یادگیری ماشین

نیم‌سال اول ۹۹ـ ۱۴۰۰

مدرس: دکتر حسینی

دانشکده‌ی مهندسی کامپیوتر

تمرین سری هفتم

موعد تحویل:۱۶ دی ۹۹

۸۰ نمره نظری + ۳۰ نمره عملی

مسئله‌ی ۱. دست‌گرمی!(۱۰ نمره)

همان‌طور که می‌دانید در مسئله‌ی K-means تابع هزینه‌ی زیر را در نظر می‌گیریم:

$$\mathcal{J} = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}||x_n - \mu_k||^2$$
$$r_{nk} = \mathbf{I}[x_n \in cluster\ k]$$

الف) (۲ نمره) فضای یک‌بعدی را در نظر بگیرید. داده‌های آموزش را $\{1, 2, 8, 9\}$ در نظر بگیرید. مقادیر اولیه‌ی μ_k ها را ۰ و ۳ در نظر بگیرید. (K=۲)

الگوریتم K-means را تا زمان همگرایی اجرا کنید. در هر مرحله تخصیص داده‌ها به خوشه‌ها و μ_k ها را مشخص کنید. مقدار تابع هزینه در زمان همگرایی چقدر است؟

ب) (۸ نمره) کاری که تا اینجا انجام دادیم تخصیص نقاط به خوشه‌ها بود. اما از روش K-means می‌توان به صورت سلسله‌مراتبی هم بهره برد. به این صورت که خوشه‌های شبیه به‌هم را در مرحله‌ی بعد به عنوان یک متاـخوشه[1] در نظر می‌گیریم.

به عوان مثال فرض کنید قرار است داده‌های $\{1, 2, 8, 9, 31, 32, 38, 39\}$ را در چهار خوشه قرار دهیم. پس از اجرای الگوریتم K-means چهار خوشه‌ی $\{38, 39\}, \{31, 32\}, \{8, 9\}, \{1, 2\}$ بدست می‌آید. حال می‌توانیم دو خوشه‌ی $\{8, 9\}, \{1, 2\}$ به عنوان متاـخوشه‌ی شماره یک و دو خوشه‌ی باقی‌مانده را متاخوشه‌ی شماره‌ی دو در نظر بگیریم.

$v_1, ..., v_M$ را مراکز این M متاخوشه‌های جدید در نظر بگیرید.

همچنین $y_{nm} = \mathbf{I}[cluster\ n \in Meta\ cluster\ m]$ در نظر بگیرید.

با این تعریف‌های جدید می‌توانیم تابع هزینه‌ی زیر را برای مسئله‌ی K-means سلسله‌مراتبی معرفی کنیم:

$$\mathcal{J} = \sum_{i=1}^{K}\sum_{m=1}^{M} y_{im}||\mu_i - v_m||^2 + \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}||x_n - \mu_k||^2$$

با در نظر گرفتن این تابع هزینه طبیعی است که از یک الگوریتم مشابه برای بهینه‌سازی استفاده کنیم. در الگوریتم جدید این چهار مرحله به صورت تکراری[2] انجام می‌شوند.

- v, r, μ ثابت هستند. تابع هزینه را بر حسب y بهینه می‌کنیم.
- y, r, μ ثابت هستند. تابع هزینه را بر حسب v بهینه می‌کنیم.
- y, v, μ ثابت هستند. تابع هزینه را بر حسب r بهینه می‌کنیم.

[1] Meta-cluster
[2] Iterative

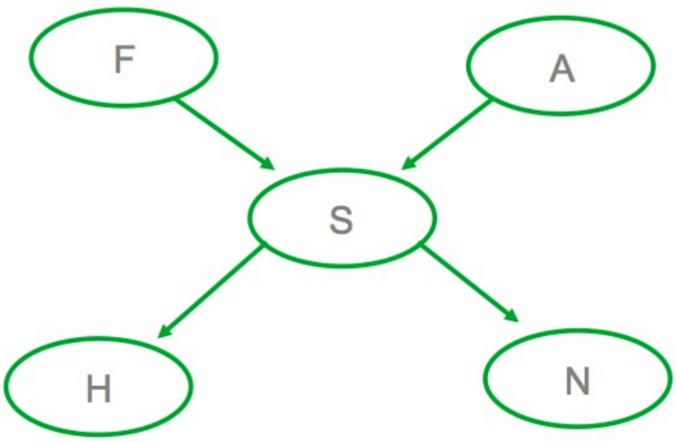
- y, v, r ثابت هستند. تابع هزینه را بر حسب μ بهینه می‌کنیم.

رابطه‌ی بروزرسانی در هر مرحله را پیدا کنید. روابط شما باید شبیه به روابط بدست آمده در مراحل E و M الگوریتم K-means باشد.

مسئله‌ی ۲. EM در شبکه‌های بیزی(۳۰ نمره)

در صورتی که با شبکه‌های بیزی آشنایی ندارید ابتدا این نوشته را مطالعه کنید.

مدل گرافیکی زیر را در نظر بگیرید. در مدل زیر تمامی متغیرها Boolean هستند. در این سوال قرار است از الگوریتم EM برای آموزش این شبکه‌ی بیزی استفاده کنیم. برای این کار دقت کنید که متغیرهای F, S, H و متغیر N کاملا قابل مشاهده[3] و متغیر A گاهی اوقات غیرقابل مشاهده[4] است.



شکل ۱: شبکه‌ی بیزین مربوط به سوال ۲

الف) (۵ نمره) توزیع شرطی مربوط به هر یک از این متغیرهای تصادفی چیست؟

ب) (۵ نمره) در گام E از الگوریتم EM، بر اساس داده‌های آموزش یک توزیع احتمال روی هر یک از متغیرهای تصادفی غیرقابل مشاهده تخمین زده می‌شود. می‌دانیم که متغیر غیرقابل‌مشاهده‌ی سوال A است. بنابراین در گام E باید توزیع $\mathcal{P}(A|F, S, H, N)$ تخمین زده شود. یک عبارت برای محاسبه‌ی $\mathcal{P}(A = 1|F, S, H, N)$ برحسب توزیع‌هایی که در شبکه‌ی بیزی وجود دارد بنویسید.

راهنمایی: از توزیع‌های شرطی موجود در گراف استفاده کنید!

پ) (۵ نمره) متغیرهای تصادفی‌ای که در Markov blanket متغیر تصادفی A هستند را پیدا کنید.[5]

ت) (۵ نمره) فرض کنید تمامی متغیرهایی که در Markov blanket متغیر A پیدا کردید مشاهده شده‌اند. بنابراین باید بتوان توزیع روی A را بوسیله‌ی تنها همین متغیرها محاسبه کرد. عبارتی که برای بخش ب نوشتید را ساده کنید به طوری که در آن تنها از متغیرهای درون Markov blanket متغیر A استفاده شده باشد.

[3] Fully Observed
[4] Unobserved
[5] به طور خلاصه این متغیرها شامل تمامی اطلاعاتی هستند که برای پیداکردن A کافی است. برای اطلاعات بیشتر اینجا را مطالعه کنید!

ث) (۵ نمره) در گام M از الگوریتم EM پارامترهای شبکه با استفاده از متغیرهای مشاهده شده و توزیع تخمین‌زده شده در گام E محاسبه می‌شوند.

در شبکه‌ی سوال برخی پارامترها تنها با استفاده از متغیرهای مشاهده شده‌ی F, S, H, N محاسبه می‌شوند و برخی دیگر با استفاده از توزیع محاسبه شده در گام E پیدا می‌شوند. پارامترهایی که محاسبه‌ی آن‌ها وابسته به محاسبه‌ی گام E است را پیدا کنید و آن‌ها را لیست کنید. برای نشان دادن یک پارامتر توزیع احتمال آن پارامتر را ذکر کنید (مثلا $(\mathcal{P}(N=۱|S=۰)$

ج) (۵ نمره) جدول ۱ را در نظر بگیرید. این جدول لیستی از داده‌های آموزشی قابل مشاهده و توزیع استنتاج شده در گام E را در دور سوم الگوریتم EM نشان می‌دهد.

برای توزیع $\mathcal{P}(A=۱)$ در گام M چه تخمینی داریم؟برای توزیع $\mathcal{P}(S=۱|F=۰, A=۱)$ چطور؟ (در حقیقت باید این تخمین در گام M را حساب کنید.)

جدول ۱: جدول داده‌های آموزشی مربوط به سوال ۲

نمونه	F	S	H	N	A	$\mathcal{P}(A=۱\|F,S,H,N)$
۱.	۰	۱	۰	۱	۱	–
۲.	۰	۰	۱	۰	۱	–
۳.	۱	۰	۱	۱	۰	–
۴.	۰	۰	۰	۱	؟	۰.۸
۵.	۰	۱	۰	۰	؟	۰.۴

مسئله‌ی ۳. EM برای مخلوطی از توزیع پوآسون(۲۰ نمره)

در این تمرین می‌خواهیم الگوریتم EM را بر روی مخلوطی از توزیع‌های پوآسون[6] پیدا کنیم.تابع جرم احتمال برای توزیع پوآسون به صورت زیر است:

$$\mathcal{P}(m;\lambda)=\frac{e^{-\lambda}\lambda^m}{m!}$$
$$m\in\mathbf{N}\cup\{۰\}$$
$$\lambda\in(۰,\infty)$$

بنابراین تابع جرم احتمال توزیع پوآسون به صورت زیر است:

$$\mathcal{P}(m;\pi,\lambda)=\sum_{k=۱}^{K}\pi_k\mathcal{P}(m;\lambda_k)$$

داده‌هایی که در اختیار داریم از این توزیع مخلوط به صورت iid تولید شده‌اند.

$$\mathcal{D}\sim\{m_۱,m_۲,...,m_N\},\ m_n\in\mathbf{N}$$

[6] Mixture of Poisson distributions

الف) (۱۰ نمره) لاگرانژی را برای بیشینه کردن کران پایین با توجه به پارامتر q_k بنویسید. ثابت کنید که مقدار بهینه‌ی پارامتر q_k برابر با توزیع احتمال پسین متغیر z_k است یعنی $q_k = \gamma_k$.

ب) (۱۰ نمره) حال مقدار q_k و γ_k را در کران پایین لگاریتم درست‌نمایی جاگذاری کنید. نشان دهید مقدار بهینه‌ی پارامترهای π_k و λ_k در حالی‌که γ_k ثابت است برابر است با:

$$\pi_k = \frac{\sum_{n=1}^{N} \gamma_{kn}}{N}$$

$$\lambda_k = \frac{\sum_{n=1}^{N} \gamma_{kn} m_n}{\sum_{n=1}^{N} \gamma_{kn}}$$

مسئله‌ی ۴. PCA (۲۰ نمره)

الف) (۵ نمره) در الگوریتم PCA داده‌ها را به صورت خطی روی محورهایی که بردار ویژه‌ی ماتریس کواریانس هستند تصویر می‌کنیم. به این صورت که بیشترین واریانس داده‌های تصویر شده مربوط به محوری است که در راستای بردار ویژه‌ی بزرگترین مقدار ویژه‌ی ماتریس کواریانس است. نشان دهید محور iام پیدا شده در الگوریتم PCA مربوط به iامین بزرگترین مقدار ویژه‌ی ماتریس کواریانس است.

ب) (۵ نمره) نشان دهید استفاده از کرنل خطی $K(x, x') = x^T x'$ معادل با الگوریتم اولیه‌ی PCA است.

پ) (۵ نمره) ماتریس داده‌ها را X می‌نامیم که هر سطر آن مربوط به یک داده است($X_{n\times d}$).
با استفاده از Singular Value Decomposition(SVD) تبدیل PCA را پیدا کنید.
در صورتی که با SVD آشنایی ندارید می‌توانید این نوشته را مطالعه کنید و به صفحه‌ی ویکی‌پدیای آن سری بزنید!

ت) (۵ نمره) در صورتی که ابعاد داده‌های مسئله بیشتر از تعداد داده‌هایی باشد که در اختیار داریم. برای محاسبه‌ی PCA استفاده از SVD روش بهتری است یا محاسبه‌ی بردار ویژه‌های ماتریس کواریانس؟ چرا؟

مسئله‌ی ۵. پیاده‌سازی K-means و GMM (۱۵ نمره)

در این تمرین قرار است الگوریتم‌های K-means و GMM را پیاده‌سازی کنید. سازوکار این تمرین در یک نوت‌بوک همراه با این تمرین منتشر شده است. لطفا کد خود را در داخل نوت‌بوک تکمیل کرده و تکمل شده‌ی نوت‌بوک را به همراه پاسخ سوالات تئوری در یک فایل فشرده آپلود نمایید.

مسئله‌ی ۶. کاهش ابعاد(۱۵ نمره)

هدف این سوال پیاده‌سازی الگوریتم PCA و ارزیابی آن روی یک مجموعه‌ی داده است.مجموعه‌ای تصاویر ۱۰ فرد به شما داده شده که از هر فرد ۹ تصویر در آن موجود است. ۵ تصویر از هر فرد را برای آموزش و تشکیل فضای Face Space و ۴ تصویر بعدی را هم برای آزمایش در نظر بگیرید.

الف) (۳ نمره) ابتدا ماتریس کواریانس را تشکیل دهید. سپس ۵ مقدار ویژه‌ی برتر این ماتریس را پیدا کنید و تصاویر مربوط به این ۵ مقدار ویژه‌ی برتر را نمایش دهید.
راهنمایی ۱: چون بردار ویژه‌ها ابعادی مشابه با داده دارند بنابراین می‌توان آن‌ها را به صورت یک تصویر نمایش داد.
راهنمایی۲: با داشتن بردار ویژه‌ی $A^T A$ می‌توان بردار ویژه‌ی AA^T را بدست آورد و برعکس.

ب) (۵ نمره) در این بخش می‌خواهیم مقایسه‌ای بین داده‌ها در فضای کاهش یافته و فضای اصلی داشته باشیم.

یک) برای این کار ابتدا بهترین بردار ویژه‌ها(Eigen Faceها) را پیدا کنید. توضیح دهید که چگونه آن‌ها را پیدا کردید.

دو) حال داده‌ها را در فضای کاهش‌یافته‌ی ساخته شده با این بردارها تصویر کنید. برای این کار ضرایب بدست آمده از ضرب داخلی داده با Eigen Face را در بردار ویژه ضرب می‌کنیم و آن‌ها را با هم جمع می‌کنیم. (در حقیقت باید داده‌ها را به فضایی با محورهای معادل با Eigen Face تصویر کنید.)

حال تصویر سه فرد مختلف در مجموعه‌ی آموزش را انتخاب کنید و آن‌ها را با بازسازی شده‌شان مقایسه کنید. این کار را برای سه فرد مختلف در مجموعه‌ی آزمایش هم انجام دهید. چه مشاهده می‌کنید؟ علت آن را بیان کنید؟

راهنمایی: دقت کنید که برای بازسازی میانگین را باید به داده‌های بازسازی شده اضافه کنید.

پ) (۳ نمره) تعداد مولفه‌های لازم برای بازسازی تصویر را به گونه‌ای پیدا کنید که تفاوت تصویر اصلی با تصویر بازسازی شده با چشم قابل مشاهده نباشد. (نیازی به استفاده از روابط ریاضی نیست اما می‌توانید از معیار cumulative explained variance ratio استفاده کنید مثلا بیش از ۹۰ درصد واریانس داده‌ها بازسازی شود.)

توضیحات اضافه:

واریانس کل [7] مجموع واریانس همه‌ی مولفه‌های اصلی [8] است.

variance explained برای یک مولفه‌ی اصلی نسبت واریانس آن مولفه به واریانس کل است.

برای چند مولفه‌ی اصلی به راحتی واریانس آن‌ها را با هم جمع و بر واریانس کل تقسیم می‌کنیم.

ت) (۴ نمره) حال داده‌ها را به فضایی با تعداد مولفه‌ای که در بخش قبل بدست آوردید ببرید. سپس توسط ماشین پشتیبان چند کلاسه‌ی خطی آن‌ها را دسته‌بندی کنید. دقت نهایی را روی داده‌های آزمایش گزارش کنید. این کار را برای حالتی که از داده‌های اصلی استفاده می‌کنید هم انجام دهید. آیا نتایج فرق می‌کند؟ چرا؟ این کار کدام مزیت استفاده از PCA را نشان می‌دهد؟

موفق باشید

[7] Total Variance

[8] Principal Components