دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

یادگیری ماشین برای بیوانفورماتیک

بهار ۱۴۰۲

استاد:علی شریفی زارچی

کوییز ۱ | تاریخ برگزاری: ۲۷ فروردین

سوالات (۱۰۰ نمره)

۱. (۵۰ نمره) پاسخ کوتاه

به سوالات زیر به صورت کوتاه پاسخ دهید:

- خوشه‌بندی (clustering) در الگوریتم GMM به صورت soft انجام می‌شود یا hard ؟
- آیا الگوریتم K-Means به مقداردهی اولیه مرکز خوشه‌ها حساس است؟ آیا این الگوریتم به صورت تضمینی همگرا می‌شود؟
- شما در حال طراحی یک مدل برای یک تسک طبقه‌بندی (classification) هستید. در ابتدا مدل خود را بر روی ۱۰۰ نمونه آموزش می‌دهید و مشاهده می‌کنید که با وجود همگرا شدن آموزش، خطای آموزش بر روی این نمونه‌ها زیاد است. پس در ادامه تصمیم می‌گیرید که شبکه خود را این‌بار روی ۱۰۰۰۰ نمونه آموزش دهید. آیا روش شما برای حل این مشکل صحیح است؟ اگر بلی، محتمل‌ترین نتایج مدل خود را در این حالت توضیح دهید. اگر خیر، راه‌حلی برای رفع این مشکل بیان کنید.
- هر چه بردارهای ویژه‌ای از ماتریس کواریانس که برای کاهش ابعاد از طریق PCA استفاده می‌کنیم دارای مقدار ویژه‌ی بزرگتری باشند، خطای بازسازی کمتر می‌شود. دلیل این موضوع را به صورت خلاصه توضیح دهید.
- خطای روی داده‌های آموزش و تست را در دو حالت overfitting و underfitting مقایسه کنید.

۲. (۷۰ نمره) رگرسیون خطی، تخمین ML و تخمین MAP

همانطور که از درس می‌دانید، در یک مدل رگرسیون خطی با ویژگی‌های x_i داریم:

$$y = \sum_{i=1}^{p} w_i x_i + \epsilon = w^T x + \epsilon$$

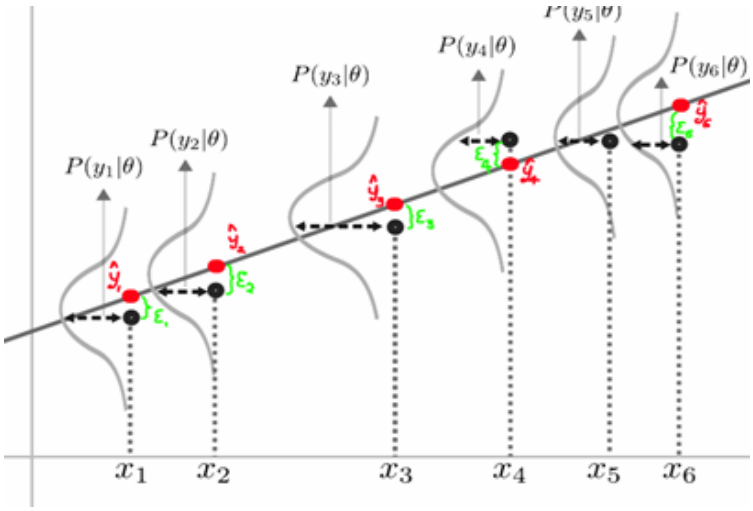
در صورتی که نویز موجود دارای توزیع $\epsilon \sim \mathcal{N}(0, \sigma^2)$ باشد، مشخصا خواهیم داشت:

$$y|x, w \sim \mathcal{N}(w^T x, \sigma^2)$$

با در نظر گرفتن تمام نمونه‌های آموزشی می‌توان این عبارت را برای همه آن‌ها بنویسیم و در نتیجه به صورت برداری خواهیم داشت:

$$Y|X, w \sim \mathcal{N}(Xw, \sigma^2 I_n)$$

که در عبارت بالا $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$ و $w \in \mathbb{R}^p$ می‌باشد.
الف) توزیع بالا به چه معناست؟ برای راهنمایی می‌توانید از شکل زیر کمک بگیرید.



نکته: در ۳ بخش بعدی جواب خود را به صورت یک مسئله بهینه‌سازی کمترین مربعات (که می‌تواند همراه با یک جمله regularizer باشد) بنویسید و نیازی به محاسبه $\hat{w}_{MAP}$ و $\hat{w}_{ML}$ نیست.

ب) تخمین ML را برای w بدست بیاورید.
این مسئله معادل با کدام حالت روش رگرسیون است؟

پ) فرض کنید برای پارامترهای w یک توزیع اولیه (Prior) در نظر می‌گیریم؛ به طوریکه $w \sim \mathcal{N}(۰, \lambda^۲ I_p)$.
تخمین MAP را برای w بدست آورید.
این مسئله معادل با کدام حالت روش رگرسیون است؟

ج) حال توزیع اولیه را تغییر می‌دهیم. فرض کنید که هر یک از وزن‌ها دارای توزیع $w_i \sim Laplace(۰, \lambda)$ باشند. تخمین MAP را برای w بدست آورید.
این مسئله معادل با کدام حالت روش رگرسیون است؟

د) تفاوت بین استفاده از این دو توزیع را از دیدگاه اثر آن‌ها بر روی اندازه w_i ها به صورت خلاصه توضیح دهید.

راهنمایی:

$$Z \sim Laplace(۰, \lambda) \rightarrow f_Z(z) = \frac{۱}{۲\lambda} exp(-\frac{|z|}{\lambda})$$

$$Z \sim \mathcal{N}(\mu, \Sigma) \rightarrow f_Z(\mathbf{z}) = \frac{۱}{(۲\pi)^{n/۲}|\Sigma|^{۱/۲}} exp(-\frac{۱}{۲}(\mathbf{z}-\mu)^T \Sigma^{-۱}(\mathbf{z}-\mu))$$