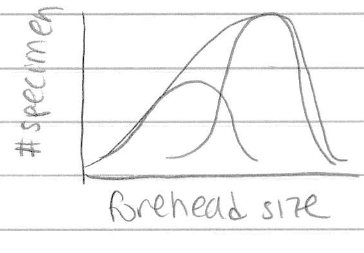
Mixture Models

Introduced by Karl Pearson in 1894

Naples crab:



Are there two species? Is the distribution a mixture of two Gaussians?

Recall: A Gaussian has pdf

$$\mathcal{N}(\mu, \sigma^2, x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

A mixture has pdf

$$F(x) = w\,\mathcal{N}(\mu_1, \sigma_1^2, x) + (1-w)\,\mathcal{N}(\mu_2, \sigma_2^2, x)$$

where $0 \leq w \leq 1$ is the mixing weight

Interpretation: Flip a biased coin to determine which component sample comes from

Other applications include modeling height, velocities in gasses etc

Pearson invented the <u>method of moments</u> to attack the learning problem

<u>def:</u> Let $M_r = \mathbb{E}_{x \in F(x)}[x^r]$

<u>Fact:</u> M_r is a polynomial in the unknown parameters, i.e. $(w, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$

In particular

① $M_1 = w\mu_1 + (1-w)\mu_2$

② $M_2 = w(\mu_1^2+\sigma_1^2) + (1-w)(\mu_2^2+\sigma_2^2)$

③ $M_3 = w(\mu_1^3 + 3\mu_1\sigma_1^2) + (1-w)(\mu_2^3 + 3\mu_2\sigma_2^2)$

etc

Let $\tilde{M}_r = \frac{1}{|S|}\sum_{i \in S} x_i^r$ denote empirical avgs

samples

<u>Sixth Moment Test</u>

- Given samples S, compute $\tilde{M}_r$ for $r = 1$ to 6
- Solve for simultaneous roots of

$$\{M_r(w, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2) = \tilde{M}_r\}_{r=1 \text{ to } 5}$$

- Among all valid solns, choose the one that is closest in sixth moment

Main Questions:

① Is it stable to sampling noise?

② Do the first six moments uniquely determine the parameters?

Milestones

Pearson (1894): method of moments (no guarantees)

Fisher (1912-1922): maximum likelihood estimator

$$\hat{\theta}_{MLE} = \operatorname{argmax}\, p(x_1, \ldots, x_s; \theta)$$

Consistent and asymptotically efficient, usually computationally hard

Teicher (1961): identifiability

i.e. if we knew the density function exactly, it determines the parameters

Proof [sketch] The component with the largest variance dominates the behavior of $F(x)$ in the tails.

Find its mean, variance and mixing weight, subtract it off from $F(x)$ and proceed. $\boxed{\text{Z}}$

Intuitively, this requires tons of samples

Dempster, Laird, Rubin (1977): expectation maximization

① initial guess $(\hat{\omega}, \hat{\mu}_1, \hat{\sigma}_1^2, \hat{\mu}_2, \hat{\sigma}_2^2)$

② Iterate:

Cluster: For each $x \in S$, calculate posterior

$$P_x = \frac{\hat{\omega}\, \mathcal{N}(\hat{\mu}_1, \hat{\sigma}_1^2, x)}{\hat{\omega}\, \mathcal{N}(\hat{\mu}_1, \hat{\sigma}_1^2, x) + (1-\hat{\omega})\, \mathcal{N}(\hat{\mu}_2, \hat{\sigma}_2^2, x)}$$

Update parameters

$$\hat{\omega} \leftarrow \frac{\sum_{x \in S} P_x}{|S|} \; ; \; \hat{\mu}_1 \leftarrow \frac{\sum_{x \in S} P_x\, x}{\hat{\omega}}$$

$$\hat{\sigma}_1^2 \leftarrow \frac{\sum_{x \in S} P_x (x - \hat{\mu}_1)^2}{\hat{\omega}}$$

and similarly for $\hat{\mu}_2, \hat{\sigma}_2^2$

This is a heuristic to maximize likelihood, but often gets stuck

---

Learning via Clustering

Dasgupta gave the first provable guarantees

Claim: If we can accurately cluster the samples into which component generated them, can estimate the parameters

But how do you cluster?

Let's start with some intuitive and counter-intuitive properties of high-dimensional Gaussians

Recall: $$\mathcal{N}(\mu, \Sigma, x) = \frac{e^{-\frac{(x-\mu)^T \Sigma^{-1} (x-\mu)}{2}}}{(2\pi)^{d/2} \det(\Sigma)^{1/2}}$$

Fact #1: $\mathcal{N}(\mu, \Sigma, x)$ is maximized at $x = \mu$

Fact #2: For $x \sim \mathcal{N}(\mu, \sigma^2 I, x)$

$$\mathbb{P}\left[\left| \|x-\mu\|^2 - \sigma^2 d \right| \geq c\,\sigma^2 \sqrt{d \log d} \right] \leq d^{-\frac{c^2}{4}}$$

How can these facts simultaneously be true?

The growth of the volume of the ball counteracts the decay in the pdf as we move away from μ

Sketch of Fact #2: First we note if

$x \sim \mathcal{N}(\mu, \sigma^2)$ then $bx + a \sim \mathcal{N}(b\mu + a, b^2\sigma^2)$

Now consider

i^{th} coordinate of random X

$$\sum_{i=1}^{d} z_i^2 \quad \text{where } z_i \triangleq \frac{(x_i - \mu_i)}{\sigma}$$

Then $\displaystyle\sum_{i=1}^{d} z_i^2 = \frac{\|x-\mu\|^2}{\sigma^2}$

Now each $z_i \sim \mathcal{N}(0,1)$ and $\displaystyle\sum_{i=1}^{d} z_i^2$ is called a χ^2-distribution

It has an explicit expression for its pdf, but all we need is

$$\frac{\sum_{i=1}^{d} z_i^2 - d}{2\sqrt{d}} \xrightarrow{\lim d \to \infty} \mathcal{N}(0,1)$$

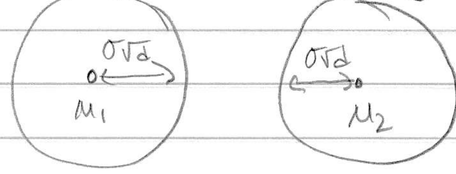
Hence $\displaystyle\sum_{i=1}^{d} z_i^2 \to \mathcal{N}(d, 4d)$

$$\Rightarrow \sum_{i=1}^{d} (x_i - \mu_i)^2 = \|x-\mu\|^2 \to \mathcal{N}(\sigma^2 d, 4\sigma^4 d)$$

Finally $\mathbb{P}\left[\left|\|x-\mu\|^2 - \sigma^2 d\right| > c\,\sigma^2\sqrt{d \ln d}\right]$

$$\lesssim e^{-\frac{c^2\sigma^4 d \ln d}{4\sigma^4 d}} = d^{-\frac{c^2}{4}} \quad \blacksquare$$

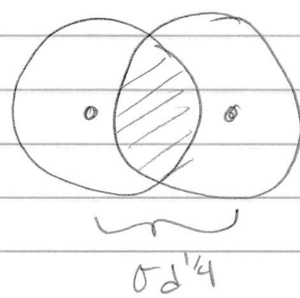
Now back to clustering: If $\|\mu_1-\mu_2\| \gg \sigma\sqrt{d}$



we should be able to cluster [Dasgupta]
[Arora, Kannan]

<u>Proposition</u>: If $\|\mu_1-\mu_2\| \gg d^{1/4}\sigma \ln^{1/2} d$ then whp all samples from first component are closer to each other than to any sample from second component and vice-versa

How can this be? Pictorially



The measure of the overlap region is negligible

<u>Proof</u>: Consider $a, a' \sim N(\mu_1, \sigma^2 I)$ and $b \sim N(\mu_2, \sigma^2 I)$. Then whp the vectors

$$a-\mu_1,\ a'-\mu_1,\ \mu_1-\mu_2 \text{ and } b-\mu_2$$

are nearly orthogonal. This follows b/c three of them are random, so pairwise inner-products are small

Now we can compute

$$\|a-a'\|^2 = \|a-\mu_1+\mu_1-a'\|^2$$

$$= \|a-\mu_1\|^2 + \|\mu_1-a'\|^2 + 2\langle a-\mu_1, \mu_1-a'\rangle$$

$$\quad ① \qquad\qquad ② \qquad\qquad ③$$

Now ① and ② are each $\sigma^2 d \pm c\sigma^2\sqrt{d\ln d}$ and ③ is $\frac{\sigma^2 d}{\sqrt{d}}$ negligible

Thus $\|a-a'\|^2 = 2\sigma^2 d \pm 2c\sigma^2\sqrt{d\ln d}$

Similarly we have

$$\|a-b\|^2 = \|a-\mu_1+\mu_1-\mu_2+\mu_2-b\|^2$$

$$= \|a-\mu_1\|^2 + \|\mu_1-\mu_2\|^2 + \|\mu_2-b\|^2$$

$$\pm \text{ lower order terms}$$

$$= 2\sigma^2 d \pm 2c\sigma^2\sqrt{d\ln d} + \sigma^2\sqrt{d}\ln d$$

Hence we have

$$\|a-b\|^2 \geq \|a-a'\|^2 + \sigma^2\sqrt{d}\ln d \quad \text{whp}$$

$\boxed{Z}$

This gives a polynomial running time / sample complexity algorithm at separation $\gtrsim d^{1/4}$

Is this the best we can do?

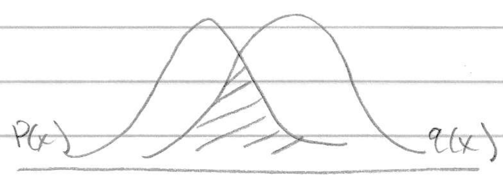
def. The total variation distance btwn pdfs $P(x)$ and $q(x)$ is

$$d_{TV} \triangleq \frac{1}{2} \int |P(x) - q(x)| \, dx$$

Fact: Clustering $\omega(1)$ samples with $w = \frac{1}{2}$ requires

$$d_{TV}(\mathcal{N}(\mu_1, \Sigma_1), \mathcal{N}(\mu_2, \Sigma_2)) \geq 1 - o(1)$$

Proof: we can couple samples from the two distributions, e.g.



Throw darts and output samples if they are below the pdf.

first dart below $p(x) \sim p(x)$
first dart below $q(x) \sim q(x)$

Now to sample from the mixture

① throw a dart

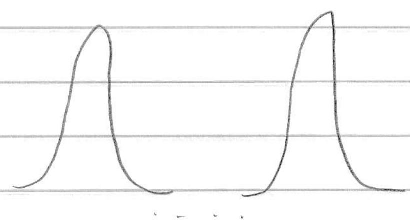
② flip a ~~biased~~ coin. on heads, output the sample if it's below $p(x)$. on tails output the sample if it's below $q(x)$

Notice that if the dart lands in the overlap region, which has area $1-d_{TV}$, then it could have come from either. ☒

For what separation is $d_{TV} = 1-o(1)$?

This holds even when $\|\mu_1 - \mu_2\| >> \sigma\sqrt{\ln d}$, and this is tight

e.g. if we could project on the line connecting μ_1 and μ_2 we'd get



Main Question: How can we find the right directions to project on?

Following [Vempala, Wang], let

$$M = \mathbb{E}_{x \sim F(x)} [x x^T]$$

Lemma: Let $u_1, \ldots, u_k$ be the top k singular vectors of M. If $F(x)$ is a mixture of k spherical Gaussians with linearly independent means $\mu_1, \ldots, \mu_k$ then

$$\mathrm{span}(u_1 \dots u_k) = \mathrm{Span}(\mu_1 \dots \mu_k)$$

Proof: we can write $x = c + z$ where

$$c = \begin{cases} \mu_1 & \text{w/ prob } w_1 \\ \vdots & \\ \mu_k & \text{w/ prob } w_k \end{cases} \quad \text{and } z \sim \mathcal{N}(0, \sigma^2 I)$$

Since $c \perp\!\!\!\perp z$ we have

$$\mathbb{E}[xx^T] = \underbrace{\mathbb{E}[cc^T]}_{\sum_{i=1}^{k} w_i \mu_i \mu_i^T} + \underbrace{\mathbb{E}[zz^T]}_{\sigma^2 I}$$

The variational characterization of singular values tells us

$$\sigma_{k+1}(M) = \min_{\dim(V)=k} \max_{u \perp V} \frac{u^T M u}{u^T u}$$

Hence $\sigma_{k+1}(M) = \sigma_{k+2}(M) = \dots = \sigma^2$

Thus all but the top k singular vectors must be $\perp$ $\mathrm{span}(\mu_1, \dots, \mu_k)$. $\boxtimes$

So if we estimate M well enough, we can reduce to a k-dimensional problem

$$\frac{\text{separation}}{d^{1/4} \sigma \sqrt{\log d}} \Longrightarrow \frac{\text{separation}}{k^{1/4} \sigma \sqrt{\log d}}$$

why is this d and not k? we still need to cluster $d^{O(1)}$ samples