دانشگاه تهران
پردیس دانشکده‌های فنی
دانشکده مهندسی برق و کامپیوتر

یادگیری ماشین

تمرین شماره ۳

نام و نام خانوادگی
مرتضی ملکی‌نژاد شوشتری

شماره دانشجویی
۸۱۰۱۰۴۲۵۶

۲۷ بهمن ۱۴۰۴

فهرست مطالب

۱ چکیده ۱

۱ K-means of Convergence ۲

۱.۱ a ۲

۲.۱ b ۳

۲ Estimation MAP and K-Means Regularized ۴

۱.۲ a ۴

۲.۲ b ۴

۳ Identity Objective K-Means The ۶

۴ Model Mixture Gaussian ۷

۱.۴ A ۷

۲.۴ B ۷

۵ EM ۸

۱.۵ a ۸

۲.۵ b ۸

۳.۵ c ۸

فهرست تصاویر

۱ در p1 و p2 هرکدام دو نقطه وجود دارد (یعنی در کل شش نقطه موجود است). اگر هرکدام از آن‌ها یکبار به c1 و یکبار به c2 متعلق شوند، خوشه‌بندی ممکن است هیچوقت تمام نشود. ۲

فهرست جداول

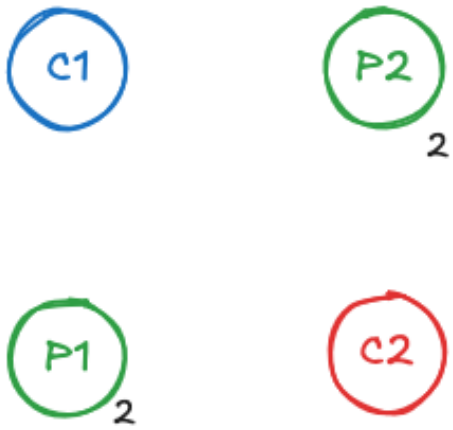
۱ ویژگی‌های انتخاب شده در هر دو روش ۷

چکیده

Convergence of K-means

1.1 a

- حتی با یک tie braking که deterministic نیست هم، مقدار فاصله درون خوشه‌ای زیاد نمی‌شود. اما deterministic نبودن باعث می‌شود تا فاصله درون خوشه‌ای بهبود نیابد، جواب قابل دستیابی دوباره نباشد، بهترین جواب بدست نیاید و یا خوشه‌بندی بیشتر طول بکشد. ولی اگر معیار توقف روی عدم بهبود فاصله درون خوشه‌ای باشد، الگوریتم حتما پایان می‌پذیرد. اگر معیار پایان الگوریتم، عدم تغییر خوشه‌ها باشد ولی ممکن است که پایان نیابد. در شکل ۱ مثالی از این مورد زده شده است.



شکل ۱: در p1 و p2 هرکدام دو نقطه وجود دارد (یعنی در کل شش نقطه موجود است). اگر هرکدام از آن‌ها یکبار به c1 و یکبار به c2 متعلق شوند، خوشه‌بندی ممکن است هیچوقت تمام نشود.

- اگر کلاس با پایین ترین شماره اساین شود هیچوقت نوسان کلاس رخ نمی‌دهد. حال تنها چیزی که باید ثابت شود این است که تابع F همواره غیر صعودی است. هدف F درواقع یافتن مقادیر c است به طوری که مقدار زیر کمینه شود:

$$F = \sum \sum ||(x - c_i)^2||$$

اگر فقط برای یک خوشه درنظر بگیریم:

$$w = \sum (x - c_i)^2 \xrightarrow{derviative} \sum 2(x - c_i) = 0 \Rightarrow nc_i = n\bar{x} \Rightarrow c_i = \bar{x}$$

یعنی میانگین هر خوشه، فاصله درون خوشه‌ای را کمینه می‌کند. و چون این میانگین معیار تشخیص مرکز جدید در KNN است پس همواره رو به کم شدن می‌رود. پس هیچوقت زیاد نمی‌شود.

1.2 b

- چون خوشه هیچ عضوی ندارد، نمی‌توان مرکز خوشه را حساب کرد.
- می‌توان نقطه‌ای که از مرکز خوشه‌ها بیشترین فاصله را دارد به عنوان مرکز خوشه جدید درنظر گرفت.
- بله. می‌توان با استفرا ثابت کرد که پس از انتخاب مرکز خوشه جدید فاصله حتما کاهش می‌یابد. به ازای هر نقطه،اگر در حالتی که مرکز جدید نبود فاصله آن با نزدیکترین مرکز d_i بوده، همچنان این مرکز موجود است و نقطه اگر به مرکز جدید نباشد، همچنان باعث می‌شود مقدار d_i به فاصله درون خوشه‌ای اضافه شود و نه بیشتر. ولی نقطه‌ای که به تازگی مرکز خوشه شده است، فاصله درون خوشه‌ای آن صفر می‌شود. پس یعنی فاصله درون خوشه‌ای در حالت کل کمتر می‌شود.

Regularized K-Means and MAP Estimation

2.1 a

۱. داریم:

$$J = \sum(x_j - \mu_i)^2 + \lambda\mu_i^2 \xrightarrow{derviative} \sum 2(\mu_i - x_j) + 2\lambda\mu_j \Rightarrow n\mu_j - n\bar{x} + \lambda\mu_j = 0 \Rightarrow \mu_j = \frac{n\bar{x}}{n+\lambda}$$

۲. اگر مقدار λ خیلی کوچک باشد، عملا Regularization ای نداریم و همان میانگین گروه درنظر گرفته می‌شود. و اگر این مقدار بسیار بزرگ باشد، میانگین ها به صفر نزدیک می‌شوند.

2.2 b

۱. برای این کار اول باید جمع را به ضرب تبدیل کرد. در MAP با گرفتن لگاریتم ضرب به جمع تبدیل می‌شود پس اینجا باید برعکس آنکار را کرد (و چون هم لگاریتم و هم توان رساندن رابطه یکنوا و صعودی هستند، بیشینه کردن هر دوعبارت یکسان است.) پس داریم:

$$(\prod e^{(x_j-\theta)^2})e^{\lambda(\theta^2)}$$

۲. برای posterior باید توزیع با پارامتر θ ای را پیدا کرد که تنا فقط در عبارت $e^{(x-\theta)^2}$ تاثیر داشته باشد. برای مثال توزیع نرمال با واریانس یک همچنین ویژگی‌ای دارد.

توزیع prior نیز شبیه یک توزیع نرمال است که واریانس آن برابر با $\frac{1}{\lambda}$. پس یعنی اگر ضرایب ثابت درنظر گرفته نشوند، مقداری که سعی در بهینه کردن آن داریم معادل یک MAP است که در آن توزیع prior یک توزیع نرمال با واریانس $\frac{1}{\lambda}$ است و توزیع likelihood یک توزیع نرمال با واریانس یک است.

۳. باتوجه به قسمت قبل داریم:

$$P_N(0,\sigma) = e^{\frac{x^2}{\sigma^2}} = e^{\lambda x^2} \Rightarrow \lambda\sigma^2 = 1 \Rightarrow \sigma^2 = \frac{1}{\lambda}$$

The K-Means Objective Identity

اگر در عبارت اولیه یک μ_i جمع و تفریق کنیم داریم:

$$Z = \sum_{l=1}^{k} \frac{1}{2|C_l|} \sum_{i,j \in C_l} ||x_i - x_j||^2 = \sum_{l=1}^{k} \frac{1}{2|C_l|} \sum_{i,j \in C_l} ||x_i - \mu_l + \mu_l - x_j||^2$$
$$= \sum_{l=1}^{k} \frac{1}{2|C_l|} \sum_{i,j \in C_l} (||x_i - \mu_l||^2 + ||\mu_l - x_j||^2 + 2(x_i - \mu_l)(\mu_l - x_j))$$

واضح است که دوعبارت اول برابر می‌شوند با دوبرابر حاصل جمع حاصل تفاضل مقادیر از میانگین به توان دو پس داریم:

$$Z = \sum_{l=1}^{k} \sum_{i \in C_l} (\frac{|C_l| ||x_i - \mu_l||^2}{|C_l|}) + \sum_{l=1}^{k} \sum_{i,j \in C_l} \frac{2}{|C_l|} (x_i - \mu_l)(\mu_l - x_j)$$
$$= \sum_{l=1}^{k} \sum_{i \in C_l} (||x_i - \mu_l||^2) + \sum_{l=1}^{k} \sum_{i,C_l} \frac{2}{|C_l|} (x_i - \mu_l) \sum_{j \in C_l} (\mu_l - x_j)$$

واضح است که حاصل جمع آخرین سیگما صفر است پس داریم:

$$Z = \sum_{l=1}^{k} \sum_{i \in C_l} (||x_i - \mu_l||^2) + \sum_{l=1}^{k} \sum_{i,C_l} \frac{2}{|C_l|} (x_i - \mu_l) * 0 \sum_{l=1}^{k} \sum_{i \in C_l} (||x_i - \mu_l||^2) + 0$$
$$= \sum_{l=1}^{k} \sum_{i \in C_l} (||x_i - \mu_l||^2)$$

که این همان عبارت سوال است.

Gaussian Mixture Model

4.1 A

جدول ۱: ویژگی‌های انتخاب شده در هر دو روش

Reason	method	Group
این مدل نواحی بیضوی شکل دارد. عملا شبیه کانتورهای دو یا چند توزیع نرمال	GMM	A
تصاویر عملا در گروه‌های دایره شکل قرار گرفته‌اند. این ویژگی knn است زیرا فاصله را از وسط حساب می‌کند و عملا نواحی به شکل دایره در می‌آیند.	KNN	B
در این مدل کمترین فاصله بین هر دو گروه متفاوت بیشینه شده است و این ویژگی این الگوریتم است.	Single Link	C

4.2 B

اول باید likelhood هر نقطه حساب شود. و سپس MLE انجام شود با این تفاوت که وزن هر نقطه برابر با likelihood آن است که توسط جمع likelhood ها نرمالایز شده. با اجرای این کار به نتایج زیر میرسیم: (محاسبات در فایل 2.ipynb هستند.)

$$\mu_1 = \begin{bmatrix} -0.239 \\ 0.965 \end{bmatrix} \mu_2 = \begin{bmatrix} 3.015 \\ -0.343 \end{bmatrix} \Sigma_1 = \begin{bmatrix} 0.439 & -0.414 \\ -0.414 & 0.679 \end{bmatrix} \Sigma_2 = \begin{bmatrix} 0.726 & -0.347 \\ -0.347 & 0.230 \end{bmatrix}$$

اگر مقدار likelhood را برای هر خوشه حساب کنیم داریم:

$$L_1 \approx 0.0063 \quad L_2 \approx 0.0147$$

پس یعنی به خوشه دوم تعلق دارد.

EM

5.1 a

اینجا باید بجای mle عملا map را انجام داد. توزیع بتا توزیع prior برای μ است و توزیع دیریکله توزیع prior برای π است.

5.2 b

در بخش expectation عملا تغییری رخ نمی‌دهد (چون likelihood حساب می‌شود و سعی در تخمین چیزی نیست.) پس داریم:

$$w_{ik} = \frac{\pi_k P(x_i|\mu_k)}{\sum_{j=1}^{K} \pi_j P(x_i|\mu_j)} = \frac{\pi_k \mu_k^{x_i}(1-\mu_k)^{1-x_i}}{\sum_{j=1}^{K} \pi_j \mu_k^{x_i}(1-\mu_k)^{1-x_i}}$$

5.3 c

توزیع بتا به عنوان prior این نقش را عمل می‌کند که به داده‌های مشاهده شده، a موفق و b ناموفق اضافه شود. همچنین فرض توزیع دیریکله برای prior بخش π باعث می‌شود تا بجای محاسبه π فقط براساس وزن، مقدار α_i را به آن اضافه کرد. اگر prior نبود داشتیم:

$$\mu_i = \sum_{j=1}^{n} \frac{w_{ij}}{total\ weight} x_i = \frac{\sum_{j=1}^{n} w_{ij} x_i}{total\ weight}$$

ولی با احتساب داده‌های فرض شده از توزیع بتا داریم:

$$\mu_i = \frac{\sum_{j=1}^{n} w_{ij} x_i + a - 1}{total\ weight + a + b - 2}$$

بطور مشابه برای توزیع دیریکله داریم:

$$\text{no prior} \Rightarrow \pi_i = \frac{\sum_{j=1}^{n} w_{ij}}{\sum_{w=1}^{K} \sum_{j=1}^{n} w_{wj}}$$

$$\text{prior} \Rightarrow \pi_i = \frac{\sum_{j=1}^{n} w_{ij} + \alpha_i}{\sum_{w=1}^{K} (\alpha_w + \sum_{j=1}^{n} w_{wj})}$$