به نام خداوند

تئوری یادگیری ماشین

دکتر سیدصالحی

جلسه ششم

دانشکده ریاضی و علوم کامپیوتر

فروردین ۱۴۰۳

Gaussian discriminant analysis

مثالی که جلسه قبل بررسی کردیم یک مثال از یک دسته از مدل ها بود به نام GDA. تو GDA ما فرضی که میکنیم که توضیع های $conditional$ ما یه سری توضیع نرمال هستند و اینکه $prior$ روی کلاس های ما برنولی هست.(فرض ما حالت دوکلاس مساله است.)

$$C_k \sim Bernouli(\phi) = \phi^y (1-\phi)^{(1-y)}$$

$$p(\boldsymbol{x}|y=k) \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\ = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\}$$

$$P(X=x|Y=1) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma_1|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu_1)^T \Sigma_1^{-1}(x-\mu_1)\right)$$

$$P(X=x|Y=0) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma_0|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu_0)^T \Sigma_0^{-1}(x-\mu_0)\right)$$

ما تا الآن $prior$ کلاس ها را به صورت فرض شده در نظر گرفتیم ولی الآن کمی مدل بهتری میتونیم اتخاذ کنیم. یک تابع $likelihood$ تشکیل میدیم با فرض iid بودن سمپل ها:

$$\begin{aligned} \ell(\phi, \mu_0, \mu_1, \Sigma_1, \Sigma_2) &= \log \prod_{i=1}^{N} p(x^{(i)}, y^{(i)}; \phi, \mu_0, \mu_1, \Sigma_0, \Sigma_1) \\ &= \log \prod_{i=1}^{N} p(x^{(i)}|y^{(i)}, \mu_0, \mu_1, \Sigma_0, \Sigma_1) p(y^{(i)}; \phi) \end{aligned}$$

با ماکزیمم کردن ℓ با توجه به پارامترها، تخمین $likelihood$ پارامترها را پیدا می کنیم.

- برای به حداکثر رساندن $\ell(\cdot)$ با توجه به پارامترها، مشتق $\ell(\cdot)$ را می‌گیریم، مشتق را برابر ۰ قرار می‌دهیم و سپس مقادیر پارامترهایی که عبارت $\ell(\cdot)$ را به حداکثر می‌رساند را حل می‌کنیم. این تخمین حداکثر احتمال ϕ را به دست می‌دهد.

توضیح شهودی در مورد مقدار ϕ برای به حداکثر رساندن $Likelihood$ به شرح زیر است. به یاد بیاورید که ϕ تخمین احتمال y برابر با ۱ است. در مثال خاص ما، این شانس که بیمار بعدی با تومور بدخیم به مطب پزشک شما می‌رود با ϕ نشان داده می‌شود. احتمال شیر آمدن در پرتاب سکه، کسری از پرتاب های دیده شده است. به همین ترتیب، برآورد حداکثر احتمال برای ϕ فقط کسری از نمونه های آموزشی شما با برچسب $y = 1$ است.

$$
\begin{aligned}
\phi &= \frac{1}{n}\sum_{i=1}^{n} 1\{y^{(i)} = 1\} \\
\mu_0 &= \frac{\sum_{i=1}^{n} 1\{y^{(i)} = 0\}x^{(i)}}{\sum_{i=1}^{n} 1\{y^{(i)} = 0\}} \\
\mu_1 &= \frac{\sum_{i=1}^{n} 1\{y^{(i)} = 1\}x^{(i)}}{\sum_{i=1}^{n} 1\{y^{(i)} = 1\}}
\end{aligned}
$$

- برای ایجاد شهود پیرامون مقدار μ_0 و μ_1 که مقدار فرمول را به حداکثر می‌رساند، به این فکر کنید که تخمین حداکثر $likelihood$ میانگین همه ویژگی‌ها برای یک کلاس خاص، مثلاً تومورهای خوش‌خیم، که با برچسب کلاس 0 در مجموعه داده ما نشان داده می‌شود، چقدر خواهد بود.

μ_0 را در نظر بگیرید. یک راه معقول برای تخمین μ_0 این است که تمام تومورهای خوش خیم $training\ set$ خود را در نظر بگیرید (همه نمونه های منفی، یعنی ورودی ها به صورت ۰) و فقط میانگین ویژگی های آنها را در نظر بگیرید. معادله فوق راهی برای نوشتن این شهود است. صورت کسر مجموع بردارهای ویژگی برای همه نمونه های تومورهای خوش خیم در $training\ set$ است (یعنی نمونه هایی با $y = 0$) در حالی که مخرج به سادگی تعداد تومورهای خوش خیم در مجموعه آموزشی است.

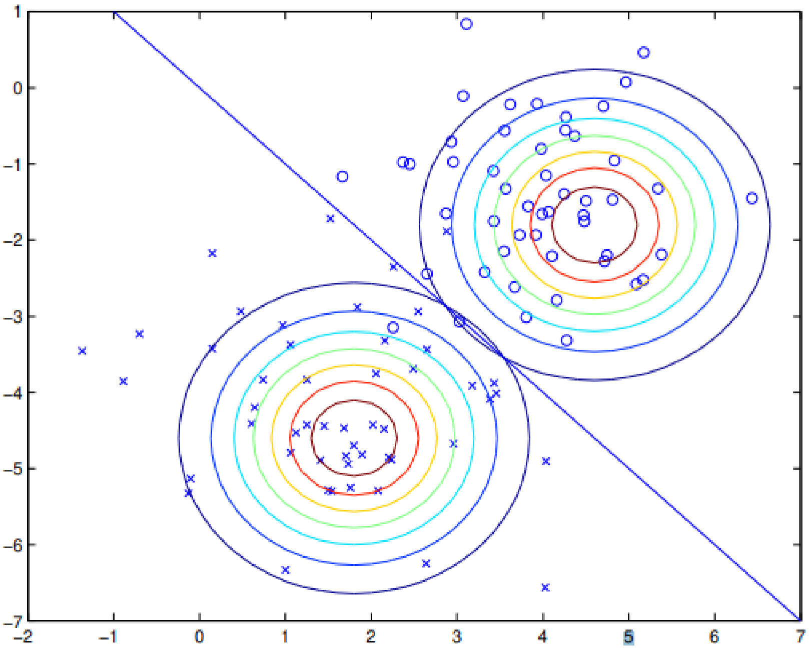
البته بدیهتا وقتی y را میدانیم(از $observation$) با توجه به ستینگ احتمالاتی ما دیگه پارامتر های دیگر توضیع رو نیاز نداریم و توضیع پارامتر $class\ conditional$ هم وجودش معنایی ندارد.

مورد خاص: $\Sigma_k = \sigma^2 I$

شکل بالا$training\ set$، و همچنین خطوط دو توزیع گاوسی که با داده های هر یک از دو کلاس مطابقت دارند را نشان میدهد. توجه داشته باشید که دو توضیع گاوسی ماتریس کوواریانس Σ مشترک دارند ولی μ

های متفاوت.

همچنین فرض داریم که $p(y=1|x) = 0.5$

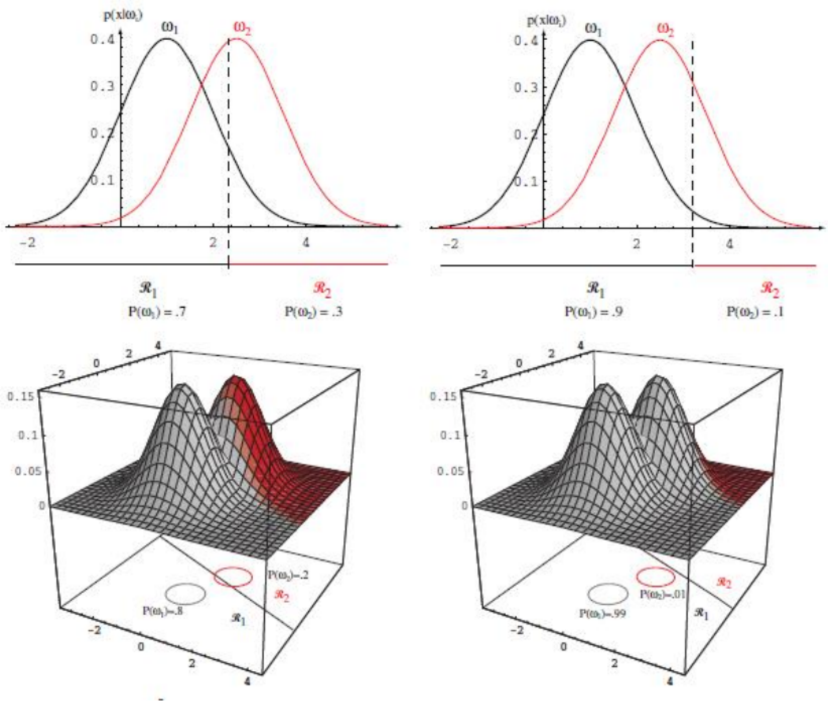


شکل ۱: حالت خاص Σ برابر

در این حالت خاص مرز تصمیم معمولاً در نقطه میانی بین دو مرکز کلاس قرار دارد، که منعکس کننده احتمال برابر برای هر کلاس است. این نشان دهنده نقطه ای است که هر دو کلاس به یک اندازه محتمل هستند و به عنوان جداکننده بین آنها عمل می کند و در نهایت یک تابع درجه ۱ است.

به طور ساده میشود گفت که چون $prior$ ها مساوی اند $posterior$ برابر $likelihood$ است و دلیل وسط قرار گرفتن خط همین بود که در آن وسط $posterior$ ها هم همدیگر را قطع میکردند.

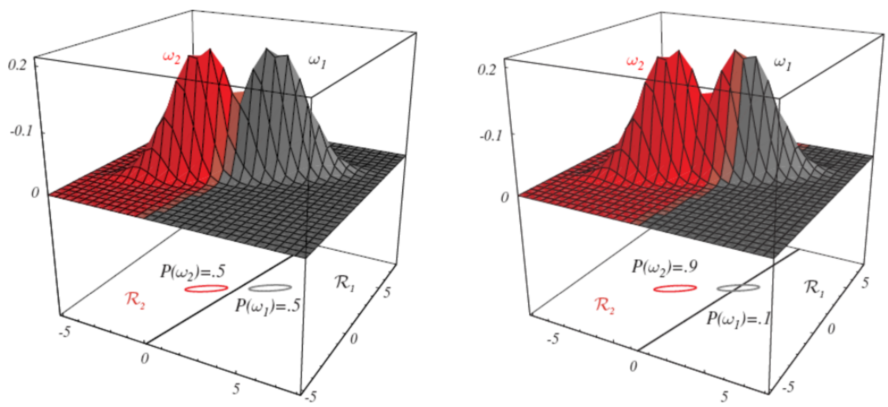
حال فرض کنیم که $prior$ کلاس سیاه بیشتر هست در این صورت مرز به سمت قرمز میرود. در واقع تمایل به قرار گیری در کلاس سیاه بیشتر میشود به دلیل $prior$.



شکل ۲: حالت خاص Σ برابر

مورد خاص: $\Sigma_k = \Sigma$

در این حالت برای توزیع های گاوسی ای را داریم با ماتریس کوواریانس های مساوی ولی دلخواه. ابرصفحه های تصمیم ما در این حالت خاص GDA معادل رویکرد LDA است. و اینکه ابرصفحه های تصمیم نباید بر خطی که میانه کلاس ها را به هم وصل می کند عمود باشند.



شکل ۳: حالت خاص $\Sigma_k = \Sigma$

$Nave\ Bayes\ classifier$

در GDA، بردارهای ویژگی x بردارهای پیوسته و $real - valued$ بودند. بیایید اکنون در مورد یک الگوریتم یادگیری متفاوت صحبت کنیم که در آن x_j ها گسسته ارزیابی می شوند. برای مثال، ساخت فیلتر $spam$ ایمیل را با استفاده از ML در نظر بگیرید. در اینجا، می‌خواهیم پیام‌ها را بر اساس ایمیل‌های تجاری ناخواسته ($spam$) یا $non - spam$ طبقه‌بندی کنیم. پس از یادگیری انجام این کار، می‌توانیم ایمیل‌خوان خود را به‌طور خودکار پیام‌های $spam$ را فیلتر کرده و شاید آنها را در یک پوشه ایمیل جداگانه قرار دهیم. طبقه بندی ایمیل ها نمونه ای از مجموعه وسیع تری از مشکلات به نام طبقه بندی متن است.

فرض کنید یک $training\ set$ داریم (مجموعه ای از ایمیل ها که به عنوان $spam$ یا $non - spam$ برچسب گذاری شده اند). ما ساخت فیلتر $spam$ خود را با مشخص کردن ویژگی‌هایی که x_j برای نمایش ایمیل استفاده می‌شود، آغاز می‌کنیم. ما یک ایمیل را از طریق یک بردار ویژگی نشان خواهیم داد که طول آن برابر با تعداد کلمات در فرهنگ لغت است. به طور خاص، اگر یک ایمیل حاوی کلمه j-ام فرهنگ لغت باشد، $x_j = ۱$ را تنظیم می کنیم. در غیر این صورت، اجازه می دهیم $x_j = ۰$. برای مثال، بردار برای نشان

$$x = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \begin{array}{l} \text{a} \\ \text{aardvark} \\ \text{aardwolf} \\ \vdots \\ \text{buy} \\ \vdots \\ \text{zygmurgy} \end{array}$$

دادن ایمیلی استفاده می شود که حاوی کلمات "a" و "buy" است، اما نه "$aardvark$"، "$aardwolf$" یا "$zygmurgy$". مجموعه کلماتی که در بردار ویژگی کدگذاری شده اند، واژگان نامیده می شود. بنابراین بعد x برابر با اندازه واژگان است.

با انتخاب بردار ویژگی خود، اکنون می خواهیم یک مدل $generative$ بسازیم. بنابراین، باید $p(x|y)$ را مدل کنیم. اما اگر مثلاً واژگانی متشکل از ۵۰۰۰۰ کلمه داشته باشیم، آنگاه $x \in \{0,1\}^{50000}$ (x یک بردار ۵۰۰۰۰ بعدی از ۰ و ۱ است)، و اگر بخواهیم x را به طور صریح با توزیع $multinomial$ بر روی 2^{50000} نتیجه ممکن مدل کنیم، سپس ما به یک بردار پارامتر $(2^{50000} - 1)$ بعدی خواهیم رسید. این به وضوح تعداد پارامترهای زیادی است.

بنابراین برای مدل کردن $p(x|y)$، به یک فرض نیاز خواهیم داشت. یک فرضی وجود دارد به نام فرض $Nave\ Bayes$ که میگوید با فرض کلاس، فیچر ها مستقل از هم تولید میشوند:

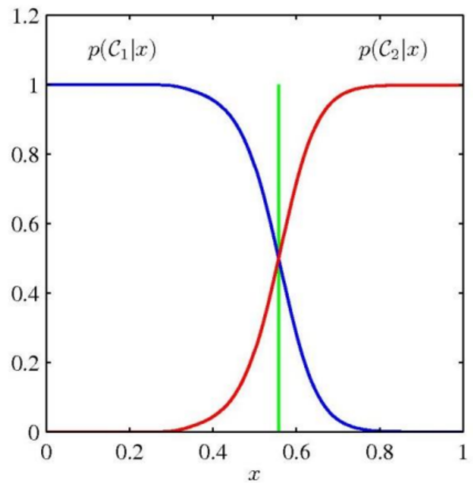
$$p(\boldsymbol{x}|C_k) = p(x_1|C_k) \times p(x_2|C_k) \times \cdots \times p(x_d|C_k)$$

برای مثال، اگر y = ۱ به معنای ایمیل $spam$ است. "buy" کلمه ۲۰۸۷ام و "$price$" کلمه ۳۹۸۳۱ است. پس فرض می کنیم که اگر شما بگوییم y = ۱ (که یک ایمیل خاص $spam$ است)، دانش x_{2087} (دانش اینکه آیا "buy" در پیام ظاهر می شود) هیچ تاثیری بر باورهای شما در مورد ارزش x_{39831} ندارد.
به طوری دیگر، این را می توان به صورت $p(x_{2087}|y x_{39831}) = p(x_{2087}|y)$ نوشت. (توجه داشته باشید که این همان چیزی نیست که بگوییم x_{2087} و x_{39381} مستقل هستند، که نوشته می‌شد ($p(x_{2087}) =$ $p(x_{2087}|x_{39381})$) بلکه فقط فرض می‌کنیم که x_{2087} و x_{39381} مستقل شرطی $given\ y$ هستند) اکنون داریم:

$$\begin{aligned} p(x_1, \dots, x_{50000}|y) &= P(x_1|y)P(x_2|y, x_1)P(x_3|y, x_1, x_2) \cdots p(x_{50000}|y, x_1, \dots, x_{49999}) \\ &= p(x_1|y)p(x_2|y)p(x_3|y) \cdots p(x_{50000}|y) \\ &= \prod_{j=1}^{50000} p(x_j|y) \end{aligned}$$

$Discriminative\ approach$

در دیدگاه $Discriminative$ مستقیم از توضیع $posterior$ استفاده میکنیم برای هر کلاس C_k.



برای یک تعداد خوبی از توابع $conditionallikelihood$ توضیع $posterior$ شبیه $logisticregression$

میشود. در واقع یکی از توابع مشهور برای مدل کردن $posterior$ همین تابع $logistic\, regression$ است.

یعنی فرضیمونو یک سیگموئید میگیریم از $w^t x$:

$$h(x; w) = \sigma(w^2 x)$$

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

داریم که:

$$x = \begin{bmatrix} 1 & x_1 & \dots & x_a \end{bmatrix}$$

$$w = \begin{bmatrix} w_0 & w_1 & \dots & w_a \end{bmatrix}$$

و خب مساله در حالت تخمین میشه تخمین زدن مقادیر w ها. و خب مقدار $h(x)$ بدیهتا بین ۰ و ۱ است و مقدار $y = 1$ را $given\ x$ نشان میدهد.