باسمه تعالی

# داده کاوی با استفاده از استنتاج ویادگیری بیزی

زیر نظر آقای دکتر رهگذر

توسط مصطفی حقیر چهرقانی

چکیده :

استدلال بیزی یک روش احتمالی برای استنتاج فراهم می‌آورد . این روش بر اساس این فرض بنا نهاده شده است که مقادیر مورد توجه از یک توزیع احتمال پیروی می‌کنند واینکه تصمیم‌های بهینه می‌توانند با استدلال بر روی این احتمالات به همراه داده‌های مشاهده شده اتخاذ شوند . بدلیل اینکه این روش یک راه کار کمی برای وزن دهی شواهدی که از فرض‌های مختلف پشتیبانی می‌کند ، فراهم آورد ، در مبحث یادگیری ماشین از اهمیت فراوانی برخوردار است . استدلال بیزی، روشی مستقیم برای کار با احتمالات برای الگوریتم‌های یادگیری فراهم آورد ، و همچنین چارچوبی برای تحلیل عملکرد الگوریتم‌هایی که مستقیماً با احتمالات سروکار ندارند ایجاد می‌نماید .

## ۱ – مقدمه

اهمیت استدلال بیزی داده کاوی را می‌توان به دو دلیل عمده نسبت داد. اول اینکه ، الگوریتم‌های یادگیری بیزی که به طور صریح بر روی احتمالات فرض‌های مختلف کار می‌کنند، مانند naïve Bayes classifier که از جمله کاراترین وعملی‌ترین الگوریتم‌های ممکن برای برخی مسائل یادگیری می‌باشد . به عنوان مثال Michie (۱۹۹۴) مقایسه کاملی بین این الگوریتم و سایر الگوریتم‌ها مانند درخت تصمیم و شبکه عصبی انجام داده است . این محقق نشان می‌دهد که الگوریتم naïve Bayes classifier قابل رقابت با سایر الگوریتم‌ها و در برخی موارد بهتر از آنها عمل می‌کند .

دلیل دوم این است که روش‌های استدلال بیزی چشم انداز مفیدی برای درک عملکرد الگوریتم‌های که مستقیماً برروی احتمالات عمل نمی‌کنند ایجاد می‌کند . برای نشان دادن این موضوع یک دلیل روشن برای یکی از مسائل یادگیری در شبکه عصبی که همان انتخاب کمترین مجموع مربعات خطا می‌باشد بیان می‌کنیم . همچنین تابع خطای دیگری به نام cross entropy ارائه می‌شود که از تابع خطای کمترین مجموع مربعات خطا در یادگیری توابع هدفی که بیانگر احتمال هستند کاراتر می‌باشد . به عنوان مثالی دیگر، اصل Minimum Description Length را با استفاده از اصول تئوری اطلاعات توضیح داده می‌شود .

از ویژگیهای یادگیری بیز می‌توان موارد زیر را نام برد :

• هر نمونه آموزشی جدید که مشاهده می‌شود می‌تواند احتمال درستی یک فرض را افزایش یا کاهش دهد . به این خاطر از روش‌های که با ناسازگاری یک نمونه فرض را کلاً حذف می کنند منعطف‌تر می‌باشد .

• دانش پیشین می‌تواند با مشاهدات ترکیب شده تا دانش جدید یا به عبارت دیگر احتمال درستی فرضیات را به وجود آورد. دانش پیشین به وسیله (۱) در نظر گرفتن احتمال هر فرض و (۲) انتساب یک توزیع احتمال برای مشاهدات ، ساخته می‌شود .

• روش‌های بیزی می‌توانند از فرض‌هایی که احتمال را پیشبینی می‌کنند بهره گیرند (به عنوان مثال « این مریض به احتمال ۹۳٪ شانس بهبودی کامل را دارد »)

• نمونه‌های جدید می‌تواند با استفاده از ترکیب وزنی نمونه‌های قبل متناسب با احتمال آنها تولید شود .

• حتی در مواردی که روش‌های بیزی از لحاظ پیچیدگی محاسبات غیر قابل استفاده باشند می‌توان ازآنها به عنوان بهترین روش (gold standard) برای مقایسه سایر روشها استفاده کرد .

یک مشکل عملی در استفاده از روش‌های بیزی این است که آنها عموماً نیاز به دانستن احتمالات پیشین بسیاری دارند . وقتی این احتمالات از قبل معلوم نباشد آنها بر اساس داده‌های موجود و دانش پیشین و توزیع احتمالی که بر روی فرض‌ها وجود دارد تخمین زده می‌شود . یک مشکل دیگر که در عمل به وجود می‌آید ، هزینه محاسباتی زیاد هنگام محاسبه بهترین فرض بیزی در حالت عمومی است (هزینه محاسبات به طور خطی با تعداد فرض‌ها افزایش می‌یابد ، یا به طور نمایی با تعداد متغیرها) . در موارد خاص این هزینه محاسبات می‌تواند به شدت کاهش یابد .

## ۲- تئوری بیز

در بسیاری از موارد به دنبال پیدا کردن بهترین فرض در فضای مفروضات H، با در اختیار داشتن داده‌های آموزشی D هستیم . یک روش برای بیان بهترین فرض این است که بگوییم ما به دنبال محتمل‌ترین فرض ، با داشتن داده D به علاوه دانش اولیه در مورد احتمالات پیشین فرض‌های H، هستیم . قضیه بیز روش مستقیم برای محاسبه این احتمالات فراهم می‌آورد .

برای تعریف قضیه بیز ابتدا کمی نماد گذاری لازم است . ما از $P(h)$ برای بیان احتمال اولیه‌ای که فرض h درست است استفاده می‌کنیم، پیش از آنکه داده‌های آموزشی را دیده باشیم . $P(h)$ را عموماً احتمال پیشین می‌نامند و بیانگر هر دانش پیشینی می‌باشد که در مورد شانس درستی فرض h سخن می‌گوید . اگر هیچ دانش اولیه‌ای از مفروضات نداشته باشیم می‌توانیم یک احتمال یکسان به کل فضای مفروضات H اختصاص دهیم . به طور مشابه از $P(D)$ برای بیان احتمال پیشین که داده‌های D مشاهده می‌شوند استفاده می‌کنیم (به عبارت دیگر احتمال مشاهده D به شرط اینکه هیچ دانشی در مورد درستی مفروضات موجود نباشد). همچنین از $P(D|h)$ برای بیان احتمال D در دنیایی که فرض h صادق است استفاده می‌کنیم . در یادگیری ماشین ما به دنبال $P(h|D)$

هستیم ، یعنی احتمال درستی فرض h به شرط مشاهده داده‌های آموزشی D . P(h|D) احتمال پسین h نام دارد ، بدین علت که بیانگر اطمینان ما از فرض h پس از مشاهده داده‌های D می‌باشد .

قضیه بیز اصلی‌ترین سنگ بنای یادگیری بیزی می‌باشد، زیرا روشی برای محاسبه احتمال پسین P(h|D) را از احتمال پیشین P(h) به همراه P(D) و P(D|h) فراهم می‌آورد .

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)} \qquad ۲,۱$$

همانطور که انتظار می‌رود ، می‌توان مشاهده کرد که P(h|D) با افزایش P(h) و همچنین P(D|h) افزایش می‌یابد . به همین ترتیب منطقی به نظر می‌رسد که P(D|h) با افزایش P(D) کاهش یابد ، زیرا با افزایش احتمال وقوع P(D) که مستقل از h می‌باشد ، شواهد کمتری در D برای پشتیبانی از h وجود خواهد داشت .

در بسیاری از سناریوهای یادگیری، یادگیرنده مجموعه‌ای از فرض‌ها H را در نظر می‌گیرد وعلاقمند به یافتن فرضی $h \in H$ می‌باشد که محتمل‌ترین باشد (یا حداقل یکی از محتمل‌ترین مفروضات اگر چندین تا وجود دارد) . هر فرضی که دارای این خصوصیت باشد به فرض (MAP) Maximum a posteriori نام دارد . ما می‌توانیم فرض MAP را با استفاده از قضیه بیز برای محاسبه احتمال پسین هر کاندیدا بیابیم . به عبارت دقیق‌تر $h_{MAP}$ فرضی است که،

$$\begin{aligned} h_{MAP} &= \arg\max_{h \in H} P(h \mid D) \\ &= \arg\max_{h \in H} \frac{P(D \mid h)P(h)}{P(D)} \\ &= \arg\max_{h \in H} P(D \mid h)P(h) \end{aligned} \qquad (۲,۱)$$

توجه کنید که در مرحله آخر بالا P(D) را حذف کردیم زیرا محاسبه آن مستقل از h می‌باشد و همیشه یک عدد ثابت است . در برخی موارد ، فرض می‌کنیم که تمام مفروضات $h \in H$ احتمال وقوع یکسانی دارند (یعنی $\forall h_i, h_j \in H; P(h_i) = P(h_j)$). دراین صورت یک ساده سازی دیگر نیز در فرمول (۲,۲) می‌توان انجام داد . به عبارت دیگر می‌توان فرضی را که P(D|h) را ماکزیمم می‌کند در نظر گرفت (Maximum Likelihood) .

$$h_{ML} = \arg\max_{h \in H} P(D \mid h) \qquad (۲,۲)$$

در اینجا، از داده‌های D به عنوان نمونه‌های آموزشی برای یک تابع هدف و از H به عنوان مجموعه‌ای از توابع هدف ممکن یاد می‌کنیم. اما در حقیقت قضیه بیز کلی‌تر از این بحث است . یعنی می‌توان آن را به گونه‌ای مشابه برای مجموعه‌ای از هر نوع مفروضات دو به دو مستقل از هم که مجموع احتمالات آنها یک می‌شود استفاده کرد .

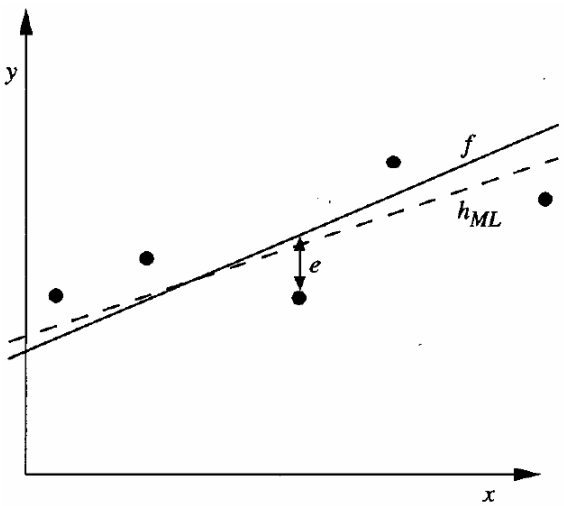
## ۳- درست‌نمایی ماکزیمم (Maximum Likelihood) و فرض کمترین مربعات خطا

در این بخش، مسأله تابع هدف پیوسته را در نظر می‌گیریم ، مسأله‌ای که در بسیاری از روش‌های یادگیری مانند شبکه عصبی، رگرسیون خطی و برازش منحنی‌های چند جمله‌ای وجود دارد . یک تحلیل بیزی ساده نشان می‌دهد که « در صورت برقرار بودن یک سری شرایط هر الگوریتم داده کاوی که مربعات خطای بین خروجی پیشبینی شده فرض و داده‌های آموزشی را کمینه کند ، منجر به خروجی یک فرض درست‌نمایی ماکزیمم می‌شود» . اهمیت این نتیجه در آن است که صحت کارایی بسیاری از الگوریتم‌هایی که خطای مربعات را کمینه می‌کنند را نشان می‌دهد .

مسئله‌ای با شرایط زیر را در نظر گیرد . الگوریتم یادگیری L در فضای نمونه X و فضای مفروضات H که شامل دسته‌ای از توابع حقیقی بر روی X (یعنی هر $h \in H$ تابعی به شکل $h : X \to \Re$ که $\Re$ مجموعه اعداد حقیقی می‌باشد) است .

مسأله‌ای که الگوریتم L باید به آن بپردازد یافتن تابع هدفی مانند $f : X \to \Re$ است که در فضای H می‌باشد . مجموعه‌ای به اندازه m از نمونه‌های آموزشی موجود می‌باشد که مقدار خروجی هر نمونه دارای مقداری نویز با توزیع نرمال می‌باشد . به عبارت دقیق‌تر هر نمونه آموزشی از یک جفت $< x_i, d_i >$ تشکیل شده که در آن $d_i = f(x_i) + e_i$ . در اینجا $f(x_i)$ مقدار بدون نویز تابع و $e_i$ یک متغیر تصادفی است که نمایانگر نویز می‌باشد . فرض بر آن است که مقادیر $e_i$ به صورت مستقل تولید شده که توزیع احتمال آن از یک توزیع نرمال با میانگین صفر پیروی می‌کند . وظیفه L بدست آوردن یک فرض درست‌نمایی ماکزیمم ، یا به طور مشابه فرض MAP هنگامی که تمام مفروضات دارای احتمال یکسان هستند می‌باشد .

یک مثال ساده ، یادگیری یک تابع خطی می‌باشد ، اگر چه تحلیل که در اینجا بیان می‌شود برای هر تابعی می‌تواند استفاده شود . شکل (۱,۳) یک تابع خطی $f$ را با خط ممتد به همراه داده‌های آموزشی دارای نویز نشان می‌دهد . خط نقطه چین شده بیانگر یک فرض $h_{ML}$ می‌باشد که دارای کمترین مربعات خطا است . توجه داشته باشید که فرض ML همیشه برابر فرض اصلی ، $f$، نمی‌باشد ، زیرا از تعداد محدودی نمونه آموزشی خطا دار منتج شده است .




شکل ۱,۳ یادگیری یک تابع حقیقی. خط ممتد متناظر با تابع هدف $f$ نقاط متناظر با نمونه‌های آموزشی می‌باشد. خط نقطه‌چین شده متناظر با فرض درست‌نمایی ماکزیمم که بر اساس کمترین مجموع مربعات خطا آموزش دیده می‌باشد.

در اینجا به نشان دادن مسأله اصلی ، یعنی اینکه فرض کمترین مربعات خطا در حقیقت فرض درست‌نمایی ماکزیمم می‌باشد می‌پردازیم . خاطر نشان می‌شود که به جای P(D|h) از p(D|h) که نشان دهنده تابع چگالی احتمال می‌باشد استفاده شده است زیرا تابع مورد نظر پیوسته است و نمی‌توان از تابع توزیع احتمال برای آن استفاده کرد،

$$h_{ML} = \arg\max_{h \in H} p(D \mid h)$$

همانند قبل مجموعه‌ای از نمونه‌های آموزشی $< x_1,...,x_m >$ را در نظر می‌گیریم و از اینجا مجموعه متناظر مقادیر هدف $D =< d_1, d_2,..., d_m >$ وجود خواهد داشت . همچنین $d_i = f(x_i) + e_i$ با فرض استقلال دو به دو نمونه ها به شرط h می‌توان نوشت،

$$h_{ML} = \arg\max_{h \in H} \prod_{i=1}^{m} p(d_i \mid H)$$

بافرض اینکه نویز $e_i$ از یک توزیع نرمال با میانگین صفر و واریانس نامعلوم $\sigma^2$ پیروی می‌کند، هر $d_i$ نیز نیز باید از یک توزیع نرمال با واریانس $\sigma^2$ که دارای میانگین $f(x_i)$ می‌باشد پیروی کند . در نتیجه $p(d_i|h)$ را می توان با یک توزیع نرمال با واریانس $\sigma^2$ و میانگین $\mu = f(x_i)$ نشان داد .

به علت آنکه ما عبارت احتمال $d_i$ را با فرض درست بودن h برای تعریف تابع هدف $f$ می‌نویسیم ،

در این فرمول نیز قرار می‌دهیم $\mu = f(x_i) = h(x_i)$ در نتیجه،

$$h_{ML} = \arg\max_{h\in H} \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(d_i-\mu)^2}$$
$$= \arg\max_{h\in H} \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(d_i-h(x_i))^2} \qquad (۳,۱)$$

حال در اینجا یک تبدیل که در محاسبات درست‌نمایی ماکزیمم معمول می‌باشد انجام می‌دهیم . به جای ماکزیمم کردن عبارت پیچیده (۳,۱) از لگاریتم این عبارت که دارای پیچیدگی کمتری می‌باشد استفاده می‌کنیم . این عمل امکان پذیر است زیرا تابع لگاریتم یک تابع اکیداً یکنوا می‌باشد . بنابراین با ماکزیمم کردن $\ln(p)$ به جای p داریم،

$$h_{ML} = \arg\max_{h\in H} \sum_{i=1}^{m} \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}\left(d_i - h(x_i)\right)^2 \qquad (۳,۲)$$

جمله اول در عبارت (۳,۲) که ثابتی مستقل از h می‌باشد و می‌توان آن را نادیده گرفت . در نتیجه،

$$h_{ML} = \arg\max_{h\in H} \sum_{i=1}^{m} -\frac{1}{2\sigma^2}\left(d_i - h(x_i)\right)^2 \qquad (۳,۳)$$

ماکزیمم کردن عبارت (۳,۳) معادل مینیمم کردن منفی آن است،

$$h_{ML} = \arg\max_{h\in H} \sum_{i=1}^{m} \frac{1}{2\sigma^2}\left(d_i - h(x_i)\right)^2$$

نهایتاً ، می‌توان ضرایب ثابتی که مستقل از h هستند را نادیده گرفت،

$$h_{ML} = \arg\max_{h\in H} \sum_{i=1}^{m} \left(d_i - h(x_i)\right)^2 \qquad (۳,۴)$$

بنابراین عبارت (۳,۴) نشان می‌دهد که فرض درست‌نمایی ماکزیمم $h_{ML}$ ، فرضی است که مجموع مربعات خطا بین داده‌های آموزشی مشاهده شده $d_i$ و پیشبینی خروجی از فرض $h(x_i)$ را مینیمم

کند . این مسأله در صورتی صادق است که داده‌های آموزشی به وسیله اضافه کردن نویز نرمال تصادفی با میانگین صفر و به طور مستقل تولید شده باشند . مشاهده می‌شود که عبارت مربع خطا مستقیماً از نمای توزیع نرمال ناشی می‌شود . می‌توان عبارات مشابهی را با فرض‌های دیگری در مورد تابع توزیع خطا بدست آورد .

توجه کنید که در روش بدست آوردن عبارت (۳,۴) از ماکزیمم کردن لگاریتم درست‌نمایی ( $\ln p(D|h)$ ) استفاده شده است . این روش منجر به رسیدن به جواب یکسان اما با عبارت ریاضی ساده‌تر می‌شود .

به چه علت فرض نرمال بودن تابع توزیع احتمال نویز معقول به نظر می‌رسد؟ یک دلیل این است که موجب تحلیل ریاضی ساده‌ای می‌شود . دلیل دوم این است که شکل توزیع نرم و زنگوله‌ای این تابع، تقریب مناسبی از نویز سیستم‌های واقعی می‌باشد. در حقیقت قضیه حد مرکزی بیانگر این است که مجموع تعداد کافی از متغیرهای تصادفی مستقل با توزیع یکسان (independent identical distributed (iid)) از یک توزیع نرمال پیروی می‌کند مستقل از اینکه توزیع تک تک متغیرهای تصادفی چه باشند . البته مسلماً ممکن است در عمل فاکتورهای مختلف تولید نویز، دارای توزیع یکسانی نباشند که در نتیجه قضیه صادق نخواهد بود .

خاطر نشان می‌شود که نویز در تحلیل فوق بر روی خروجی فرض تاثیر می‌گذاشت . اگر مایل به بررسی شرایطی باشیم که نویز بر روی خصوصیات (پارامترهای ورودی) نیز تاثیر داشت ، تحلیل‌های بسیار پیچیده تری مورد نیاز بود .

## ۴ – اصل طول توصیف مینیمم

در این بخش ما یک دیدگاه بیزی در اصل Ockham's razor و رابطه آن با اصل طول توصیف مینیمم بیان می‌کنیم . اصل Ockham's razor یک اصل معروف برای بایاس استنتاج به گونه‌ای است که :

«خلاصه‌ترین توضیح را برای بیان مشاهدات انتخاب کن» .

اصل طول توصیف مینیمم با بررسی تعریف $h_{MAP}$ با استفاده از مفاهیم تئوری اطلاعات بخوبی بیان می‌شود .

تعریف آشنای $h_{MAP}$ را در نظر گیرید،

$$h_{MAP} = \arg\max_{h \in H} P(D \mid h) P(h)$$

که می توان آنرا به صورت زیر نیز نوشت،

$$h_{MAP} = \arg\max_{h \in H} \log_2 P(D \mid h) + \log_2 P(h)$$

یا به عبارت دیگر،

$$h_{MAP} = \arg\min_{h \in H} -\log_2 P(D \mid h) - \log_2 P(h) \qquad (۴,۱)$$

با توجه به فرمول (۴,۱) می‌تواند این‌گونه توجیح کرد که فرض‌های کوتاه‌تر با در نظر گرفتن یک ارائه خاص و روشی برای کدگذاری فرض و داده بر دیگر فرض‌ها ترجیح دارند . برای توضیح این مسأله باید ابتدا با مفاهیمی از تئوری اطلاعات آشنا شویم .

فرض کنید باید رمزی برای ارسال یک سری پیام که به طور تصادفی انتخاب می‌گردد طراحی کنیم . به عبارت دیگر احتمال انتخاب پیام i، $p_i$ می‌باشد، البته ما به فشرده‌ترین رمز علاقه‌مندیم؛ یعنی رمزی که امید تعداد بیت‌های ارسالی برای فرستادن پیام را مینیمم کند . به وضوح می‌توان دریافت که برای مینیمم کردن امید طول رمز باید تعداد بیت کمتری به پیام‌هایی که با احتمال بیشتر ظاهر می‌شوند اختصاص داد. Weaver و Shamon (۱۹۴۹) نشان دادند که کد بهینه

$-\log_2 p_i$ بیت به پیام i، اختصاص می‌دهد . از این پس به تعداد بیتهای لازم برای رمز گذاری پیام i به روش کدگذاری C طول توصیف پیام i در روش C می‌گوییم و آن را با $L_C(i)$ نشان می دهیم .

حال با تحلیل معادله (۴,۱) با استفاده از نتایج بدست آمده از تئوری اطلاعات داریم :

• $-\log_2 P(h)$ طول توصیف بهینه برای h در فضای فرض‌های H می‌باشد . به عبارت دیگر این مقدار، اندازه توصیف بهینه برای فرض h با استفاده از بهترین ارائه ممکن می‌باشد . در نماد گذاری ما،

$$L_{C_H} = -\log_2 P(h)$$

که در آن $C_H$ که بهینه برای فضای فرض H می‌باشد .

• $-\log_2 P(D|h)$ طول توصیف داده‌های آموزشی D به شرط درستی فرض h ، در کد گذاری بهینه می‌باشد . در نماد گذاری ما،

$$L_{C_{D|H}} = -\log_2 P(D|h)$$

• بنابراین می‌توان معادله (۴,۱) را به گونه‌ای بازنویسی کرد که نشان دهد $h_{MAP}$ فرضی را که مجموع طول توصیف فرض و طول توصیف داده به شرط داشتن فرض را مینیمم می‌کند بیان می‌نماید،

$$h_{MAP} = \arg\min_{h \in H} L_{C_H}(h) + L_{C_{D|H}}(D|h)$$

که در آن $C_H$ و $C_{D|H}$ متناظراً کد گذاری‌های بهینه برای H و D به شرط H می باشند .

بنابراین اصل MDL روشی را برای متعادل کردن پیچیدگی فرض وتعداد خطاهایی که مرتکب می‌شود فراهم می‌نماید . یعنی ممکن است فرض کوتاه‌تری را که کمی خطا می‌کند به فرض بلندی که هیچ خطایی مرتکب نمی‌شود ترجیح دهد . به این ترتیب روشی را برای برخورد با برازش بیش از حد مدل با داده ارائه می‌نماید .

Quinlan ,Rivest (۱۹۸۹) آزمایش‌هایی برای تعیین اندازه درخت تصمیم‌های ساخته شده به روش MDL توصیف می‌نمایند . نتایج آنها حاکی از قابل مقایسه بودن دقت طبقه بندی درخت تصمیم‌های ساخته شده به روش MDL با درخت‌های ساخته شده با روش‌های استاندارد هرس کردن می‌باشد . این نتایج در کارهای Mehta(۱۹۹۵) نیز دیده می‌شود .

آیا از مباحث انجام شده در اصل MDL می‌توان نتیجه گرفت که همواره فرض کوتاه‌تری که توسط این اصل ترجیح داده می‌شود بهترین فرض مورد نظر است؟ خیر . تنها مطلبی که دراینجا نشان دادیم این است که اگر ارائه فرض به گونه‌ای باشد که اندازه آن بر اساس $-\log_2 P(h)$ وارائه داده‌ها به گونه‌ای باشد که طول داده‌های کدگذاری شده D به شرط دانستن فرض h برابر $-\log_2 P(D|h)$ باشد آنگاه فرضی که بر اساس اصل MDL باشد فرض MAP را تولید می‌نماید . اما ، برای داشتن چنین ارائه‌ای باید احتمالات پیشین P(h) وهمچنین درست‌نمایی P(D|h) را در اختیار داشته باشیم . در غیر این صورت هیچ دلیلی وجود ندارد که فرض MDL را با هر کدگذاری دلخواه $C_1$ و $C_2$ بر سایر فرض‌ها ترجیح دهیم.

## ۵- طبقه بندی بهینه بیزی (Bayes Optimal Classifier)

تا به حال ما این سوال را مورد توجه قرار دادیم که «چه فرضی محتمل‌ترین فرض با داشتن داده‌های آموزشی D می‌باشد؟» .

در حقیقت سوال بسیار مهم دیگری که خیلی نزدیک به این سوال است این است که «محتمل‌ترین دسته داده جدید با در اختیار داشتن داده‌های آموزشی D چیست ؟» . ممکن است به نظر رسد که استفاده از فرض MAP برای دسته بندی داده جدید کفایت می‌کند , اما از آن بهتر نیز می‌توان عمل کرد .

برای درک این موضوع فرض کنید فضای مفروضات ما دارای سه فرض $h_1, h_2, h_3$ می‌باشد . همچنین احتمال پسین این فرضیات با توجه به داده‌های آموزشی به ترتیب برابر ۰/۴ و ۰/۳ و ۰/۳ می‌باشد . بنابراین $h_1$ فرض MAP است . فرض کنید نمونه جدید x مشاهده می‌شود که به وسیله فرض $h_1$ مثبت و توسط سایر فرض‌ها منفی دسته بندی می‌شود . با در نظر گرفتن کل مفروضات احتمال اینکه x مثبت باشد ۰/۴ و احتمال اینکه منفی باشد ۰/۶ می‌باشد. محتمل‌ترین دسته‌بندی در این مسأله با دسته‌ای که محتمل‌ترین فرض ( فرض MAP ) را تولید می‌کند متفاوت می‌باشد .

در حالت کلی محتمل‌ترین دسته بندی نمونه جدید به وسیله ترکیب احتمالات تمام مفروضات به وسیله وزن دهی آنها بر اساس احتمال پسین مفروضات انجام می‌گیرد. اگر دسته محتمل $v_j$ از یک مجموعه مانند V انتخاب گردد آنگاه احتمال $P(v_j|D)$ که بیانگر احتمال درستی دسته نمونه جدید به شرط مشاهده D می‌باشد به صورت زیر محاسبه می‌گردد،

$$P(v_j \mid D) = \sum_{h_i \in H} P(v_j \mid h_i) P(h_i \mid D) \qquad (۵,۱)$$

بهترین دسته بندی برای نمونه جدید دسته‌ای خواهد بود که عبارت ۵,۱ را ماکزیمم کند , یعنی،

$$\arg\max_{v_j \in V} \sum_{h_i \in H} P(v_j \mid h_i) P(h_i \mid D) \qquad (۵,۲)$$

برای اینکه فرمول (۵,۲) را برای مثال قبل توضیح دهیم مجموعه $V = \{\oplus, \otimes\}$ را برای دسته بندی نمونه جدید در نظر گیرید،

$$\begin{aligned} &P(h_1 \mid D) = .4, \quad P(\otimes \mid h_1) = 0, \quad P(\oplus \mid h_1) = 1 \\ &P(h_2 \mid D) = .3, \quad P(\otimes \mid h_2) = 1, \quad P(\oplus \mid h_2) = 0 \\ &P(h_3 \mid D) = .3, \quad P(\otimes \mid h_3) = 1, \quad P(\oplus \mid h_3) = 0 \end{aligned}$$

بنابر این،

$$\sum_{h_i \in H} P(\oplus \mid h_i) P(h_i \mid D) = .4$$
$$\sum_{h_i \in H} P(\otimes \mid h_i) P(h_i \mid D) = .6$$

**و در نتیجه،**

$$\arg\max_{v_j \in \{\oplus, \otimes\}} \sum_{h_i \in H} P(v_j \mid h_i) P(h_i \mid D) = \otimes$$

**هر سیستمی که نمونه‌های جدید را بر اساس فرمول (۵,۲) دسته‌بندی نماید یک روش طبقه بندی بهینه بیزی خواهد بود . هیچ روش دیگر وجود ندارد که با داشتن فضای مفروضات یکسان و دانش پیشین مفروضات یکسان بتواند بهتر از این روش عمل نماید . این روش با توجه به داده‌های موجود ، فضای مفروضات و احتمالات پیشین بر روی مفروضات ، احتمال درست دسته‌بندی کردن نمونه جدید را ماکزیمم می‌کند .**

**یک نکته جالب در مورد طبقه بندی بهینه بیزی است این است که تخمین‌های آن می‌تواند مربوط به فرضی باشد که خارج از فضای مفروضات H است! در نظر گیرید که با استفاده از فرمول (۵,۲) تمام نمونه‌ها در X را دسته‌بندی کردیم . بر چسب گذاری نمونه‌ها به این صورت نیازی ندارد معادل با برچسب گذاری یکی از فرض‌ها مانند h از فضای مفروضات H باشد . می‌توان اینگونه تصور کرد که طبقه بندی بهینه بیزی از فضای مفروضات H که متفاوت و ابرمجموعه مفروضات H است استفاده می‌کند . در حقیقت فضای H شامل هر ترکیب خطی از فضای H می‌باشد .**

## ۶- الگوریتم Gibbs

**اگر چه روش طبقه بندی بهینه بیزی بهترین کارایی را دارد ، اما از نظر محاسباتی بسیار پیچیده است . این هزینه زیاد به علت محاسبه احتمال پسین تمامی مفروضات و وزن دهی بر اساس آن برای دسته بندی هر نمونه جدید می‌باشد .**

یک روش که بهینگی و پیچیدگی کمتری دارد الگوریتم Gibbs می‌باشد( Haussler و Opper، ۱۹۹۱) که به صورت زیر است :

۱- فرض h را از فضای H به صورت تصادفی، براساس توزیع احتمال پسین بر روی H انتخاب کن.

۲- از h برای دسته‌بندی نمونه جدید x استفاده کن .

هنگام دریافت نمونه جدید الگوریتم Gibbs فرض را به تصادف براساس توزیع احتمال پسین کنونی انتخاب می‌کند . می‌توان نشان داد که در چنین شرایطی امید دسته بندی کردن اشتباه حداکثر دو برابر خطای طبقه‌بندی بهینه بیزی می‌باشد (Haussler ،۱۹۹۴).

## ۷- طبقه‌بندی ساده بیزی

یک روش بسیار کاربردی یادگیری بیزی روش یادگیرنده ساده بیزی می‌باشد که عموماً روش طبقه‌بندی ساده بیزی نامیده می‌شود . در برخی زمینه‌ها نشان داده شده است که کارایی آن قابل قیاس با کارایی روش‌هایی مانند شبکه عصبی و درخت تصمیم می‌باشد . این بخش، روش طبقه بندی ساده بیزی را معرفی می‌کند . بخش بعدی کاربرد این روش را در یک مثال عملی نشان می‌دهد .

طبقه‌بندی ساده بیزی برای مسائلی که هر نمونه x درآن توسط مجموعه‌ای از مقادیر صفات و تابع هدف f(x) از مجموعه‌ای مانند V انتخاب می‌گردد کاربرد دارد . مجموعه‌ای از داده‌های آموزشی و خروجی تابع هدف و یا طبقه‌ای که نمونه جدید به آن تعلق دارد مورد نظر است .

روش بیزی برای طبقه‌بندی نمونه جدید این است که محتمل‌ترین طبقه یا مقدار هدف $v_{MAP}$ را با داشتن مقادیر صفات $< a_1, a_2, ..., a_n >$ که توصیف کننده نمونه جدید است شناسایی کند،

$$v_{MAP} = \arg\max_{v_j \in V} P(v_j \mid a_1, a_2, ..., a_n) \qquad (۷,۱)$$

با استفاده از قضیه بیز می‌توان عبارت (۷,۱) را به صورت زیر بازنویسی کرد،

$$\begin{aligned} v_{MAP} &= \arg\max_{v_j \in V} \frac{P(a_1, a_2, ..., a_n, \mid v_j) P(v_j)}{P(a_1, a_2, ..., a_n)} \\ &= \arg\max_{v_j \in V} P(a_1, a_2, ..., a_n, \mid v_j) P(v_j) \qquad (۷,۲) \end{aligned}$$

حال با استفاده از داده‌های آموزشی سعی می‌کنیم دو جمله معادله (۷,۲) را تخمین بزنیم . محاسبه از روی داده‌های آموزشی به این صورت که میزان تکرار $v_j$ در داده‌ها چقدر است، آسان می‌باشد . اما محاسبه جملات مختلف $P(a_1, a_2, ..., a_n \mid v_j)$ به این صورت قابل قبول نخواهد بود مگر اینکه حجم بسیار بسیار زیادی از داده‌های آموزشی در اختیار داشته باشیم . مشکل اینجاست که تعداد این جملات برابر تعداد نمونه‌های ممکن ضرب در تعداد مقادیر تابع هدف می‌باشد . بنابراین باید هر نمونه را چندین بار مشاهده کنیم تا تخمین مناسبی از آن بدست آید .

فرض روش طبقه‌بندی ساده بیزی بر اساس این ساده سازی است که مقادیر صفات با داشتن مقادیر تابع هدف از یکدیگر مستقل شرطی می‌باشند . به عبارت دیگر ، این فرض بیانگر این است که به شرط مشاهده خروجی تابع هدف احتمال مشاهده صفات $a_1, a_2, ..., a_n$ برابر ضرب احتمالات هر صفت به طور جداگانه می‌باشد . اگر این را جایگزین معادله (۷,۲) کنیم روش طبقه‌بندی ساده بیزی را نتیجه می‌دهد،

$$v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_i P(a_i \mid v_j) \qquad (۷,۳)$$

که $v_{NB}$ خروجی طبقه‌بندی ساده بیزی برای تابع هدف می‌باشد . توجه کنید که تعداد جملات $P(a_i \mid v_j)$ که در این روش باید محاسبه شوند برابر تعداد صفات ضرب در تعداد دسته‌های خروجی برای تابع هدف می‌باشد که این مقدار از تعداد جملات $P(a_1, a_2, ..., a_n \mid v_j)$ بسیار کمتر است .

نتیجه اینکه یادگیری ساده بیزی سعی در تخمین مقادیر مختلف $P(v_j)$ و $P(a_i|v_j)$ با استفاده از میزان تکرار آنها در داده‌های آموزشی دارد . این مجموعه تخمین‌ها متناظر با فرض یاد گرفته شده است . سپس از این فرض برای طبقه بندی نمونه‌های جدید استفاده می‌شود که این کار با استفاده از فرمول (۷,۳) صورت می‌گیرد . هر گاه فرض مستقل شرطی بودن روش طبقه بندی ساده بیزی بر آورده شود طبقه ساده بیزی معادل طبقه MAP خواهد بود .

## ۷,۱- یک مثال در توضیح طبقه بندی ساده بیزی

برای درک بهتر روش طبقه‌بندی ساده بیزی مثالی در رابطه با یادگیری مفهومی به این صورت بیان می‌کنیم :

«طبقه بندی روزها بر اساس اینکه فردی تنیس بازی می‌کند یا خیر!»

جدول (۷,۱) نشان دهنده ۱۴ نمونه آموزشی از مفهوم تنیس بازی کردن می‌باشد که هر روز با صفات وضعیت ، دما ، رطوبت و باد توصیف می‌شود .

جدول (۷,۱) داده‌های آموزشی برای مفهوم تنیس بازی کردن

| تنیس بازی کردن | باد | رطوبت | دما | وضعیت | روز |
|---|---|---|---|---|---|
| خیر | کند | زیاد | گرم | آفتابی | ۱ |
| خیر | تند | زیاد | گرم | آفتابی | ۲ |
| بله | کند | زیاد | گرم | ابری | ۳ |
| بله | کند | زیاد | متوسط | بارانی | ۴ |
| بله | کند | نرمال | خنک | بارانی | ۵ |

| | | | | | |
|---|---|---|---|---|---|
| خیر | تند | نرمال | خنک | بارانی | ۶ |
| بله | تند | نرمال | خنک | ابری | ۷ |
| خیر | کند | زیاد | متوسط | آفتابی | ۸ |
| بله | کند | نرمال | خنک | آفتابی | ۹ |
| بله | کند | نرمال | متوسط | بارانی | ۱۰ |
| بله | تند | نرمالب | متوسط | آفتابی | ۱۱ |
| بله | تند | نرمال | متوسط | ابری | ۱۲ |
| بله | کند | زیاد | گرم | ابری | ۱۳ |
| خیر | کند | نرمال | متوسط | بارانی | ۱۴ |

در اینجا ما می‌خواهیم با استفاده از طبقه بندی ساده بیزی و داده‌های آموزشی نمونه جدید زیر را طبقه بندی کنیم .

‹ وضعیت=آفتابی، دما=خنک، رطوبت=زیاد، باد=تند ›

وظیفه ما تعیین مقدار هدف (‹بله› یا ‹خیر›) مفهوم تنیس بازی کردن برای این نمونه جدید می‌باشد . با استفاده از فرمول (۷,۳) می‌توان $v_{NB}$ را اینگونه محاسبه کرد،

$$\begin{aligned} v_{NB} &= \arg\max_{v_j \in \{yes,no\}} P(v_j) \prod_i P(a_i \mid v_j) \\ &= \arg\max_{v_j \in \{yes,no\}} P(v_j) P(outlook = sunny \mid v_j) P(temprature = cool \mid v_j \\ &\quad P(humidity = high \mid v_j) P(wind = strong \mid v_j) \end{aligned} \qquad (۷,۱,۱)$$

برای محاسبه $v_{NB}$ باید مقادیر ۱۰ احتمال تعیین شوند که با استفاده از داده‌های آموزشی تخمین زده می‌شوند ابتدا احتمال مقادیر هدف را به راحتی با استفاده از فرکانس تکرار آن در ۱۴ نمونه آموزشی تخمین می‌زنیم،

$$P(playTennis = yes) = 9/14 = .64$$
$$P(playTennis = no) = 5/14 = .36$$

به طور مشابه می‌توانیم احتمالات شرطی را تخمین بزنیم . برای مثال آنهایی که مربوط به ‹باد=تند› می‌شوند عبارتند از،

$$P(wind = strong \mid playTennis = yes) = 3/9 = .33$$
$$P(wind = strong \mid playTennis = no) = 3/5 = .60$$

با استفاده ازاین احتمالات و تخمین‌های مشابهی برای سایر صفات ، از فرمول (۷,۱,۱) برای محاسبه $v_{NB}$ استفاده می‌کنیم (در اینجا نام صفات برای سادگی حذف شده‌اند )

$$P(yes)P(sunny \mid yes)P(cool \mid yes)P(high \mid yes)P(strong \mid yes) = .0053$$
$$P(no)P(sunny \mid no)P(cool \mid no)P(high \mid no)P(strong \mid no) = .0206$$

بنابراین طبقه‌بندی ساده بیزی مقدار ‹خیر= تنیس بازی کردن› را بر اساس مقادیر یاد گرفته شده از داده‌های آموزشی به نمونه جدید اختصاص می‌دهد . علاوه بر آن با نرمالیزه کردن مقادیر بالا بگونه‌ای که مجموع آن یک شود می توان احتمال شرطی آنکه مقدار هدف ‹خیر› شود را محاسبه نمود. برای مثال در این نمونه احتمال مورد نظر برابر است با $\frac{.0206}{.0206+.0053}$ .

## ۷,۲- تخمین زدن احتمالات

تا به حال، احتمالات را با استفاده از کسری از دفعات که مشاهده می‌شد نسبت به تعداد کل مشاهدات تخمین می‌زدیم . به عنوان مثال ما تخمین (خیر = تنیس بازی کردن| تند =باد )P را با

کسر $\frac{n_c}{n}$ محاسبه کردیم که $n$=۵ تعداد کل داده‌های آموزشی که ‹خیر=تنیس بازی کردن› است و $n_c$=۳ تعداد داده‌هایی از آنهاست که برای آنها ‹تند=باد› می‌باشد .

با وجود اینکه این کسر در بسیاری از موارد تخمین خوبی از احتمال مورد نظر است، هنگامی که $n_c$ بسیار کوچک باشد تخمین خوبی بدست نمی‌دهد. برای دیدن این مشکل فرض کنید مقدار واقعی (خیر=تنیس بازی کردن)P برابر ۰/۰۸ باشد و در داده‌های آموزشی ، فقط ۵ نمونه وجود دارد که ‹خیر = تنیس بازی کردن› باشد . محتمل‌ترین مقدار $n_c$ برابر صفر خواهد بود . این اثر، دو مسأله را به وجود می‌آورد . اول اینکه $\frac{n_c}{n}$ یک احتمال بایاس شده تولید می‌کند . دوم اینکه وقتی تخمین این احتمال صفر باشد ، این احتمال بر نمونه‌هایی که درآینده مقدار ‹تند=باد› داشته باشند در طبقه بندی بیز احتمال غالب خواهد بود . این مشکل به علت ضرب این احتمال صفر در فرمول (۷,۳) با دیگر به وجود می‌آید . برای جلوگیری از این امر می توان ازیک راه کار بیزی به نام m_estimate که به صورت زیر تعریف می گردد استفاده شود،

$$\frac{n_c + mp}{n + m} \qquad (۷,۲,۱)$$

دراینجا $n$ و $n_c$ مانند قبل می باشند و $p$ تخمین اولیه از احتمالی است که باید محاسبه شود. $m$ ضریبی است که به «اندازه نمونه متناظر» معروف است و کار آن وزن دهی به میزان اعتقاد ما از احتمال پیشینی که در نظر گرفتیم می‌باشد . یعنی اگر صفتی دارای k حالت ممکن می‌باشد $p = \frac{1}{k}$ .

برای مثال ، در تخمین (خیر=تنیس بازی کردن|تند=باد)P صفت باد دارای ۲ مقدار ممکن می‌باشد ، بنابراین $p$=۰/۵. دقت کنید که اگر $m$ در اینجا برابر صفر اختیار شود ، فرمول (۷,۲,۱) به تخمین معمولی تبدیل می‌گردد . اگر هر دوی $m$ و $n$ غیر صفر باشند آنگاه ضریب $\frac{n_c}{n}$ و $p$

متناسب با وزن $m$ با یکدیگر ترکیب می‌شوند . دلیل نامگذاری $m$ به «اندازه نمونه متناظر» این است که فرمول (۷,۲,۱) می‌تواند به این صورت تعبیر شود که $n$ مشاهده واقعی را با $m$ نمونه مجازی که با احتمال $p$ توزیع شده‌اند ترکیب می‌کنیم.

## ۸- الگوریتم EM:

به طور کلی الگوریتم EM در مسائلی به کار می رود که می خواهیم مجموعه ای از پارامترهای $\theta$ را تخمین بزنیم که مبتنی بر یک توزیع احتمالی اند و تنها بخشی از داده هایی که توسط این توزیع احتمالی تولید شده اند را داریم نه کل آنها را . مجموعه داده های مشاهده شده را به صورت $X=\{x_1,x_2,...,x_m\}$ که مستقل از همند و مجموعه داده های نامرئی را به صورت $Z=\{z_1,z_2,...,z_m\}$ و کل داده ها را به صورت $Y=X\bigcup Z$ در نظر می گیریم . الگوریتم EM فرض با بیشینه نزدیکی (ML) $h'$ می گردد که $P(Y\mid h')$ را بیشینه کند . این امید ریاضی روی توزیع احتمال Y که توسط پارامترهای معلوم $\theta$ اندازه گیری می شود . الگوریتم EM برای تخمین توزیع احتمالی روی Y به جای پارامترهای واقعی $\theta$ از فرض فعلی $h$ استفاده می کند . تابع $Q(h\mid h')$ را به گونه ای تعریف می کنیم که $E(\ln P(Y\mid h'))$ را به صورت تابعی از $h'$ تعریف کند . با فرض $\theta=h$ و اینکه داده های X از کل داده های Y مشاهده شده اند ، داریم :

| | |
|---|---|
| $Q(h\mid h')=E[\ln p(Y\mid h')\mid h,X]$ | (۸,۱) |

مرحله ۱ : مرحله تخمین (E) . با استفاده از مقدار فرض فعلی $h$ و داده های مشاهده شده X مقدار $Q(h\mid h')$ را محاسبه می کند تا توزیع احتمالی روی Y را تخمین زند .

مرحله ۲ : مرحله بیشینه سازی (M) . فرض $h$ را با فرض $h'$ ای جایگزین می کند که تابع Q را بیشینه کند .

| | |
|---|---|
| $h \leftarrow \arg\max_{h'} Q(h' \mid h)$ | (۸,۲) |

**۸,۱ تولید الگوریتم K-Mean از الگوریتم EM :**

**مسئله** K-Mean **عبارتست از تخمین پارامتر های** $\theta = \langle \mu_1, \mu_2, ..., \mu_k \rangle$ **. این پارامترها میانگین** K **توزیع نرمال را تعریف می کنند . مجموعه داده های مشاهده شده** $X = \{\langle x_i \rangle\}$ **و داده های نامرئی** $Z = \{\langle z_{i,1}, z_{i,2}, z_{i,3}, ..., z_{i,k} \rangle\}$ **را در در نظر بگیرید . داده های نامرئی مشخص می کنند که کدام یک از** K **توزیع نرمال برای تولید** $x_i$ **به کار رفته اند . برای استفاده از الگوریتم** EM **برای عبارت** $Q(h \mid h')$ **باید مقداری مناسب با مسئله** K-Mean **بیابیم . احتمال یک نمونه** $y_i = \langle x_i, z_{i,1}, z_{i,1}, ..., z_{i,k} \rangle$ **از کل داده ها می تواند به صورت زیر نوشته شود :**

| | |
|---|---|
| $p(y_i \mid h') = p(x_i, z_{i,1}, z_{i,1}, ..., z_{i,k} \mid h') = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}\sum_{j=1}^{k} z_{ij}(x-\mu'_j)^2}$ | (۸,۱,۱) |

**تنها یکی از** $z_{ij}$ **می تواند ۱ باشد وبقیه باید ۰ باشند . لذا این عبارت توزیع احتمال تولید** $x_i$ **توسط توزیع انتخاب شده را نشان می دهد . لگاریتم احتمال** $P(Y \mid h')$ **برای همه** m **نمونه به صورت زیر است :**

| | |
|---|---|
| $P(Y \mid h') = \ln \prod_{i=1}^{m} p(y_i \mid h')$ <br> $= \sum_{i=1}^{m} \ln p(y_i \mid h')$ <br> $= \sum_{i=1}^{m} (\ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2} \sum_{j=1}^{k} z_{ij}(x_i - \mu'_j)^2)$ | (۸,۱,۲) |

اکنون باید امید ریاضی $P(Y|h')$ را روی توزیع احتمال Y یا به طور معادل روی قسمتهای نامرئی Y یعنی مقادیر $z_{ij}$ حساب کنیم . عبارت بالا برای $P(Y|h')$ تابع خطی از مقادیر $z_{ij}$ است . لذا خواهیم داشت :

$$E[P(Y|h')] = E[\sum_{i=1}^{m}(\ln\frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}\sum_{j=1}^{k} z_{ij}(x_i-\mu'_j)^2)]$$
$$= \sum_{i=1}^{m}(\ln\frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}\sum_{j=1}^{k} E[z_{ij}](x_i-\mu'_j)^2)$$

(۸,۱,۳)

لذا تابع $Q(h|h')$ به صورت زیر در خواهد آمد :

$$Q(h'|h) = \sum_{i=1}^{m}(\ln\frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}\sum_{j=1}^{k} E[z_{ij}](x_i-\mu'_j)^2)$$

(۸,۱,۴)

که $h' = \langle \mu'_1, \mu'_2, ..., \mu'_k \rangle$ فرض جدید (بهبود یافته) است . $E[z_{ij}]$ بر اساس مقادیر فعلی h و مقادیر مشاهده شده X محاسبه می شود .

$$E[z_{ij}] = \frac{e^{-\frac{1}{2\sigma^2}(x_i-\mu_j)^2}}{\sum_{n=1}^{k} e^{-\frac{1}{2\sigma^2}(x_i-\mu_n)^2}}$$

(۸,۱,۵)

مرحله دوم الگوریتم EM مقادیر $\langle \mu'_1, \mu'_2, ..., \mu'_k \rangle$ را به گونه ای تعیین می کند که مقدار تابع Q بیشینه شود :

| | |
|---|---|
| $\arg\max_{h'} Q(h' \mid h) = \arg\max_{h'} \sum_{i=1}^{m} (\ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2} \sum_{j=1}^{k} E[z_{ij}](x_i - \mu'_j)^2$ <br> $= \arg\min_{h'} \sum_{i=1}^{m} \sum_{j=1}^{k} E[z_{ij}](x_i - \mu'_j)^2$ | (۸,۱,۶) |

لذا فرض ML (بیشینه کردن نزدیکی) به صورت کمینه کردن مجموع مربعات خطاها در می آید .

مقدار فرمول وقتی کمینه می شود که هر $\mu_j$ به صورت زیر مقدار دهی شود :

| | |
|---|---|
| $\mu_j = \frac{\sum_{i=1}^{m} E[z_{ij}]x_i}{\sum_{i=1}^{m} E[z_{ij}]}$ | (۸,۱,۷) |

## مراجع


[۱] Buntine W. L, (۱۹۹۴). Operations for learning with graphical models. *Journal of Artificial Intelligence Research*, ۲, ۱۵۹-۲۲۵.

[۲] R. Agrawal, J. Gehrke, D. Gunopolos, and Prabhakar Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In *ACM SIGMOD Conference*, ۱۹۹۸.

[۳] R. Neal and G. Hinton. A view of the EM algorithm that justifies incremental, sparse and other variants. Technical report, Dept. of Statistics, University of Toronto, ۱۹۹۳.

[۴] PELLEG, D. and MOORE, A. ۲۰۰۰. X-means: Extending K-means with Efficient Estimation of the Number of Clusters. In Proceedings ۱۷th ICML, Stanford University.

[۵] LEE, C-Y. and ANTONSSON, E.K. ۲۰۰۰. Dynamic partitional clustering using evolution strategies. In Proceedings of the ۳rd Asia-Pacific Conference on Simulated Evolution and Learning, Nagoya, Japan.

[۶] Casella, G., & Berger, R. L. (۱۹۹۰). *Statistical inference* . Pacific Grove, CA: Wadsworth & Brooks/Cole.

[۷] L. Xu and M. Jordan. On convergence properties of the EM algorithm for Gaussian mixtures. *Neural Computation*, ۷, ۱۹۹۵.

[۸] Cédric Archambeau, John A. Lee, Michel Verleysen. On Convergence Problems of the *EM* Algorithm for Finite Gaussian Mixtures. ESANN'۲۰۰۳ proceedings – European Symposium on Artificial Neural Networks Bruges (Belgium), ۲۳-۲۵ April ۲۰۰۳, d-side publi., ISBN ۲-۹۳۰۳۰۷-۰۳-X, pp. ۹۹-۱۰۶

[۹] P. Langley, W. Iba, and K. Thompson. An Analysis of Bayesian Classifiers. Proc. ۱۰th Nat. Conf. on Artificial Intelligence, ۲۲۳–۲۲۸, AAAI Press and MIT Press, USA ۱۹۹۲

[۱۰] P. Langley and S. Sage. Induction of Selective Bayesian Classifiers. Proc. ۱۰th Conf. on Artificial Intelligence, ۱۹۹۴

[۱۱] B.W. Silverman: Density Estimation for Statistics and Data Analysis. Chapman and Hall, London (۱۹۸۶).

[۱۲] E. Parzen: On Estimation of a Probability Density Function and Mode. Annals of Math. Statistics, ۳۳, ۱۰۶۵–۱۰۷۶ (۱۹۶۲).

[۱۳] J. Bilmes: A Gentle Tutorial of the EM Algorithm and its Application to Parameter stimation for Gaussian Mixture and Hidden Markov Models. Technical Report of the International Computer Science Institute, Berkeley, CA (۱۹۹۸).

[۱۴] Adwait Ratnaparkhi. ۱۹۹۸. *Maximum Entropy Models for Natural Language Ambiguity Resolution*. Ph.D. thesis, the University of Pennsylvania.

[۱۵] Robert E. Schapire and Yoram Singer. ۲۰۰۰. Boostexter: A boosting-based system for text categorization. *Machine Learning*, ۳۹(۲/۳):۱۳۵–۱۶۸.

[۱۶] Andrew McCallum and Kamal Nigam. ۱۹۹۸. A comparison of event models for naive bayes text classification. *In AAAI-۹۸ Workshop on Learning for Text Categorization*.

[۱۷] Thorsten Joachims. ۱۹۹۹. Transductive inference for text classification using support vector machines. In *Proc. ۱۶th International Conf. on Machine Learning*, pages ۲۰۰–۲۰۹. Morgan Kaufmann, San Francisco, CA.