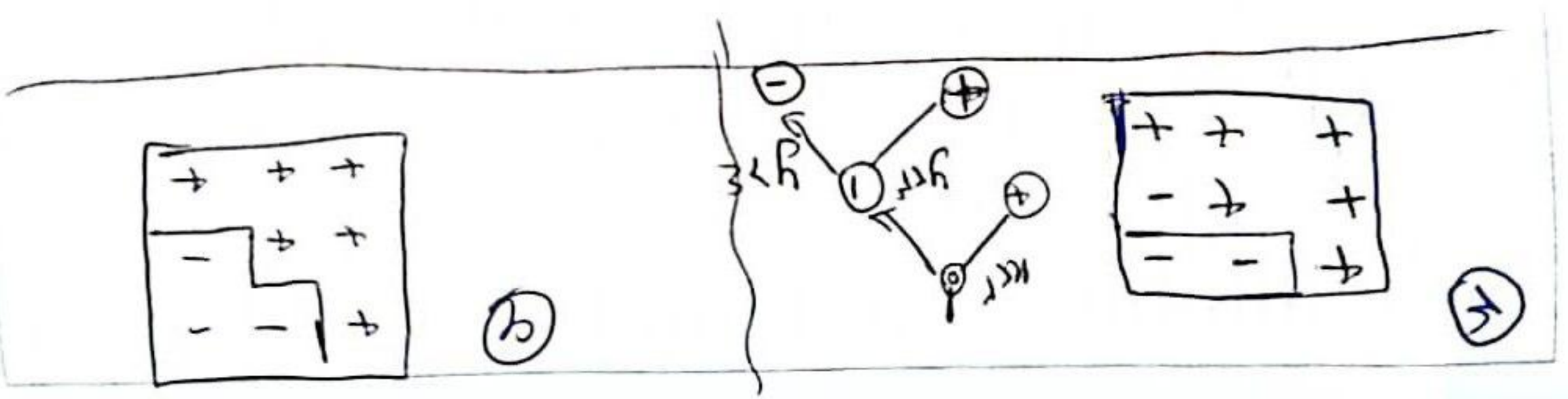
نوشتن دو معادله بالا (۵ نمره)

سوال ۶:

باتوجه به زمان نیاز برای به روزرسانی وزن‌ها و خلاقیت الگوریتم از ۱۰ نمره داده شود.

DATE

① ② ③ ④ ⑤ ⑥



آورید.

$$\frac{\partial L(w)}{\partial w} = -2X^Ty + 2X^TXw + 2\Gamma^T\Gamma w = 0$$

میدانیم $\left(L(w) = y^Ty + w^TX^TXw - 2w^TX^Ty + w^T\Gamma^T\Gamma w \right)$

و از مشتق ترم عبارت بالا میشود

$$\boxed{w^* = \left(X^TX + \Gamma^T\Gamma\right)^{-1} X^Ty}$$

بهینه برای $\hat{w}$ دارد.

برای اینکه مطمئن شویم دو بردار داشته باشیم این ستون‌های $X^TX+\Gamma^T\Gamma$ مستقل خطی باشند

بین بردارهای $X^TX_j+\Gamma^T\Gamma_j$ مستقل خطی باشند

X می باشد.

$$l(w) = \mathbb{E}\left[\|y - (D \odot X)w\|^2\right]$$

$D \sim \text{bernoulli}(p)$ ← هر درایه X با احتمال p است

$$= \mathbb{E}\left[(y^T y - 2w^T (D \odot X)^T y + w^T (D \odot X)^T (D \odot X) w)\right]$$

$$= y^T y - 2w^T \underbrace{\mathbb{E}(D \odot X)^T}_{p} y + w^T \mathbb{E}\left(\|D \odot X\|^2\right) w$$

$$= \|y - pX\hat{w}\|^2 + p(1-p)\|\hat{w}\|^2$$

$$\mathbf{L}(\hat{w}) = \|y - \hat{X}\hat{w}\|_2^2 + \lambda\|\hat{w}\|_2^2$$

$$L(w) = \|y - Xw\|^2 + \|\Gamma w\|^2$$

تغییر متغیر $\Rightarrow$ $\hat{w} = \frac{(\Gamma w)}{\sqrt{\lambda}} \Rightarrow \Gamma w = \sqrt{\lambda}\hat{w}$

$$\|\Gamma w\|^2 = \lambda\|\hat{w}\|^2$$

Γ وارون‌پذیر $\Rightarrow w = \sqrt{\lambda}\,T^{-1}\hat{w} \Rightarrow \|y - Xw\|^2 =$

$$= \|y - \underbrace{\sqrt{\lambda}\,XT^{-1}}_{\hat{X}}\hat{w}\|^2 = \|y - \hat{X}\hat{w}\|^2$$

$$\Rightarrow L(\hat{w}) = \|y - \hat{X}\hat{w}\|^2 + \lambda\|\hat{w}\|^2$$

پاسخ دادم بندی :

1. اشاره به تغییرات لازم در String های داده شده لازم است. (0.5 نمره)

جواب یکتایی وجود ندارد. یک مثال می تواند استفاده از k-means و یا رگرسیون باشد (2.5 نمره)

2. کشن Centeralize مهم است (1 نمره)

میزان تأثیرگذاری را برای ویژگی ها یکسان کنیم یا به عبارتی نرمال سازی ویژگی ها را داشته باشیم (2 نمره)

مطابق با توضیحات و دلایل برای موارد بالا و یا جواب های دیگر (1 نمره)

3.

$$C = \begin{bmatrix} 3 & -4 \\ -4 & 3 \end{bmatrix} \rightsquigarrow \text{هدف اولیه: یافتن مقادیر ویژه}$$

$$\Rightarrow \det(C - \lambda I) = 0 \rightsquigarrow (3-\lambda)^2 - 16 = 0 \rightsquigarrow \lambda^2 - 6\lambda - 7 = 0$$

$$(\lambda + 1)(\lambda - 7) = 0$$

مقادیر ویژه ما: $\lambda_1 = 7$ $\lambda_2 = -1$ (2 نمره)

$$Cv_1 = \lambda_1 v_1 \rightarrow v_1 = \begin{bmatrix} \sqrt{2}/2 \\ -\sqrt{2}/2 \end{bmatrix}$$

$$Cv_2 = \lambda_2 v_2 \rightarrow v_2 = \begin{bmatrix} \sqrt{2}/2 \\ \sqrt{2}/2 \end{bmatrix}$$

(2 نمره)

4. ابتدا باید ماتریس کواریانس را بیابید. سپ مقادیر ویژه را به دست آورد $(\lambda_1, \lambda_2, \dots, \lambda_n)$ با ترتیب نزولی $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$

با توجه به مقدار اطلاعاتی که نیاز داریم اگر $k\%$ باشد $\leftarrow$ $m \leftarrow \min m \left(\frac{\sum_{i=1}^{m} \lambda_i}{\sum_{j=1}^{n} \lambda_j} \geq k\% \right)$ به صورت نیاز ماست (3 نمره)

$\arg\max \; Var(Xv) \rightsquigarrow Var(Xv) = v^T X^T X v$
$\|v\|_2^2 = 1$

طبق لاگرانژ : $L = v^T X^T X v - \lambda (\|v\|_2^2 - 1)$

$\rightsquigarrow \frac{\partial L}{\partial v} = 0 = 2 X^T X v - 2\lambda v \rightsquigarrow n\Sigma v = \lambda v$ (2 نمره)

$n\Sigma$ ← Cov mat

$\Rightarrow Var(Xv) = v^T X^T X v = n\, v^T (\Sigma v) = \lambda v^T v = 1$ ← $\frac{\lambda v}{n}$

بزرگترین مقدار ویژه جالاترین واریانس را دارد (2 نمره)

6. طبق خلاصیت و جالب بودن سؤال مطرح شده. (2 نمره)