دانشکده‌ی مهندسی کامپیوتر

یادگیری ماشین

تاریخ تحویل: ۷ خرداد

مدرس: دکتر حسینی

تمرین سری ششم

۱. (خوشه بندی k-means) (۱۵ نمره)

- (۴ نمره) ثابت کنید الگوریتم k-means همگرا می شود.
- (۵ نمره) برای یک دیتاستی با n داده و k خوشه ($k > ۲$)، نصف داده ها در ناحیه متمرکز و نصف دیگر در ناحیه ای با چگالی کمتر قرار گرفته اند و این دو ناحیه تقریبا جدا از هم اند، پس از خوشه بندی دیتاست (با احتساب MSE) آیا در نهایت مرکز خوشه ها به صورت یکنواخت بین دو ناحیه ذکر شده توزیع می شوند؟ یا در ناحیه ای مراکز تجمع بیشتری دارند؟ کدام ناحیه؟
- (۶ نمره) الگوریتم برگرفته از k-means را اینگونه در نظر بگیرید که در اولین مرحله انتخاب مرکز ها، اولین مرکز به صورت تصادفی انتخاب شود ولی مرکز بعدی نقطه ای با بیشترین احتمال بدست آمده از فرمول زیر باشد. مراکز بعدی نیز به همین صورت. در انتخاب مرکز در مراحل بعدی (هنگام بروز کردن مرکز ها) نیز نقطه با بیشترین احتمال فرمول زیر را به عنوان مرکز جدید برمی گزینیم.

$$\frac{D^{۲}(x)}{\sum_{x\in X} D^{۲}(x)}$$

$D(x)$ طول کوتاه ترین فاصله بین داده x تا نزدیکترین مرکزی است که قبلا انتخاب شده است. X نیز مجموعه تمام داده هاست.
نتیجه خوشه بندی این الگوریتم را نسبت به نتیجه k-means مقایسه کنید. از لحاظ سرعت همگرایی نیز مقایسه کنید.

۲. (خوشه بندی EM) (۱۷ نمره)

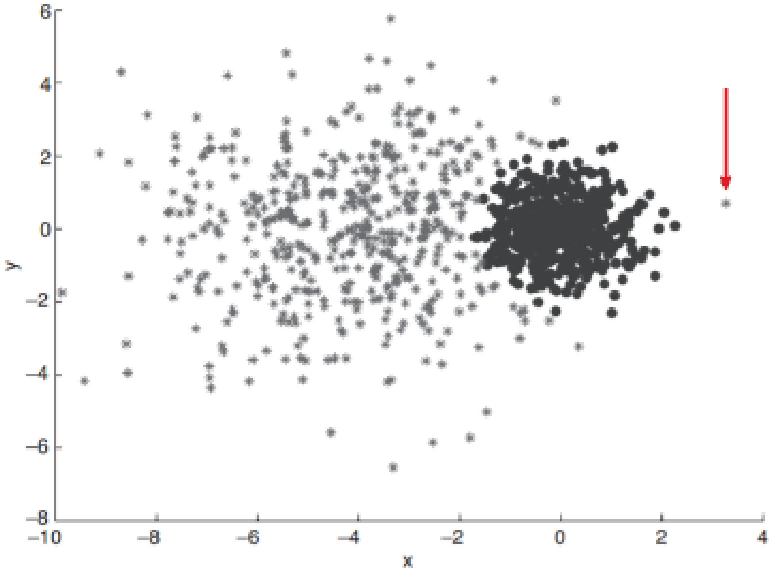
- (۷ نمره) در شکل ۱ که یک دیتاست با دو ویژگی و دو خوشه هست، خوشه سمت چپ گسترده تر و خوشه سمت راست متمرکزتر است. یکی از داده های خوشه سمت چپ که با علامت پیکان نشان داده شده است دورتر و سمت راست خوشه متمرکز قرار گرفته است. توضیح دهید چرا خوشه بندی EM مناسب تر از k-means برای این مورد است؟
- (۱۰ نمره) در این تمرین قصد داریم برای پیدا کردن تخمین بیشینه درست نمایی، الگوریتم EM را روی مخلوطی از توزیع های پواسن [۱] اعمال کنیم. توزیع پواسن احتمال رخ دادن m اتفاق در یک بازه زمانی هنگامی که نرخ میانگین رخ دادن λ است را مدل می کند. تابع جرم احتمال برای یک توزیع پواسن به صورت زیر است.

$$P(m,\lambda) = \frac{e^{-\lambda}\lambda^m}{m!}$$

بنابراین تابع جرم احتمال یک مخلوط از توزیع های پواسن به صورت زیر خواهد بود.

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

[۱] mixture of Poisson distributions



شکل ۱: مسئله ۲ مورد اول

در طول حل سوال فرض کنید مجموعه داده های ما iid [۲] هستند.
لاگرانژی را برای بیشینه کردن کران پایین با توجه به پارامتر q_k بنویسید. ثابت کنید مقدار بهینه پارامتر $q_k = \gamma_k$ برابر با احتمال پسین متغییر z_k است.
مقدار q_k را با γ_k در کران پایین لگاریتم درست نمایی جایگذاری کنید و نشان دهید مقادیر بهینه λ_k و π_k در حالی که γ_k ثابت است برابر $\frac{\sum_{n=1}^{N} \gamma_n m_n}{\sum_{n=1}^{N} \gamma_n}$ و $\frac{\sum_{n=1}^{N} \gamma_n}{N}$ می باشد.

۳. (خوشه بندی EM) (۱۲ نمره)

گام های بروزرسانی E و M را برای تخمین پارامتر های توزیع مخلوطی زیر بدست آورید:

$$p(x) = \sum_{k=1}^{K} \lambda_k p_k(x)$$

به گونه ای که اجزای مخلوطی $p_k(x)$، توزیع رایلی [۳] (این لینک) با پارامتر θ_k هستند.

۴. (بخش عملی) (۲۵ نمره)

در این سوال قصد داریم دو الگوریتم k-means و GMM را پیاده سازی کنیم. برای پیاده سازی این دو الگوریتم از کتابخانه های پایتون که این دو الگوریتم در آن ها پیاده سازی شده اند استفاده نکنید. توضیحات تکمیلی در فایل ژوپیتر آمده است.

توجه: برای مطالعه مبحث ضرایب لاگرانژ می‌توانید این لینک و یا از صفحه ۷۰۷ این لینک را مطالعه کنید.

[۲] Independent and identically distributed
[۳] Rayleigh distribution