Lecture 9

Review * Scheffé selector for learning 1-dimensional mixtures of Gaussian:

→ $\Omega \equiv \epsilon/2k$ grid of the prob. simplex $\mathcal{P}_k$

$\mathcal{L} = \{ P_{\omega,\mu} : \omega \in \Omega, \ \mu_i \in \{X_1, \dots, X_n\} \text{ for all } 1 \le i \le k \}$

Use Scheffé selector to find the best match

* ML selector can also work if our guess-list satisfies more conditions, but fails in general.

Agenda * Learning d-dimensional Gaussian mixtures

[A] Distance based clustering for learning Gaussian mixtures in higher dimensions

Consider $X_1, \dots, X_n$ generated iid from

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

How does the distance b/w two samples generated from the same mean differ from that b/w two samples from a different mean?

If $X, Y \sim N(\mu, \sigma^2 I_{d\times d})$,

$$\mathbb{E} \|X-Y\|_2^2 = \sum_{i=1}^{d} \mathbb{E}[(X_i - Y_i)^2] = 2d\sigma^2$$

If $X \sim N(\mu_1, \sigma^2 I_{d\times d})$ and $Y \sim N(\mu_2, \sigma^2 I_{d\times d})$,

$$\mathbb{E}\,\|X-Y\|_2^2 = \sum_{i=1}^{d} \mathbb{E}\left[X_i^2 + Y_i^2 - 2X_iY_i\right]$$

$$= \sum_{i=1}^{d}\left(\sigma^2 + \mu_{1i}^2 + \sigma^2 + \mu_{2i}^2 - 2\mu_{1i}\mu_{2i}\right)$$

$$= 2d\sigma^2 + \|\mu_1 - \mu_2\|^2$$

Also, in the first case, using Chebyshev's inequality

$$P\left(\left|\,\|X-Y\|_2^2 - \mathbb{E}\left[\|X-Y\|_2^2\right]\right| > t\right) \le \frac{\mathrm{Var}\left(\|X-Y\|_2^2\right)}{t^2}$$

$$= \frac{d}{t^2}\,\mathrm{Var}\left((X_1-Y_1)^2\right)$$

$$= O\left(\frac{d}{t^2}\cdot\sigma^4\right).$$

Thus, in this case, with significant probability

$$\|X-Y\|_2^2 = 2d\sigma^2 \pm O\left(\sqrt{d}\,\sigma^2\right)$$

On the other hand, in the second case

$$P\left(\left|\,\|X-Y\|_2^2 - \mathbb{E}\left[\|X-Y\|_2^2\right]\right| > t\right)$$

$$\le \frac{1}{t^2}\sum_{i=1}^{d}\mathrm{Var}\left((X_i-Y_i)^2\right)$$

$$= \frac{1}{t^2}\sum_{i=1}^{d}\left[\left((\mu_{1i}-\mu_{2i})^4 + 12(\mu_{1i}-\mu_{2i})^2\sigma^2 + 12\sigma^4\right) - \left(2\sigma^2 + (\mu_{1i}-\mu_{2i})^2\right)^2\right]$$

$$= \frac{1}{t^2}\sum_{i=1}^{d} 8(\mu_{1i}-\mu_{2i})^2\sigma^2 + 8\sigma^4 = 8\frac{\sigma^2}{t^2}\left(\|\mu_1-\mu_2\|_2^2 + d\sigma^2\right)$$

Thus, in the second case, with large prob.

$$\|X-Y\|_2^2 = 2d\sigma^2 + \|\mu_1-\mu_2\|_2^2 \pm O\left(\sigma\left(\|\mu_1-\mu_2\|_2 + \sqrt{d}\,\sigma\right)\right)$$

Therefore, we can reliably distinguish two points from the same cluster from those coming from two different clusters iff

$$\|\mu_1-\mu_2\|_2^2 - O\left(\sigma\,\|\mu_1-\mu_2\|_2\right) \gg O\left(\sqrt{d}\,\sigma^2\right)$$

which holds if

$$\|\mu_1-\mu_2\|_2 > c\,d^{1/4}\sigma.$$

* The effort in this line of research is to seek algorithms that circumvent the requirement of $\|\mu_1-\mu_2\|_2$ to grow d.

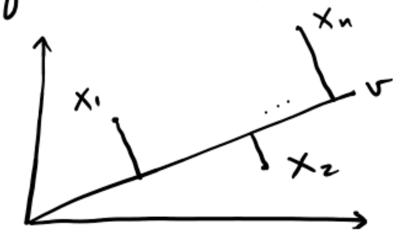
B Projection to a lower dimensional space

* Idea (Vempala-Wang '02) Do distance based clustering after projecting on a lower dimensional space.

Review of Singular Value Decomposition (SVD)

Chapter 3 of "Foundations of data science," Blum, Hopcroft and Kannan

(a) Best fitting line



Denote by A the $(n\times d)$ data matrix obtained by stacking the data vectors $X_i \in \mathbb{R}^d$ as rows.

We need to minimize $\sum_{i=1}^{n} \|X_i - (X_i \cdot v)\,v\|_2^2$ over v,

which by Pythagorean Theorem is the same as obtaining

$$v_1 = \underset{v \text{ s.t. } \|v\|_2 = 1}{\operatorname{argmax}} \sum_{i=1}^{n} (X_i \cdot v)^2 = \underset{v \text{ s.t. } \|v\|_2 = 1}{\operatorname{argmax}} \|Av\|_2^2$$

Denote $\sigma_1(A) = \|Av_1\|_2 \longrightarrow$ this corresponds to the spectral norm $\|A\|_2 = \underset{v \text{ s.t. } \|v\|_2 \leq 1}{\max} \|Av\|_2$

(b) Best fitting k-rank space

We now move to the more general problem of finding a k-dimensional vector space V that minimizes

$$\sum_{i=1}^{d} \| X_i - \text{proj}_V(X_i) \|_2^2.$$

Such a space can be found by the following greedy procedure:

1) $v_1 = \underset{v \text{ s.t. } \|v\|_2 = 1}{\operatorname{argmax}} \|Av\|_2$

2) Repeat for $i = 2, \dots, k$

$$v_i = \underset{\substack{v \text{ s.t. } v \perp \{v_1, \dots, v_{i-1}\} \\ \|v\|_2 = 1}}{\operatorname{argmax}} \|Av\|_2$$

Suppose that we continue the process above till we cannot find a v in step(2). Suppose $\{v_1, \dots, v_r\}$ denote the obtained vectors. Then,

→ row-rank of A is r

→ $\{v_1, \dots, v_r\}$ constitutes an orthonormal basis of the row-space of A ($\equiv$ span$\{X_1, \dots, X_n\}$)

Note that $\|X_i\|_2^2 = \sum_{j=1}^{r} (X_i \cdot v_j)^2$ and summing over i we get

$$\sum_{i=1}^{n} \|X_i\|_2^2 = \sum_{i=1}^{n}\sum_{j=1}^{r} (X_i \cdot v_j)^2 = \sum_{j=1}^{r}\sum_{i=1}^{n} (X_i \cdot v_j)^2$$

$$= \sum_{j=1}^{r} \|A v_j\|_2^2 = \sum_{j=1}^{r} \sigma_j^2(A),$$

where $\sigma_j(A) = \|A v_j\|_2 \equiv j^{th}$ singular value.

Thus, $\sum_{j=1}^{r} \sigma_j^2(A) = \|A\|_F^2 \equiv$ Frobenius norm of A

$\rightarrow$ The vectors $v_1, \dots, v_r$ are called the right-singular vectors.

Let $u_i = \dfrac{1}{\sigma_i(A)} A v_i$, $1 \le i \le r$.

The vectors $u_1, \dots, u_r$ are the left-singular vectors.

(c) Singular Value Decomposition (SVD)

Consider $\sum_{i=1}^{r} \sigma_i(A) u_i v_i^T \equiv \tilde{A}$.

Then, $\tilde{A} v_i = \sigma_i(A) u_i = A v_i$ for $1 \le i \le r$. Since $v_1, \dots, v_r$ constitutes a basis for the row-space of A, $A = \tilde{A}$.

(d) Best rank-k approximation

Let $A_k = \text{span}\{v_1, \dots, v_k\}$.

Theorem. For any matrix B of rank at most k,

(i) $\|A - A_k\|_F \le \|A - B\|_F$, (ii) $\underbrace{\|A - A_k\|_2}_{\sigma_{k+1}} \le \|A - B\|_2$.