Lecture 8

Review: Estimating the mean of a Gaussian using Scheffé (for $d=1$)

Step 1: Guesslist $\mathcal{L} = \{P_{\hat{\mu}},\ \hat{\mu} \in \{x_1, \dots, x_n\}\}$

$$\left(\text{Since } P\left(\exists\, i \text{ s.t. } |x_i - \mu| \le \sqrt{\frac{2\sigma^2}{n} \log \frac{2}{\epsilon}}\right) \ge 1 - \epsilon/2\right)$$

Step 2: Use Scheffé to find a $\hat{P} \in \mathcal{L}$ s.t.

$$d(\hat{P}, P) \le 9 \cdot \sqrt{\frac{1}{2n} \log \frac{2}{\epsilon}} + 8\Delta$$

Here Scheffé set $A(i,j)$ correspond to intervals of the form $(-\infty, a)$ or (a, ∞).

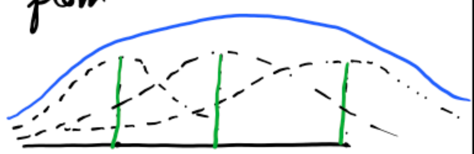
Thus, $\Delta \le d_{KS}(P_\mu, \underbrace{\mu_n}_{\text{empirical measure}}) \le O\left(\sqrt{\frac{1}{n} \log \frac{1}{\epsilon}}\right)$.

Agenda: * Learning Gaussian Mixtures

$\boxed{A}$ Learning the distribution vs learning parameters

Given samples $X_1, \dots, X_n$ generated iid from

$$\sum_{j=1}^{k} \omega_j N(\mu_j, \sigma_j^2 I_{d \times d}),$$



find:

(a) the weights ω, means μ, variances σ^2 (parameter learning)

(b) a mixture $\hat{P}$ that is close to P (proper learning of distribution)

(c) any distribution $\hat{P}$ that is close to P (improper learning)

We will go with (b).

A difficulty in (a) is identifying which samples are associated with which component. This can only be done if we assume sufficient separation between the means.

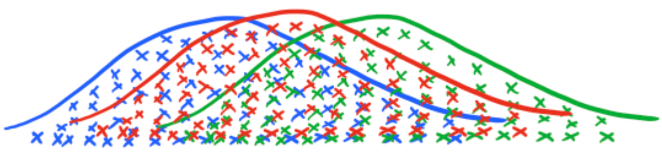
→ This difficulty can be circumvented in (b).

PAC formulation (We assume $\sigma_j^2 = \sigma^2$ is known)

$$\epsilon(n, \delta, d) = \min_{\hat{\theta}=(\hat{\omega},\hat{\mu})} \max_{\theta=(\omega,\mu)} P_\theta\left(d\left(P_\theta, P_{\hat{\theta}(x^n)}\right) > \delta\right)$$

$\boxed{B}$ The 1-dimensional case

→ Difficulty in using empirical mean estimation:



↓ seen as

which samples are coming from blue, red, and green

If we have many, many samples, we can detect subtle differences b/w histograms corresponding to different mixtures.

Large Sample + poly(n) computational complexity

⇓

infeasibility of the algorithm

→ A Scheffé based approach

Step (i) Form a guesslist

Note that

$$
\begin{aligned}
& d\left(P_{\omega,\mu}, P_{\omega',\mu'}\right) \\
& \leq \sum_{i=1}^{k} |\omega_i - \omega_i'| + \sum_{i=1}^{k} \omega_i \, d\left(P_{\mu_i}, P_{\mu_i'}\right) \\
& \leq \sum_{i=1}^{k} |\omega_i - \omega_i'| + \sum_{i=1}^{k} \frac{\omega_i |\mu_i - \mu_i'|}{2\sigma}
\end{aligned}
$$

Let $\Omega = \{ \omega : \quad \omega_i \in \{0, \frac{\delta}{2k}, \frac{2\delta}{2k} \dots, 1\}, \; 1 \leq i \leq k, \sum_{i=1}^{k} \omega_i = 1 \}$.

Our desired guesslist is then given by

$$\mathcal{L} = \{ P_{\omega,\mu} : \omega \in \Omega, \quad \mu_i \in \{X_1, \dots, X_n\} \; \forall \quad 1 \leq i \leq k \}.$$

Then,

$$
\begin{aligned}
& P_{\omega,\mu}\left(\forall \, \hat{P} \in \mathcal{L}, \quad d\left(\hat{P}, P_{\omega,\mu}\right) > \delta\right) \\
& \leq P_{\omega,\mu}\Big(\forall \; (\hat{\mu}_1, \dots, \hat{\mu}_k) \text{ s.t. } \hat{\mu}_i \in \{X_1, \dots, X_n\}, \\
& \qquad\qquad \sum_{j=1}^{k} \omega_j \, d\left(P_{\mu_j}, P_{\hat{\mu}_j}\right) > \frac{\delta}{2}\Big) \\
& \leq P_{\omega,\mu}\Big(\exists \; 1 \leq j \leq k \text{ s.t. } d\left(P_{\mu_j}, P_{X_i}\right) > \frac{\delta}{2} \text{ for} \\
& \qquad\qquad \text{every } 1 \leq i \leq n\Big) \\
& \leq \sum_{j=1}^{k} P_{\omega,\mu}\left(\forall \; 1 \leq i \leq n, \quad |\mu_j - X_i| > \sigma\delta\right)
\end{aligned}
$$

Denoting by N_j the number of samples from P_{μ_j},

$$
\begin{aligned}
& P_{\omega,\mu}\left(\forall \, 1 \leq i \leq n, |\mu_j - X_i| > \sigma\delta\right) \\
& \leq \sum_{\ell=0}^{n} P_{\omega,\mu}\left(N_j = \ell\right) e^{-\frac{\ell\delta^2}{2}} \leq P_{\omega,\mu}\left(N_j \leq t\right) + e^{-t\delta^2/8},
\end{aligned}
$$

for every t. We choose $t = n\omega_j \cdot \theta$, $\theta < 1$.

The right-side is bounded by (assuming $\omega_j \leq 1/2$)

$$e^{-c \cdot n \frac{\omega_j^2 (1-\theta)^2}{\omega_j}} + e^{-n\omega_j \theta \delta^2/8},$$

which is less than ϵ if $n = O\left(\frac{1}{\omega_j \delta^2} \log \frac{1}{\epsilon}\right)$.

Thus,

$$n = O\left(\frac{1}{\min_j \omega_j} \cdot \frac{1}{\delta^2} \cdot \log \frac{k}{\epsilon}\right)$$

suffice. We can improve by ignoring weights $\omega_i \leq \frac{\delta}{4k}$.

Then, our required prob. is bounded above by

$$\sum_{j \in [k],\ \omega_j > \delta/4k} P_{\omega,\mu}\left(\forall i \quad d(P_{\mu_j}, P_{x_i}) > \frac{\delta}{4}\right)$$

which can be bounded as above to get

$$n = O\left(\frac{k}{\delta^3} \log \frac{k}{\epsilon}\right) \text{ suffice.}$$

<u>Step (ii)</u> <u>Scheffé Selector</u>

Our list now contains $\left(\frac{k}{\delta}\right)^k \cdot \binom{n}{k} = O\left(\frac{k^{2k}}{\delta^{4k}} \log^k \frac{k}{\epsilon}\right)$

Also, note that the Scheffé sets $A(i,j)$ still appear to be intervals. Thus, we can bound Δ using $c \cdot d_{KS}(P_\mu, \mu_n)$.

$\boxed{C}$ <u>Can we use ML for selection?</u>

<u>The Yes Case</u> We can do this if we assume bounds for

the probabilities in the guesslist over the support of the unknown P.

<u>Theorem</u> Consider a set $\mathcal{L}$ of distributions on $\mathcal{X}$ with densities w.r.t. ν s.t. $\exists \hat{P} \in \mathcal{L}$ satisfying

$$D(P \| \hat{P}) \leq \delta .$$

Furthermore, assume $\alpha \leq Q(x)$ for all $x \in supp(P)$, for every $\hat{P} \in \mathcal{L}$. Denote by $\hat{P}_{ML}$ the maximizer of $\sum_{i=1}^{n} \ln \frac{1}{Q(X_i)}$ over the set $\mathcal{L}$. Then,

$$P\left(D(P \| Q_{ML}) > 4\delta \right) \leq (|\mathcal{L}|+1) \exp\left(- \frac{2n \cdot \delta^2}{\log^2(1/\alpha)} \right).$$

<u>Proof</u>. The function $\Lambda(Q) = -\sum_{i=1}^{n} \ln Q(X_i)$ is called the score function. Note that

$$\mathbb{E}_P[\Lambda(Q)] = n\left(H(P) + D(P \| Q)\right).$$

Note that $\hat{P}_{ML}$ will satisfy the required condition if

$$\Lambda(\hat{P}) < \min\{\Lambda(Q) : D(P \| Q) > 4\delta\},$$

which in turn will hold if

$\Lambda(\hat{P}) \leq n(H(P) + 2\delta)$ and

$\Lambda(Q) \geq n(H(P) + 3\delta)$ for every Q with $D(P \| Q) > 4\delta$.

We bound the complement of this latter event.

$$P\left(\Lambda(\hat{P}) > n\left(H(P)+2\delta\right)\right)$$
$$= P\left(\Lambda(\hat{P}) > n\left(H(P)+\delta\right) + n\delta\right)$$
$$\leq P\left(\Lambda(\hat{P}) > n\left(H(P)+D(P\|\hat{P})\right) + n\delta\right)$$
$$= P\left(\Lambda(\hat{P}) > \mathbb{E}_P[\Lambda(\hat{P})] + n\delta\right)$$
$$\leq \exp\left(-\frac{n\delta^2}{\log^2(1/\alpha)}\right) \quad \text{(using Hoeffding's inequality)}$$

Similarly, for a Q s.t. $D(P\|Q) > 4\delta$,

$$P\left(\Lambda(Q) \leq n\left(H(P)+3\delta\right)\right)$$
$$\leq P\left(\Lambda(Q) \leq \mathbb{E}_P[\Lambda(Q)] - n\delta\right)$$
$$\leq \exp\left(\frac{-n\delta^2}{\log^2(1/\alpha)}\right).$$

The claimed bound follows from the union bound. ■

The No example

$P_1 = \text{unif}[-1,1]$, $P_2 = \text{unif}[\delta, 1+\delta]$, $P = \text{unif}[0,1]$.

We see n samples from P.

Let $N \equiv \#$ of samples in $[0,\delta)$.

ML chooses P_1 if $N > 0$, which happens with prob.

$$P(N>0) = 1 - P(N=0) = 1-(1-\delta)^n$$

But $d(P_1, P) = 1/2$ and $d(P_2, P) = \delta \Rightarrow$ ML doesn't choose the closer P_i with prob. $\to 1$ as $n \to \infty$.