تمرین سری ۲ - یادگیری ماشین برای بیوانفورماتیک

مهدی کاظمی ۹۹۲۱۰۷۵۳

1

۱.۱ - الف)

$P(C_1) = \pi \,,\quad P(C_2) = 1-\pi$

$P(x|C_1) = \mathcal{N}(\mu_1, \Sigma_1) \,,\quad P(x|C_2) = \mathcal{N}(\mu_2, \Sigma_2)$

Decision Boundary $\Rightarrow P(C_1|x) = P(C_2|x) \Rightarrow \dfrac{P(x|C_1)P(C_1)}{P(x)} = \dfrac{P(x|C_2)P(C_2)}{P(x)}$

$\Rightarrow P(x|C_1)P(C_1) = P(x|C_2)P(C_2)$

$\Rightarrow \ln[P(x|C_1)P(C_1)] = \ln[P(x|C_2)P(C_2)] = \ln P(x|C_1) + \ln P(C_1) = \ln P(x|C_2) + \ln P(C_2)$

$$\Rightarrow -\frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_1| - \frac{1}{2}(x-\mu_1)^T \Sigma_1^{-1}(x-\mu_1) + \ln\pi$$

$$= -\frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_2| - \frac{1}{2}(x-\mu_2)^T \Sigma_2^{-1}(x-\mu_2) + \ln(1-\pi)$$

$$\Rightarrow \ln|\Sigma_1| + (x-\mu_1)^T \Sigma_1^{-1}(x-\mu_1) - 2\ln(\pi) = \ln|\Sigma_2| + (x-\mu_2)^T \Sigma_2^{-1}(x-\mu_2) - 2\ln(1$$

$$\Rightarrow \ln|\Sigma_1| + x^T \Sigma_1^{-1} x + \mu_1^T \Sigma_1^{-1}\mu_1 - 2\mu_1^T \Sigma_1^{-1} x - 2\ln(\pi)$$

$$= \ln|\Sigma_2| + x^T \Sigma_2^{-1} x + \mu_2^T \Sigma_2^{-1}\mu_2 - 2\mu_2^T \Sigma_2^{-1} x - 2\ln(1-\pi)$$

۱.۱ - ب) $\Sigma_1 = \Sigma_2 = \Sigma$

$$\Rightarrow \ln|\Sigma| + x^T \Sigma^{-1} x + \mu_1^T \Sigma^{-1}\mu_1 - 2\mu_1^T \Sigma^{-1} x - 2\ln(\pi)$$

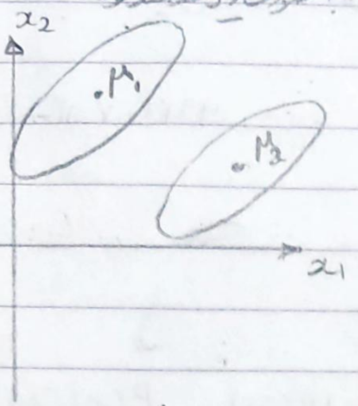
$$= \ln|\Sigma| + x^T \Sigma^{-1} x + \mu_2^T \Sigma^{-1}\mu_2 - 2\mu_2^T \Sigma^{-1} x - 2\ln(1-\pi)$$

$$\Rightarrow \underbrace{[-2(\mu_1^T - \mu_2^T)\Sigma^{-1}]}_{A}\, x + \underbrace{\mu_1^T \Sigma^{-1}\mu_1 - \mu_2^T \Sigma^{-1}\mu_2 - 2\ln(\pi) + 2\ln(1-\pi)}_{b} = 0$$

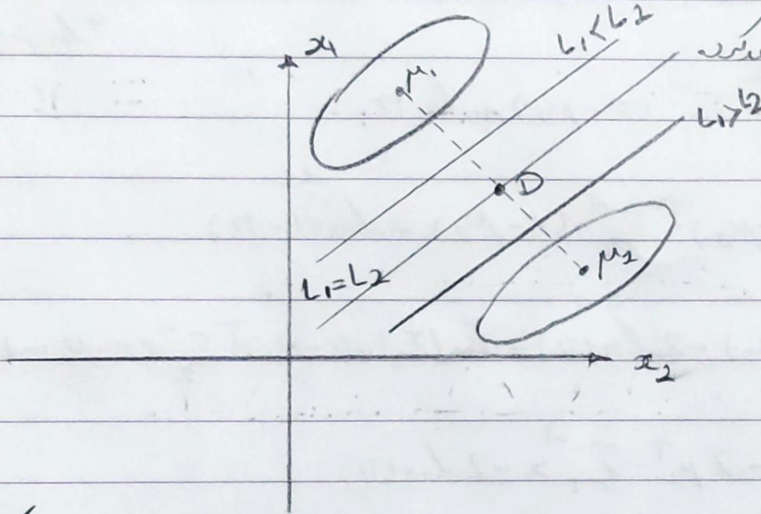
* فرم به صورت ابرصفحه درآمد.

s.a.m

ج) اگر هر دو توزیع دارای ماتریس کوواریانس یکسانی باشند، خط آن‌ها به صورت زیر خواهد بود:



حال اگر به ازای misclassification بخواهیم جریمه‌ای برابر l_1 و l_2 در نظر بگیریم به ازای ۳ حالتی که l_1 و l_2 می‌توانند نسبت به هم داشته باشند، مرز تصمیم به صورت زیر تغییر خواهد کرد.



$\mu_1 D = \mu_2 D$

اگر خطای misclassification به ازای هر ۲ کلاس برابر باشد ($l_1 = l_2$) مرز دسته‌بندی دقیقاً بین دو توزیع می‌افتد زیرا که به ازای خطای دسته‌بندی هر ۲ کلاس به یک میزان جریمه می‌شویم.

اگر $l_2 < l_1$ باشد، به این معنی است که اگر داده کلاس اول به کلاس دوم نسبت داده شود خطای بزرگتری مرتکب شده‌ایم.

بنابراین باید سعی کنیم این خطا کمتر رخ بدهد پس باید مرز را به توزیع با میانگین μ_2 نزدیک کنیم که کمتر به اندازه l_1 جریمه شویم.

اگر $L_2 > L_1$ باشد با استدلالی مشابه مرحله قبل باید مرز را به توزیع با میانگین μ_1 نزدیک کنیم.

در حالت ج) چون ما در صورتی مرز دقیقاً بین توزیع ها می افتد که احتمال π برابر با 0.5 باشد و در حالت کلی اگر که مقدار $(1-\pi)\times L_2$ بیشتر از $\pi \times L_1$ باشد، داده را به دسته C_1 اختصاص میدهیم زیرا که احتمال خطای احتمالی برای خروجی C_1 کمتر از C_2 است.

د) $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

$$\underbrace{P(X|C_1)}_{N(\mu_1, \Sigma)} P(C_1) \ ? \ \underbrace{P(X|C_2)}_{N(\mu_2, \Sigma)} P(C_2)$$

$$\frac{1}{\sqrt{(2\pi)^2|\Sigma|}} \exp\left(-\frac{(x-\mu_1)^T \Sigma^{-1}(x-\mu_1)}{2}\right)\pi \ ? \ \frac{1}{\sqrt{(2\pi)^2|\Sigma|}} \exp\left(-\frac{(x-\mu_2)\Sigma^{-1}(x-\mu_2)}{2}\right)(1-\pi)$$

$$\exp\left(-\frac{(x-\mu_1)^T \sigma^2 I (x-\mu_1)}{2}\right)\pi \ ? \ \exp\left(-\frac{(x-\mu_2)^T \sigma^2 I (x-\mu_2)}{2}\right)(1-\pi)$$

$$\exp\left(-\frac{\sigma^2 (x-\mu_1)^T(x-\mu_1)}{2}\right)\pi \ ? \ \exp\left(-\frac{\sigma^2 (x-\mu_2)^T(x-\mu_2)}{2}\right)(1-\pi)$$

$$\exp\left(\frac{-\sigma^2 [\text{distance between } x \text{ and } \mu_1]^2}{2}\right)\pi \ ? \ \exp\left(-\frac{\sigma^2 [\text{dist between } x \text{ and } \mu_2]^2}{2}\right)(1-\pi)$$

if $\pi = 0.5$ and $dist(x, \mu_1) < dist(x, \mu_2) \Rightarrow P(C_1|X) > P(C_2|X)$

s.a.m

$D = \{(x^i, y^i)\}_{i=1}^{N}$ $\qquad y = \{0, 1\}$

if $c = 1 \Rightarrow y = 1$ and if $c = 2 \Rightarrow y = 0$

$N1 = \sum_{i=1}^{N} y^i \Rightarrow \pi = \frac{N1}{N}$

$X = \begin{bmatrix} x_1^1 & x_1^2 & \cdots & x_1^N \\ x_2^1 & x_2^2 & \cdots & x_2^N \end{bmatrix}$

dimension $= d = 2$

log likelihood $= l(\mu, \Sigma | X) = \ln \prod_{i=1}^{N} f_{x^i}(x^i | \mu, \Sigma)$

$= \ln \prod_{i=1}^{N} \frac{1}{\sqrt{(2\pi)^2 |\Sigma|}} \exp\left(-\frac{(x^i-\mu)^T \Sigma^{-1} (x^i-\mu)}{2}\right)$

$= \sum_{i=1}^{N} \left(-\ln(2\pi) - \frac{1}{2}\ln|\Sigma| - \frac{1}{2}(x^i-\mu)^T \Sigma^{-1}(x^i-\mu)\right)$

$= -N\ln(2\pi) - \frac{N}{2}\ln|\Sigma| - \frac{1}{2}\sum_{i=1}^{N}(x^i-\mu)^T \Sigma^{-1}(x^i-\mu) = l(\mu, \Sigma | x^i)$

μ:

$\frac{\partial l(\mu, \Sigma | x^i)}{\partial \mu} = +\frac{1}{2}\sum_{i=1}^{N} 2 \times \Sigma^{-1}(\mu - x^i) \Rightarrow \sum_{i=1}^{N} \Sigma^{-1}(\mu - x^i) = 0$

$\Rightarrow 0 = N\mu - \sum_{i=1}^{N} x^i \Rightarrow \hat{\mu} = \frac{\sum_{i=1}^{N} x^i}{N}$

$\mu_1 = \frac{\sum_{i=1}^{N} y^i x^i}{N_1}$, $\mu_2 = \frac{\sum_{i=1}^{N} (1-y^i) x^i}{N_2}$ $\quad (N_2 = N - N_1)$

s.a.m

$$l(\mu,\Sigma|x^i) = -N\ln(2\pi) - \frac{N}{2}\ln|\Sigma| - \frac{1}{2}\sum_{i=1}^{N}(x^i-\mu)^T\Sigma^{-1}(x^i-\mu)$$

$$= -N\ln(2\pi) + \frac{N}{2}\ln|\Sigma^{-1}| - \frac{1}{2}\sum_{i=1}^{N} tr[\Sigma^{-1}(x^i-\mu)(x^i-\mu)^T]$$

$$\frac{\partial l(\mu,\Sigma|x^i)}{\partial\Sigma^{-1}} = \frac{N}{2}\Sigma^T - \frac{1}{2}\sum_{i=1}^{N}((x^i-\mu)(x^i-\mu)^T)^T$$

$$= \frac{N}{2}\Sigma - \frac{1}{2}\sum_{i=1}^{N}(x^i-\mu)(x^i-\mu)^T \Rightarrow \frac{\partial l(\mu,\Sigma|x^i)}{\partial\Sigma^{-1}} = 0$$

$$\Rightarrow 0 = N\Sigma - \sum_{i=1}^{N}(x^i-\hat{\mu})(x^i-\hat{\mu})^T \Rightarrow \Sigma = \frac{\sum_{i=1}^{N}(x^i-\hat{\mu})(x^i-\hat{\mu})^T}{N}$$

* در قسمت الف، معادله‌ای که برای مرز تقسیم به دست می‌آید یک معادله درجه دوم بر حسب x است و باعث می‌شود که یک Decision Boundary از مرتبه ۲ داشته باشیم ولی اگر مانند قسمت ب، ماتریس‌های کوواریانس یکسان شود بخش درجه دوم معادله حذف می‌شود و مرز تقسیم به صورت خطی و یک ابر صفحه درخواهد آمد.

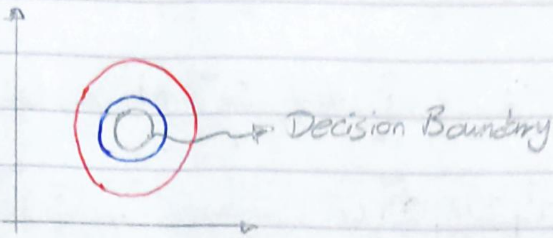
۱-۲)

الف) در حالت گسسته فرض استقلال شرطی برقرار نیست زیرا با دانستن مقدار ویژگی x_2 احتمال ویژگی ۱ تحت تاثیر قرار می‌گیرد: $P(x_1|x_2=0) \neq P(x_1)$

ب) در این حالت کانتورها به شکل دایره درآمده‌اند که نشان دهنده ماتریس کوواریانس به صورت $\Sigma = \sigma I$ است و بنابراین ویژگی‌ها دارای استقلال شرطی هستند و برای پیدا کردن مرز با توجه به اینکه یکی از کلاس‌ها داخل دیگری قرار گرفته و کانتورهای با احتمال 0.5 مشخص شده از مرز تقسیم باید درداخل دایره دسته ۱ (رنگ آبی)

s.a.m

قرار بگیرد به صورت روبرو:



ج) در این شکل کلاسترها به صورت بیضی هایی هستند که بردارهای اصلی آن‌ها موازی محورهای مختصات نیست بنابراین ماتریس کوواریانس نیز به صورت قطری نبوده است و استقلال شرطی برقرار نیست.

۳-۱) در مدل های regression اگر فیچرها را بتوان از روی یکدیگر به دست آورد حالتی ایجاد می‌شود که به آن multicollinearity گفته می‌شود و نشان از داشتن correlation بالا بین فیچرها است، این پدیده باعث می‌شود در مدل logistic regression، ۲ مشکل پیش بیاید، یکی اینکه بردار وزن های مدل بسیار غیرپایدار می‌شوند و با تغییرات کوچک در داده‌ها، وزن ها به صورت ناگهانی تغییر می‌کنند و مشکل دوم این است که با داشتن فیچرهای correlated، نمی‌توان اهمیت فیچرها را به صورت مستقیم از بردار وزن ها (ضرایب) متوجه شد؛ اما در این مسئله با توجه به اینکه گفته شده است فقط یک فیچر داریم و آن فیچر خودش تکرار شده است به نظر می‌رسد که تنها مشکل اول در اینجا ایجاد می‌شود و بردار وزن ها با تغییرات کوچک فیچرها به صورت ناگهانی به مقدار زیادی تغییر می‌کنند و ناپایدار می‌شوند.

s.a.m

۲

الف)

• در SVM های حاشیه نرم؛ نادرست - در SVM با حاشیه نرم، پارامتر C با مقدار کوچک اجازه می‌دهد محدودیت حاشیه راحت‌تر نادیده گرفته شود و حاشیه بزرگتر می‌شود و مقدار بالای C باعث می‌شود که حاشیه باریک شود به طوری که مقدار C برابر با بی‌نهایت منجر به SVM با حاشیه سخت می‌شود.

• هسته (Kernel) گاوسی؛ نادرست - در هسته گاوسی هر چه مقدار σ کوچک‌تر باشد باعث می‌شود که حاشیه‌ها کوچک‌تر شوند و در واقع به داده‌های آموزشی می‌چسبند و باعث overfitting می‌شوند و مقدار بزرگ‌تر برای σ باعث حاشیه بزرگ‌تر می‌شود که می‌تواند باعث شود که مدل خیلی ساده شود و در واقع آموزش صورت نگیرد؛ بنابراین این پارامتر باید توسط تکنیک‌هایی مانند cross validation بهینه شود تا مدل بهترین عملکرد را داشته باشد.

ب)

ج)

A: Linear - مرز ایجاد شده صرفاً یک خط است.

B: RBF, σ=2 - مرزی که ایجاد شده، دیاگرام یک SVM با کرنل گاوسی است و با مقایسه با سایر کرنل‌های گاوسی، این دیاگرام حاشیه بزرگتری دارد بنابراین σ آن از همه بزرگتر است.

C: Second Order Polynomial - مرز ایجاد شده به صورت یک چندجمله‌ای درجه دوم است.

D: RBF, σ=0.08 - دیاگرام نشان دهنده پیچیده‌ترین مدل است و به داده‌ها بسیار fit شده است بنابراین مربوط به یک کرنل گاوسی با کوچکترین پارامتر σ است.

E: RBF, σ=0.5 - دیاگرام نشان دهنده کرنل گاوسی است که در مقایسه با ۲ دیاگرام گاوسی دیگر مشخص می‌شود که دارای σ ای بین پارامتر σ دو مورد دیگر است.

(۳) برای اینکه اثبات کنیم یک کرنل valid است کافی است که آن را به صورت یک ضرب داخلی بنویسیم. برای

اثبات valid بودن این کرنل که برای رشته‌ها اعمال می‌شود به صورت زیر عمل می‌کنیم:

۱) در ابتدا یک ماتریس Φ به صورت زیر برای دو رشته فرضی s_1 و s_2 می‌سازیم به طوری که هر سطر ماتریس یکی از رشته‌ها

است و ستون‌های ماتریس، تمام زیررشته‌های موجود رشته‌ها هستند و اگر زیررشته در رشته موجود باشد درایه ۱ می‌گذاریم و

اگر نباشد درایه صفر می‌گذاریم. $\Sigma = \{a, u, e\}$ $\quad s_1 = aue \quad s_2 = auue$

زیررشته‌های مشترک $= \{a, u, e, au, ue\} = A \Rightarrow |A| = 5$

$\Phi =$

	a	u	e	au	ue	uu	auu	uue	aue	auue
aue	1	1	1	1	1	0	0	0	1	0
auue	1	1	1	1	1	1	1	1	0	1

حال برای داشتن خروجی تعداد زیررشته‌های مشترک کافی است که ماتریس Φ را در Φ^T ضرب کنیم که در اینصورت

درایه‌های $K(s_1, s_2)$ و $K(s_2, s_1)$ در ماتریس حاصل برابر ۵ خواهند بود که همان تعداد رشته‌های مشترک و یا خروجی

کرنل مورد نظر است. بنابراین نشان دادیم که کرنل را می‌توان به صورت ضرب داخلی نوشت. بنابراین این کرنل یک کرنل

valid است. (می‌توانستیم به جای ضرب کل ماتریس Φ، فقط سطر اول را در سطر دوم ضرب داخلی کنیم و

خروجی باز هم همان عدد 5 خواهد بود.)