سمینار توزیع های آمیخته :   شنبه ۹۴/۹/۸ (ثبت اسم)

۱- معرفی ساختار توزیع های آمیخته :

* تعریف : هر ترکیب محدب به صورت $\sum_{i=1}^{k} p_i f_i(x)$ که در آن $\sum_{i=1}^{k} p_i = 1$ و f_i ها هر توزیعی باشد، یک توزیع آمیخته است.

← هم چنین یک توزیع آمیخته پیوسته به صورت $g(x)=\int_{\Theta} f(x|\theta)\,h(\theta)\,d\theta$ بیان می شود.

← در واقع توزیع های آمیخته فرم سلسله مراتبی دارند :

گسسته :

$$y|w_i \sim f(y|w_i) \equiv N(w_i,1)$$

$$w_i \sim Multinomial(1,p_1,\dots,p_k)$$

$$g(y)=\sum_{i=1}^{k} p_i\, f(y|w_i)$$

مثلاً توزیع آمیخته گاما-پواسن : (پیوسته)

$$y|w \sim poisson(\lambda w)$$

$$w \sim gamma(3,2)$$

$$f(y)=\int h(y|w)\,g(w)\,dw$$

← اگر توزیع w پیوسته باشد ← آمیخته پیوسته

اگر توزیع w گسسته باشد ← آمیخته گسسته ← دو امکان وجود دارد :

finite mixture ①

infinite mixture ②

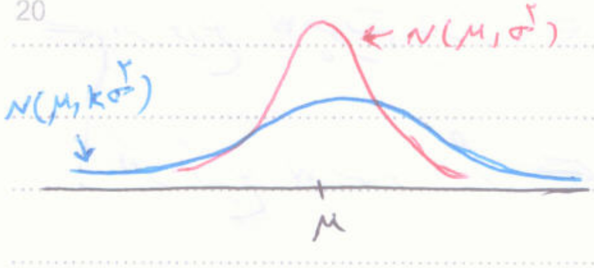
۲- اهمیت و کاربرد توزیع های آمیخته :

۲-۱- تحلیل خوشه بندی برمبنای درستنمایی : وقتی جمعیت را به k دسته تقسیم می‌کنیم که داخل دسته‌ها خصوصیت مشابه و بین دسته‌ها خصوصیات متفاوت باشد (مثلاً واریانس داخل دسته‌ها کم و واریانس بین دسته‌ها زیاد است)

۲-۲- مدل بندی outlier : یکی از راه‌های مدل بندی و توجیه outlier در مدلسازی استفاده از mixture هاست. مثلاً فرض کنید توزیع داده‌های ما $N(\mu,\sigma^2)$ است ولی تعدادی outlier داریم: $0<p_2<1$ را کوچک در نظر می‌گیریم و به یکی از دو صورت زیر عمل می‌کنیم :

روش اول | فرض می‌کنیم توزیع داده‌ها به جای $N(\mu,\sigma^2)$ به صورت زیر است :

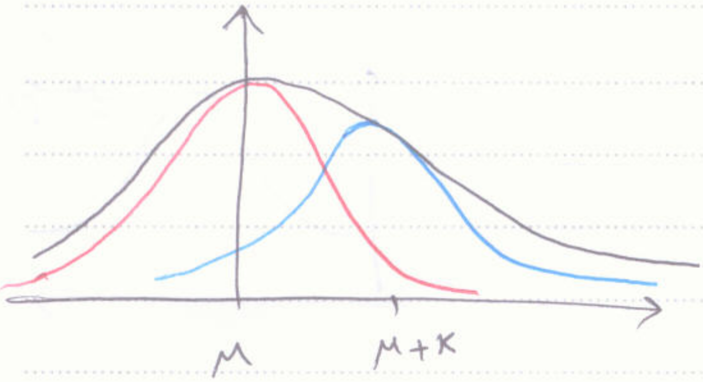
$$Y \sim (1-p_2)N(\mu,\sigma^2) + p_2 N(\mu,k\sigma^2) \qquad k>>1$$



چون تعداد outlier ها کم بوده، می‌گیم احتمال کمی داده‌ها از توزیع $N(\mu,k\sigma^2)$ می‌آیند و با احتمال زیادی از $N(\mu,\sigma^2)$ می‌آیند.

$\boxed{\text{روش دوم}}$ فرض کنیم توزیع داده ها به جای $N(\mu,\sigma^2)$ به صورت زیر است:

$$Y \sim (1-\epsilon_2)\, N(\mu,\sigma^2) + \epsilon_2\, N(\mu+k,\sigma^2)$$



← در واقع outlier های من فقط یک طرف هستند.

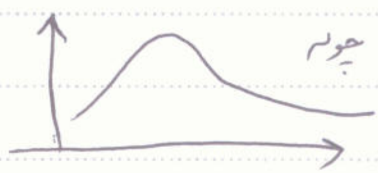
۲-۳- در مدل بندی داده های گمشده:

← در pattern-mixture model

۲-۴- در density Estimation: از آنجایی که هر توزیعی را می توان با آمیخته k بعدی از نرمال (k زیاد) تقریب زد، می توان از آمیخته ای از توزیع های نرمال، برای برآورد چگالی داده ها استفاده کرد.

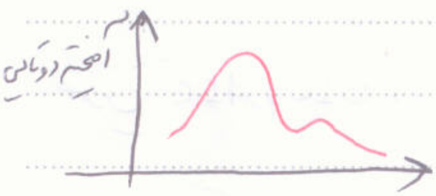
۳- ویژگی ها و گشتاورهای توزیع های آمیخته:

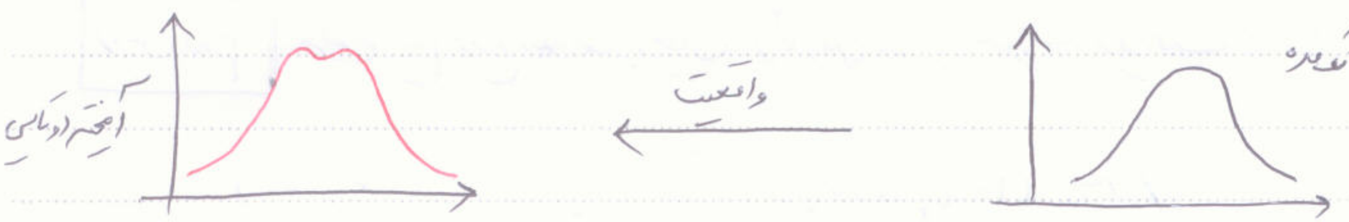
۳-۱- چند مد بودن: گاهی هیستوگرام یک سری داده را رسم می کنیم که:

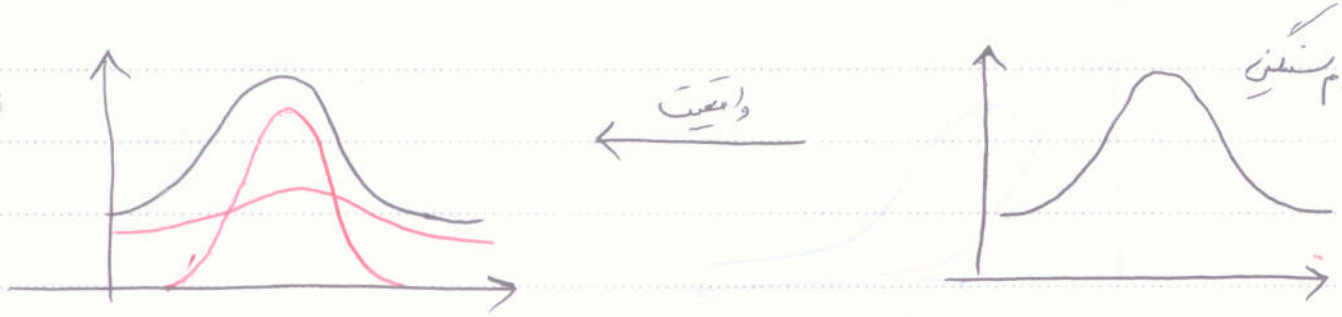


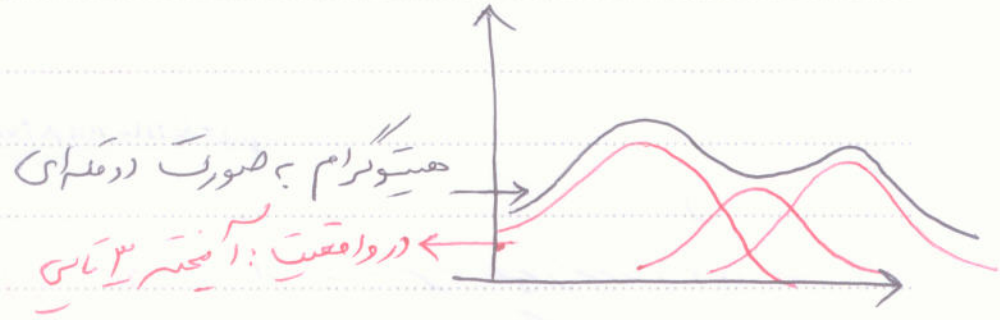
در صورتیکه در واقعیت ← به صورت روبه رو است

(یعنی میشه این مدل را برازش داد.)









(← گاهی هم یک توزیع مثلاً k مده داریم ولی وقتی رسمش می‌کنیم می‌بینیم k^* تا قله داره (یعنی برعکس حالت قبل) ⇐ پس تعداد قله نشان دهنده تعداد k در آمیخته‌ها نیست)

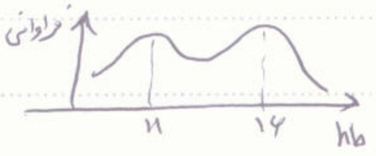
⇐ با رسم هیستوگرام نمی‌تونیم تعداد k را تشخیص دهیم. در مورد انتخاب k، راه‌های زیادی وجود دارد که ساده‌ترین آن، برازش مدل با kهای مختلف و مقایسه آنها از طریق AIC و BIC و ... است

(ولی تقریب با آمیخته‌ای از نرمال‌ها، راه مناسب‌تری است و می‌تونیم AIC و BIC و ... این مدل هم، با قبلی‌ها مقایسه کنیم.)

* اما وقتی هیستوگرام داده های ما چندقله‌ای می‌شود، چه دلیلی داشته

① گاهی دلیل چندقله‌ای بودن مشخص است: مثلاً مثال هموگلوبین خون.

وقتی هیستوگرام داده ها را رسم می‌کنیم به صورت



است یعنی

به تعدادی حول هموگلوبین ۱۱ و به تعدادی حول ۱۶ متمرکز شده اند. با استفاده از اطلاعات پزشکی مشخص می‌شود که هموگلوبین در زنان کمتر از مردان است و دلیل دوقله‌ای بودن این هیستوگرام جنسیت است.

← حالا که دلیل دوقله‌ای بودن مشخص است، در هر زیرگروه، تحلیل های جداگانه صورت می‌گیرد. مثلاً در گروه زنان توزیع را $N(\mu_1,\sigma_1^2)$ و در گروه مردان توزیع را $N(\mu_2,\sigma_2^2)$ درنظر می‌گیریم و p_1 و $1-p_1$ هم که مشخص است.

② دلیل چندقله‌ای بودن مشخص نیست: در این صورت p_i ها نامشخص اند و نیاز به برآوردیابی داریم که در بخش برآوردیابی به آن اشاره می‌کنیم.

۳-۲ - بدست آوردن گشتاورهای توزیع های آمیخته:

$$Y \sim f(y) = \sum_{k=1}^{K} \eta_k h_j(y)$$

$$E(H(Y)) = \sum_{k=1}^{K} \eta_k E(H_k(Y_k))$$

E(H(Y)) در زیرجامعه kام

اثبات: $E(H(Y)) = \int_{-\infty}^{+\infty} H(y) f(y)\, dy$

$$= \int_{-\infty}^{+\infty} H(y) \sum_{k=1}^{K} \eta_k h_k(y)\, dy$$

(برای کشاهی به جابجایی $\sum$ و $\int$) $= \sum_{k=1}^{K} \eta_k \int_{-\infty}^{+\infty} H_k(y)\, dy$

$$= \sum_{k=1}^{K} \eta_k E_k(H(Y)) \quad \square$$

$$\Rightarrow \mu = E(Y) = \sum_{k=1}^{K} \eta_k \underbrace{E_k(Y_k)}_{\mu_k = \text{میانگین زیرجامعه kام}}$$

$$E((Y-\mu)^i) = \sum_{k=1}^{K} \eta_k E_k(Y \overset{\pm\mu_k}{-} \mu)^i$$

$$= \sum_{k=1}^{K} \sum_{r=0}^{i} \binom{i}{r} (\mu_k - \mu)^{i-r} E_k(Y - \mu_k)^r \eta_k$$

⇐ زمانیکه ۲ = i ، (یعنی واریانس توزیع آمیخته) داریم:

$$\sigma^2 = E(Y-\mu)^2 = \sum_{k=1}^{K} \eta_k \left((\mu_k - \mu)^2 + \sigma_k^2 \right)$$

واریانس زیرجامعه kام $= E_k(Y-\mu_k)^2$

۴- برآورد بیزی:
→ رویکرد کلاسیک
→ رویکرد بیز

۴-۱- رویکرد کلاسیک: ما فقط آمیخته‌ای از توزیع‌های نرمال را در نظر می‌گیریم.

$$Y \sim f(y) = \sum_{k=1}^{K} p_k \, N(\mu_k , \sigma_k^2)$$

پارامترهای مدل $= \underline{\theta} = (\mu_1, \dots, \mu_K, \sigma_1^2, \dots, \sigma_K^2, p_1, \dots, p_{K-1})$

① برآوردگشتاوری: که فقط در حالاتی خاص جواب می‌دهد و ما هم آن را بیان نمی‌کنیم.

② برآورد ML: که فقط برای آمیخته‌ای از دو نرمال آن هم در حالتی که $\sigma_1^2 = \sigma_2^2$ بوده است بدست آمده است ⇐ بنابراین در حالت کلی از الگوریتم EM استفاده می‌کنیم:

گام نخست) بدست آوردن درستنمایی پارامترها: $Y_1, \dots, Y_N \sim f(y)$

$$p(y|\theta) = L(\theta|y) = \underbrace{\prod_{i=1}^{N} \sum_{k=1}^{K} p_k \, N(y|\mu_k , \sigma_k^2)}_{\text{نمی‌توان کاری کرد}}$$

باید متغیری که نشان‌گر تعلق به هر زیر جامعه باشد، وجود داشته باشد. ولی چون این متغیر مشاهده نشده است به آن latent variable می‌گویند. برای سهولت محاسبات آن را وارد درستنمایی می‌کنند و الگوریتم EM بر مبنای latent variable است.

الگوریتم EM : Expectation-maximization Algorithm

یک متغیر تصادفی گنگ W که $w_i \in \{1, 2, \dots, k\}$ (k fix) و مشخص می‌کند y_i از کدام گروه آمده است، در نظر می‌گیریم. در این حالت درستنمایی داده‌های کامل (یعنی (y, w)) به صورت زیر است:

$$P(y, w|\theta) = L(\theta|y, w) = P(y|w, \theta)\, P(w|\theta)$$ ← هم سلسله مراتبی خودش را نشون میده

دقت کنیم که هر y_i می‌تونه فقط متعلق به یک زیرگروه باشد. بنابراین:

w_i	1	2		k
$P(W_i = w_i)$	$P(w_i=1)$	$P(w_i=2)$	-	$P(w_i=k)$

$\xrightarrow{\text{هم بسته}}$ $f_{W_i}(w_i) = \prod_{k=1}^{k} \left(P(W_i = k)\right)^{I(w_i = k)}$ $k = 1, 2, \dots, k$

$$P(y, w|\theta) = f(y|w, \theta)\, P(w|\theta)$$

$$= \prod_{i=1}^{N} f(y_i | w_i, \theta)\, f(w_i|\theta)$$

$$= \prod_{i=1}^{N} \prod_{j=1}^{k} \left\{ N(y_i | \mu_j, \sigma_j^2)\, P(w_i = j) \right\}^{I(w_i = j)}$$

چرا نرمال هم توان $I(w_i = j)$ می‌گیرد؟

دلیل: $y_i | w_i : \prod_{j=1}^{k} \left\{ N(\mu_j, \sigma_j^2) \right\}^{I(w_i=j)}$

$$\Rightarrow \ell(\theta|y, w) = \sum_{i=1}^{N} \sum_{j=1}^{k} \left(I(w_i = j)\right) \ln N(y_i | \mu_j, \sigma_j^2)\, \pi_j$$

در برآورد ML ما درستنمایی را max می‌کنیم ولی در EM ما امید درستنمایی (برحسب W) را ماکزیمم می‌کنیم. چون W تصادفی است از درستنمایی برحسب W امید می‌گیریم تا از شر تصادفی بودن W خلاص شویم و سپس آن امید را max می‌کنیم.

الگوریتم EM یک روش تکراری است. پس با نقطه شروع اولیه $\underline{\theta}^{(0)}$ شروع می‌کنیم.

$$\underline{\theta}^{(0)} = (\mu_1^{(0)}, \dots, \mu_k^{(0)}, \sigma_1^{2(0)}, \dots, \sigma_k^{2(0)}, \eta_1^{(0)}, \dots, \eta_{k-1}^{(0)})$$

و الگوریتم EM شامل دو step است:

E-Step: $E_W(\ell(\theta|y,w))$

M-Step: $Q(\theta,\theta^{(p-1)}) = E_W\left[\ell(\theta|y,w) \,\middle|\, y,\theta^{(p-1)}\right]$

$$= \sum_{w\in S_w} p(w|y,\theta^{(p-1)}) \log p(y,w|\theta)$$

$$\theta^{(p)} = \arg\max_{\theta} Q(\theta,\theta^{(p-1)})$$

این الگوریتم تا جایی ادامه پیدا می‌کند که اختلاف $\theta^{(p)}$ و $\theta^{(p-1)}$ کم شود.