بسمه تعالی

بازشناسی آماری الگو

دانشگاه صنعتی خواجه نصیرالدین طوسی ۱۳۰۷ | تمرین سری دهم | تاریخ تحویل: ۱۴۰۴/۱۰/۶ | آزمایشگاه بینایی ماشینی و پردازش تصاویر پزشکی

۱- فرض کنید s یک معیار شباهت متریک تعریف شده روی مجموعه X باشد و داشته باشیم: $s(x,y)>0,\ \forall x,y \in X$ چنانچه $f: R^+ \to R^+$ یک تابع پیوسته کاهشی یکنواخت باشد بطوریکه داشته باشیم:

$$f(x) + f(y) \geq f\left(\frac{1}{\frac{1}{x}+\frac{1}{y}}\right), \quad \forall x,y \in R^+$$

نشان دهید $f(s)$ یک معیار عدم شباهت متریک روی مجموعه X است.

۲- (اختیاری) ثابت کنید مقادیر بیشینه و کمینه معیار شباهت فازی $s_F^q(\mathbf{x},\mathbf{y}) = \left(\sum_{\mathrm{i}=1}^{l} s(x_i,y_i)^q\right)^{1/\mathrm{q}}$ بترتیب با $l^{1/q}$ و $0.5l^{1/\mathrm{q}}$ برابر است.

۳- (اختیاری) شبه کد الگوریتم‌های خوشه‌بندی ترتیبی (BSAS) و خوشه‌بندی ترتیبی اصلاح شده (MBSAS) را با استفاده از یک معیار شباهت بجای معیار عدم شباهت بنویسید.

۴- (اختیاری) الف- معادلات ۱۴.۵۶ و ۱۴.۵۷ کتاب تئودوریدیس را برای الگوریتم پوسته‌های بیضیگونی فازی (FCES) بدست آورید. ب- روابط مربوط به تعیین پارامترهای الگوریتم FCES را بدست آورید.

۵- (اختیاری) مجموعه داده زیر را در نظر بگیرید.

$$x_1 = (-1,1), x_2 = (1,1), x_3 = (-1,-1), x_4 = (1,-1)$$

می‌خواهیم از خوشه‌بندی k-means با فاصله اقلیدسی استفاده کنیم با K=2. فرض کنید در ابتدا دو نقطه که به شکل تصادفی انتخاب می‌شوند به عنوان مراکز خوشه‌ها در نظر گرفته می‌شوند.

الف- تمام نتایج ممکن پس از اعمال الگوریتم خوشه‌بندی را بنویسید.

ب- از بین همه نتایج ممکن، کدامیک کمترین هزینه را دارد؟ هزینه هر نتیجه را مجموع فاصله اقلیدسی هر نقطه از مرکز دسته‌ای که به آن نسبت داده شده است در نظر بگیرید.

ج- در شرایطی مثل حالت فوق، چطور می‌توان با استفاده از الگوریتم k-means به نتیجه‌ای رسید که کمترین هزینه را داشته باشد؟

۶- (اختیاری) مساله امتحان پایان ترم بازشناسی آماری الگو سال ۱۳۸۹

در یک مساله خوشه‌بندی از معیار مجموع مربعات خطا استفاده می‌شود:

$$J = \sum_{i=1}^{M} J_i, \qquad J_i = \sum_{\mathbf{x}\in C_i} \|\mathbf{x} - \hat{\mathbf{m}}_i\|^2$$

که در آن M تعداد خوشه‌ها و C_i مجموعه نمونه‌های خوشه i-ام می‌باشد. چنانچه یک نمونه $\mathbf{x}$ از خوشه i-ام به خوشه j-ام منتقل شود، نشان دهید مقادیر جدید مرکز ثقل خوشه‌ها و نیز تابع هزینه متناظر این خوشه‌ها به فرم زیر تغییر می‌یابند:

$$\widehat{\mathbf{m}}_j' = \widehat{\mathbf{m}}_j + \frac{\mathbf{x} - \widehat{\mathbf{m}}_j}{N_j + 1} \qquad , \qquad J_j' = J_j + \frac{N_j}{N_j + 1} \left\| \mathbf{x} - \widehat{\mathbf{m}}_j \right\|^2$$
$$\widehat{\mathbf{m}}_i' = \widehat{\mathbf{m}}_i - \frac{\mathbf{x} - \widehat{\mathbf{m}}_i}{N_i - 1} \qquad , \qquad J_i' = J_i - \frac{N_i}{N_i - 1} \left\| \mathbf{x} - \widehat{\mathbf{m}}_i \right\|^2$$

$\widehat{\mathbf{m}}_j$ و $\widehat{\mathbf{m}}_i$: مراکز ثقل خوشه‌ها قبل از انتقال نمونه $\mathbf{x}$ از خوشه i-ام به خوشه j-ام

$\widehat{\mathbf{m}}_j'$ و $\widehat{\mathbf{m}}_i'$: مراکز ثقل خوشه‌ها بعد از انتقال نمونه $\mathbf{x}$ از خوشه i-ام به خوشه j-ام

N_j و N_i : تعداد نمونه‌های خوشه‌های j-ام و i-ام قبل از انتقال نمونه

۷- (اختیاری) مساله امتحان پایان ترم سال ۱۳۹۶

در فضای n بعدی فرض کنید N نمونه درون M خوشه مجزا دسته‌بندی شده باشند و تابع هزینه خوشه‌بندی که باید کمینه شود بصورت زیر تعریف شده باشد:

$$J_e = \sum_{i=1}^{M} \sum_{j=1}^{N_i} \left\| \mathbf{x}_{ij} - \mathbf{m}_i \right\|^2$$

که در آن M تعداد خوشه‌ها، N_i تعداد نمونه‌ها در خوشه i-ام، $\mathbf{m}_i$ مرکز ثقل نمونه‌ها در خوشه i-ام و $\mathbf{x}_{ij}$ نمونه j-ام از خوشه i-ام می‌باشد. بازای N>M نشان دهید که چنانچه یک خوشه حاوی هیچ نمونه‌ای نباشد، این خوشه‌بندی بهینه نیست. به عبارت دیگر برای حداقل شدن J_e ضروری است که حداقل یک نمونه در هر خوشه وجود داشته باشد. (راهنمایی: ابتدا فرض کنید یکی از خوشه ها حاوی هیچ نمونه ای نیست و سپس نمونه ای را از یکی از خوشه ها به این خوشه منتقل کنید.)

۸- (اختیاری) مساله پایان ترم ۱۳۹۷

در فضای n بعدی فرض کنید فاصله بین دو خوشه‌ی A و B بصورت زیر تعریف شده باشد:

$$d(A, B) = \| \mathbf{m}_A - \mathbf{m}_B \|^2$$

که در آن $\mathbf{m}_A$ و $\mathbf{m}_B$ مراکز ثقل خوشه‌ها هستند. با استفاده از تعریف فوق چنانچه بخواهیم فاصله بین خوشه‌ی K و یک خوشه حاصل از ادغام دو خوشه‌ی I و J را محاسبه کنیم نشان دهید که می‌توان نوشت:

$$d(K, I + J) = \frac{N_I}{N_I + N_J} d(I, K) + \frac{N_J}{N_I + N_J} d(J, K) - \frac{N_I N_J}{(N_I + N_J)^2} d(I, J)$$

که در آن N_I و N_J به ترتیب تعداد نمونه‌ها در خوشه‌ی I و J هستند.

۹- (اختیاری) سوال امتحانی پایان ترم ۱۳۹۸

در فضای n بعدی فرض کنید N نمونه در اختیار است و می‌خواهیم با استفاده از الگوریتم K-means آنها را در K خوشه‌ی مجزا خوشه‌بندی نماییم. مطابق این الگوریتم، مجموعه بردارهای مراکز خوشه‌ها $\boldsymbol{\mu} = \{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_K\}$ با کمینه‌سازی فاصله‌ی متوسط نمونه‌ها از نزدیکترین مرکز خوشه بدست می‌آید. به عبارت دیگر در این روش تابع هزینه‌ی زیر کمینه می‌شود:

$$L(\boldsymbol{\mu}) = \sum_{i=1}^{N} \min_{j \in \{1, \ldots, K\}} \left\| \boldsymbol{x}_i - \boldsymbol{\mu}_j \right\|^2$$

و اختصاص هر نمونه $\boldsymbol{x}_i$ به خوشه j-ام بر اساس متغیر $z_i = \arg min_{j \in \{1, \ldots, K\}} \left\| \boldsymbol{x}_i - \boldsymbol{\mu}_j \right\|^2$ انجام می‌شود. یعنی z_i معرف برچسب نزدیکترین مرکز خوشه به نمونه‌ی $\boldsymbol{x}_i$ می‌باشد. الگوریتم K-means بصورت تکراری در دو مرحله‌ی به‌روزسازی برچسب‌های z_i (مرحله‌ی

برچسب زنی) و به‌روزسازی مراکز خوشه‌ها $\boldsymbol{\mu}_j = \frac{1}{|\{i:z_i=j\}|}\sum_{i:z_i=j} \boldsymbol{x}_i$ (مرحله بازتنظیم) اجرا می شود. الگوریتم هنگامی متوقف می‌شود که در مرحله‌ی برچسب زنی تغییری در برچسب نمونه‌ها بوجود نیاید. نشان دهید که الگوریتم K-means همواره به یک نقطه کمینه (محلی یا سراسری) همگرا می‌شود.

راهنمایی: کافی است نشان دهید که تابع هزینه بصورت تکراری تا رسیدن به نقطه همگرایی بصورت یکنواخت کاهش می‌یابد. این روند کاهشی در هر تکرار الگوریتم را برای مرحله‌ی برچسب زنی و مرحله بازتنظیم بطور جداگانه نشان دهید.

۱۰- سوال امتحانی پایان ترم ۱۳۹۹

تابع هزینه برای مساله خوشه بندی k-میانگین با k خوشه، و نمونه های $\mathbf{x}_1,\ldots,\mathbf{x}_n$ و مراکز خوشه $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_n$ بصورت زیر تعریف شد:

$$L = \sum_{j=1}^{k} \sum_{\mathbf{x}_i \in S_j} \left\| \mathbf{x}_i - \boldsymbol{\mu}_j \right\|^2$$

که در آن S_j زیرمجموعه نمونه‌هایی است که به $\boldsymbol{\mu}_j$ نسبت به سایر مراکز خوشه ها نزدیک‌تر هستند.

الف- به جای به روز سازی $\boldsymbol{\mu}_j$ بر اساس متوسط‌گیری از نمونه های S_j، می‌خواهیم از روش گرادیان نزولی دسته‌ای (batch gradient descent algorithm) که در آن زیرمجموعه‌های S_j ثابت نگه داشته می‌شوند استفاده کنیم. رابطه به روز سازی $\boldsymbol{\mu}_j$ را چنانچه نرخ آموزش ε باشد بدست آورید.

ب- رابطه به روز سازی $\boldsymbol{\mu}_j$ را بر اساس الگوریتم گرادیان نزولی برخط (به ازای یک تک نمونه) و نرخ آموزش ε بنویسید.

ج- در این قسمت می خواهیم معادل بودن رابطه بدست آمده در قسمت الف را با الگوریتم استاندارد k-میانگین بررسی کنیم. در الگوریتم استاندارد، میانگین نمونه‌های هر خوشه را به عنوان مرکز آن خوشه در نظر می‌گیریم. به نظر می‌رسد چنانچه در الگوریتم بدست آمده در بند الف، یک مقدار خاص برای نرخ آموزش (یعنی ε) در نظر گرفته شود، رابطه به روز سازی مراکز خوشه ها در هر دو روش یکسان خواهد شد. چنانچه $\boldsymbol{\mu}_1$ مرکز خوشه شماره ۱ باشد، مقدار ε را طوری تعیین کنید که رابطه به روز سازی مرکز این خوشه در هر دو روش یکسان شود. (توجه شود که مقدار ε می تواند بطور خاص برای به روز سازی خوشه شماره ۱ بدست آید و برای دیگر خوشه ها متفاوت باشد).

۱۱- (اختیاری) فرض کنید داده‌هایی به ما داده شده که شامل چند کلاس مختلف است و هر کلاس دارای یک توزیع احتمال متفاوت می‌باشد. برچسب نمونه‌ها در دسترس نیست و قرار است از k-means به منظور خوشه‌بندی استفاده شود. شرایطی که باعث تضعیف اثربخشی k-means می‌گردند را علامت بزنید.

الف. برخی کلاس‌ها دارای توزیع نرمال نباشند

ب. واریانس توزیع‌ها از تمام جهات دارای واریانس کمی باشد

ج. میانگین کلاس‌ها با هم برابر باشد.

د. k=n انتخاب شود که n تعداد نمونه‌های در دسترس است.

۱۲- (پایان ترم ۱۴۰۲) معیار زیر برای خوشه‌بندی بردارهای $\mathbf{x}$ در داخل c خوشه بکار می‌رود:

$$J_G = \sum_{i=1}^{c} \sum_{\mathbf{x} \in R_i} (\mathbf{x} - \mathbf{m}_i)^t \Sigma_T^{-1} (\mathbf{x} - \mathbf{m}_i)$$

در رابطه فوق:

$$\mathbf{m}_i = \frac{1}{N_i} \sum_{\mathbf{x} \in R_i} \mathbf{x} \qquad \Sigma_T = \sum_{\mathbf{x} \in X} (\mathbf{x} - \mathbf{m})\,(\mathbf{x} - \mathbf{m})^t \qquad \mathbf{m} = \frac{1}{N} \sum_{\mathbf{x} \in X} \mathbf{x}$$

در روابط فوق X معرف مجموعه کل نمونه‌ها، $\mathbf{m}$ معرف مرکز ثقل عمومی، Σ_T ماتریس کواریانس عمومی، R_i ناحیه مربوط به خوشه iام، N_i تعداد نمونه‌های قرار گرفته در خوشه iام و $\mathbf{m}_i$ مرکز ثقل خوشه iام را مشخص می‌کند. حداقل‌سازی معیار فوق خوشه‌بندی بر اساس حداقل مربعات فاصله با معیار فاصله ماهالانوبیس را نتیجه می‌دهد. نشان دهید اگر یک نمونه $\mathbf{y}$ از خوشه i به خوشه j منتقل شود، مقدار تابع هزینه بصورت زیر تغییر می‌کند:

$$J'_G = J_G + \left[\frac{N_j}{N_j+1}(\mathbf{y}-\mathbf{m}_j)^t\Sigma_T^{-1}(\mathbf{y}-\mathbf{m}_j) - \frac{N_i}{N_i-1}((\mathbf{y}-\mathbf{m}_i)^t\Sigma_T^{-1}(\mathbf{y}-\mathbf{m}_i)\right]$$

۱۳- (پایان ترم ۱۴۰۳)

الف- نتایج خوشه‌بندی (اعضای هر خوشه و مراکز خوشه‌ها) با استفاده از الگوریتم k-means، مجموعه داده زیر، پس از یک تکرار را گزارش دهید. (تعداد خوشه‌ها را ۳ در نظر بگیرید و در ابتدا داده‌های ۱، ۴ و ۷ را مراکز خوشه در نظر بگیرید و از فاصله اقلیدسی استفاده کنید).

A1 = (2,10), A2 = (2,5), A3 = (8,4), A4 = (5,8), A5 = (7,5), A6 = (6,4), A7 = (1,2), A8 = (4,9).

در زیر ماتریس فاصله اقلیدسی داده‌ها نسبت به یکدیگر، جهت افزایش سرعت حل این مسئله، به شما داده شده‌است.

	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	$\sqrt{25}$	$\sqrt{36}$	$\sqrt{13}$	$\sqrt{50}$	$\sqrt{52}$	$\sqrt{65}$	$\sqrt{5}$
A2		0	$\sqrt{37}$	$\sqrt{18}$	$\sqrt{25}$	$\sqrt{17}$	$\sqrt{10}$	$\sqrt{20}$
A3			0	$\sqrt{25}$	$\sqrt{2}$	$\sqrt{2}$	$\sqrt{53}$	$\sqrt{41}$
A4				0	$\sqrt{13}$	$\sqrt{17}$	$\sqrt{52}$	$\sqrt{2}$
A5					0	$\sqrt{2}$	$\sqrt{45}$	$\sqrt{25}$
A6						0	$\sqrt{29}$	$\sqrt{29}$
A7							0	$\sqrt{58}$
A8								0

ب- شبه کد انجام محاسبات بند الف را در حالت کلی یعنی C خوشه و تعداد نمونه های ورودی برابر N بنویسید و در آن ورودی ها، خروجی ها و حلقه عملیات تا رسیدن به نتیجه را مشخص کنید.

موفق باشید