CSCI B609: “Foundations of Data Science”

Lecture 5: Dimension Reduction, Separating and Fitting Gaussians

Slides at http://grigory.us/data-science-class.html

Grigory Yaroslavtsev

http://grigory.us

Gaussian Annulus Theorem

- Gaussian in $\boldsymbol{d}$ dimensions $(N_{\boldsymbol{d}}(0^{\boldsymbol{d}}, 1))$:

$$\Pr[\boldsymbol{x} = (z_1, \dots, z_{\boldsymbol{d}})] = (2\pi)^{-\frac{\boldsymbol{d}}{2}} e^{-\frac{z_1^2 + z_2^2 + \cdots + z_{\boldsymbol{d}}^2}{2}}$$

Nearly all mass in annulus of radius $\sqrt{\boldsymbol{d}}$ and width $O(1)$:

- **Thm.** For any $\boldsymbol{\beta} \leq \sqrt{\boldsymbol{d}}$ all but $3e^{-\boldsymbol{c}\boldsymbol{\beta}^2}$ probability mass satisfies $\sqrt{\boldsymbol{d}} - \boldsymbol{\beta} \leq ||\boldsymbol{x}||_{\mathbf{2}} \leq \sqrt{\boldsymbol{d}} + \boldsymbol{\beta}$ for constant $\boldsymbol{c}$

Nearest Neighbors and Random Projections

- Given a database $\boldsymbol{A}$ of n points in $\mathbb{R}^{\boldsymbol{d}}$
  - Preprocess $\boldsymbol{A}$ into a small data structure $\boldsymbol{D}$
  - Should answer following queries fast:

Given $\boldsymbol{q} \in \mathbb{R}^{\boldsymbol{d}}$ find closest $\boldsymbol{x} \in \boldsymbol{A}$: $argmin_{x \in \boldsymbol{A}} ||\boldsymbol{q} - \boldsymbol{x}||_2$

- Project each $\boldsymbol{x} \in \boldsymbol{A}$ onto $f(\boldsymbol{x})$, where $f: \mathbb{R}^{\boldsymbol{d}} \to \mathbb{R}^{\boldsymbol{k}}$
- Pick $\boldsymbol{k}$ vectors $\boldsymbol{u}_1, \dots, \boldsymbol{u}_{\boldsymbol{k}}$ i.i.d: $\boldsymbol{u}_i \sim N_{\boldsymbol{d}}(0^{\boldsymbol{d}}, 1)$

$$f(\boldsymbol{v}) = (\langle \boldsymbol{u}_1, \boldsymbol{v} \rangle, \dots, \langle \boldsymbol{u}_{\boldsymbol{k}}, \boldsymbol{v} \rangle)$$

- Will show that w.h.p. $||f(\boldsymbol{v})||_2 \approx \sqrt{\boldsymbol{k}}||\boldsymbol{v}||_2$

Return: $argmin_{x \in A}||f(\boldsymbol{q}) - f(\boldsymbol{x})||_2 = argmin_{x \in A}||f(\boldsymbol{q} - \boldsymbol{x})||_2 \approx \sqrt{\boldsymbol{k}}\, argmin_{x \in A}||\boldsymbol{q} - \boldsymbol{x}||_2$

Random Projection Theorem

- Pick $\boldsymbol{k}$ vectors $\boldsymbol{u}_1, \dots, \boldsymbol{u}_{\boldsymbol{k}}$ i.i.d: $\boldsymbol{u}_i \sim N_{\boldsymbol{d}}(0^{\boldsymbol{d}}, 1)$
$$f(\boldsymbol{v}) = (\langle \boldsymbol{u}_1, \boldsymbol{v} \rangle, \dots, \langle \boldsymbol{u}_{\boldsymbol{k}}, \boldsymbol{v} \rangle)$$
- Will show that w.h.p. $\left|\left|f(\boldsymbol{v})\right|\right|_2 \approx \sqrt{\boldsymbol{k}}\left|\left|\boldsymbol{v}\right|\right|_2$

Thm. Fix $\boldsymbol{v} \in \mathbb{R}^{\boldsymbol{d}}$ then $\exists c > 0$: for $\boldsymbol{\epsilon} \in (0,1)$:
$$\Pr_{\boldsymbol{u}_i \sim N_{\boldsymbol{d}}(0^{\boldsymbol{d}},1)}\left[\left|\left|\left|f(\boldsymbol{v})\right|\right|_2 - \sqrt{\boldsymbol{k}}\left|\left|\boldsymbol{v}\right|\right|_2\right| \geq \boldsymbol{\epsilon}\sqrt{\boldsymbol{k}}\left|\left|\boldsymbol{v}\right|\right|_2\right] \leq 3\, e^{-c\boldsymbol{k}\boldsymbol{\epsilon}^2}$$

- Scaling: $\left|\left|\boldsymbol{v}\right|\right|_2 = 1$
- **Key fact**: $\langle \boldsymbol{u}_i, \boldsymbol{v} \rangle = \sum_{j=1}^{\boldsymbol{d}} \boldsymbol{u}_{ij} \boldsymbol{v}_j \sim N(0, \left|\left|\boldsymbol{v}\right|\right|_2^2) = N(0,1)$
- Apply "Gaussian Annulus Theorem" with $\boldsymbol{k} = \boldsymbol{d}$

Nearest Neighbors and Random Projections

Thm. Fix $\boldsymbol{v} \in \mathbb{R}^{\boldsymbol{d}}$ then $\exists c > 0$: for $\boldsymbol{\epsilon} \in (0,1)$:

$$\Pr_{\boldsymbol{u}_i \sim N_{\boldsymbol{d}}(0^{\boldsymbol{d}},1)}\left[\left|\,||f(\boldsymbol{v})||_2 - \sqrt{\boldsymbol{k}}\,||\boldsymbol{v}||_2\right| \geq \boldsymbol{\epsilon}\sqrt{\boldsymbol{k}}\,||\boldsymbol{v}||_2\right] \leq 3\, e^{-c\boldsymbol{k}\boldsymbol{\epsilon}^2}$$

Return: $argmin_{x\in A}||f(\boldsymbol{q}) - f(\boldsymbol{x})||_2 \approx \sqrt{\boldsymbol{k}}\, argmin_{x\in A}||\boldsymbol{q} - \boldsymbol{x}||_2$

- Fix and let $\boldsymbol{v} = \boldsymbol{q} - \boldsymbol{x}_i$ for $\boldsymbol{x}_i \in \boldsymbol{A}$ and let $\boldsymbol{k} = O\left(\frac{\gamma \log n}{\boldsymbol{\epsilon}^2}\right)$

$(1 \pm \boldsymbol{\epsilon})\sqrt{\boldsymbol{k}}\,||\boldsymbol{q} - \boldsymbol{x}_i||_2 \approx ||f(\boldsymbol{q}) - f(\boldsymbol{x})||_2$ (prob. $1 - n^{-\gamma}$)

- Union bound:

For fixed $\boldsymbol{q}$ distances to $\boldsymbol{A}$ preserved with prob. $1 - n^{-\gamma+1}$

Separating Gaussians

- One-dimensional mixture of Gaussians:
$$p(x) = w_1 p_1(x) + w_2 p_2(x)$$
- E.g. modeling heights of men/women
- **Parameter estimation problem**:
  - Given samples from a mixture of Gaussians
  - **Q:** Estimate means and (co)-variances
- **Sample origin problem**:
  - Given samples from **well-separated** Gaussians
  - **Q**: Did they come from the same Gaussian?

Separating Gaussians

- Gaussian in $\boldsymbol{d}$ dimensions ($N_{\boldsymbol{d}}(0^{\boldsymbol{d}}, 1)$):

$$\Pr[\boldsymbol{x} = (z_1, \dots, z_{\boldsymbol{d}})] = (2\pi)^{-\frac{\boldsymbol{d}}{2}} e^{-\frac{z_1^2 + z_2^2 + \cdots + z_{\boldsymbol{d}}^2}{2}}$$

Nearly all mass in annulus of radius $\sqrt{\boldsymbol{d}}$ and width $O(1)$:

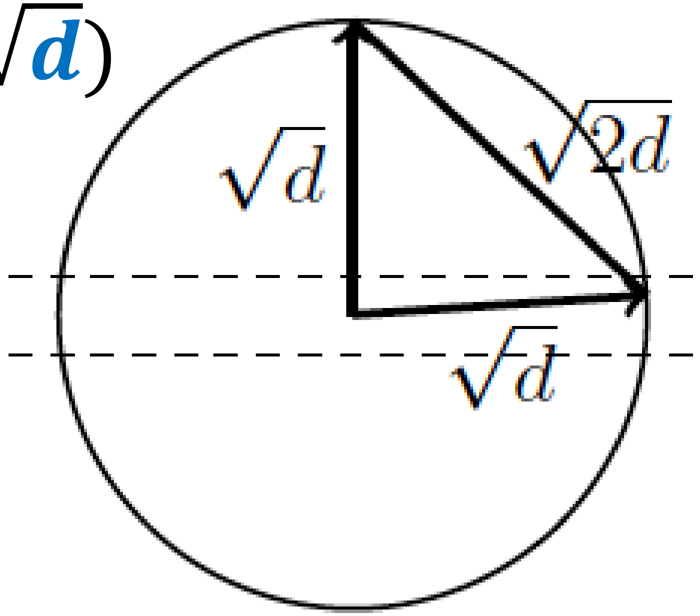
- Almost all mass in a slab $\{\boldsymbol{x} | -c \leq x_1 \leq c\}$ for $c = O(1)$
- Pick $\boldsymbol{x}$ ~Gaussian and rotate coordinates to make it x_1
- Pick $\boldsymbol{y}$ ~Gaussian, w.h.p. projection of $\boldsymbol{y}$ on $\boldsymbol{x}$ is $\in [-c, c]$

$$||\boldsymbol{x} - \boldsymbol{y}||_2 \approx \sqrt{||\boldsymbol{x}||_2^2 + ||\boldsymbol{y}||_2^2}$$

Separating Gaussians

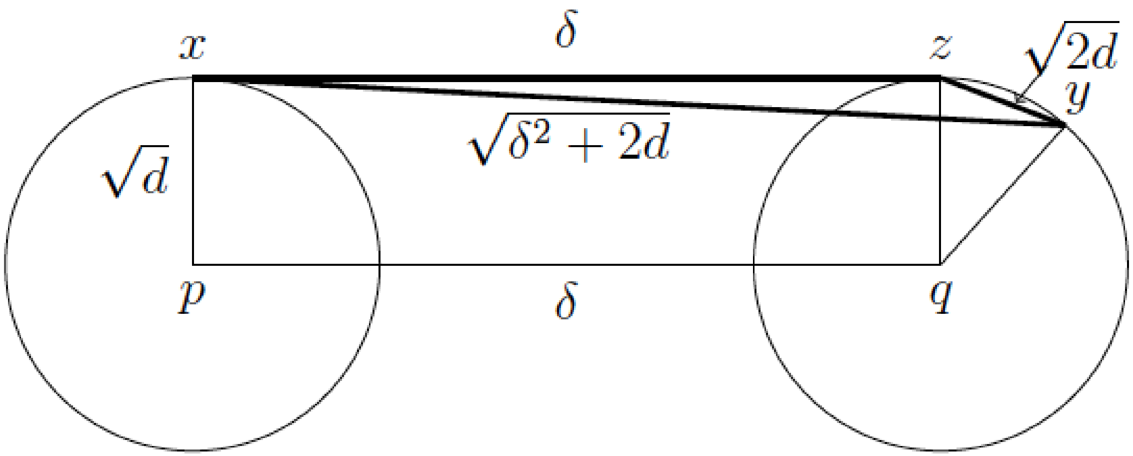
In coordinates:

- $\boldsymbol{x} = (\sqrt{\boldsymbol{d}} \pm O(1), 0, 0, \ldots, 0)$
- $\boldsymbol{y} = (\pm O(1), \sqrt{\boldsymbol{d}} \pm O(1), 0, \ldots, 0)$
- W.h.p: $\left\|\boldsymbol{x} - \boldsymbol{y}\right\|_2^2 = 2\boldsymbol{d} \pm O(\sqrt{\boldsymbol{d}})$



Separating Gaussians

- Two spherical unit variance Gaussians centered at $\boldsymbol{p}, \boldsymbol{q}$
- $\delta = \left|\left|\boldsymbol{p} - \boldsymbol{q}\right|\right|_{\boldsymbol{2}}$
- $(\boldsymbol{x} \sim N(\boldsymbol{p}, 1), \boldsymbol{y} \sim N(\boldsymbol{q}, 1))$
- $\boldsymbol{x} = \left(\sqrt{\boldsymbol{d}} \pm O(1), 0,0,0 \ldots, 0\right)$
- $\boldsymbol{y} = (\pm O(1), \delta \pm O(1), \sqrt{\boldsymbol{d}} \pm O(1), 0, \ldots, 0)$
- $\left|\left|\boldsymbol{x} - \boldsymbol{y}\right|\right|_{\boldsymbol{2}}^{\boldsymbol{2}} = \delta^2 + 2\boldsymbol{d} \pm O\left(\sqrt{\boldsymbol{d}}\right)$



Separating Gaussians

- Same Gaussian:

$$||\boldsymbol{x}-\boldsymbol{y}||_{\mathbf{2}}^{\mathbf{2}} = 2\boldsymbol{d} \pm O(\sqrt{\boldsymbol{d}})$$

- Different Gaussians:

$$||\boldsymbol{x}-\boldsymbol{y}||_{\mathbf{2}}^{\mathbf{2}} = \delta^2 + 2\boldsymbol{d} \pm O(\sqrt{\boldsymbol{d}})$$

- Separation requires:

$$2\boldsymbol{d} \pm O(\sqrt{\boldsymbol{d}}) < \delta^2 + 2\boldsymbol{d} \pm O(\sqrt{\boldsymbol{d}})$$

$$O(\sqrt{\boldsymbol{d}}) < \delta^2$$

$$\omega(\boldsymbol{d}^{\mathbf{1/4}}) < \delta$$

Fitting Spherical Gaussian to Data

- Given samples $\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_n}$
- **Q:** What are parameters of best fit $N(\boldsymbol{\mu}, \sigma)$?

$$\Pr[\boldsymbol{x_i} = (z_1, \dots, z_{\boldsymbol{d}})]$$

$$= (2\pi)^{-\frac{\boldsymbol{d}}{2}} e^{-\frac{(\boldsymbol{\mu}_1 - z_1)^2 + (\boldsymbol{\mu}_2 - z_2)^2 + \cdots + (\boldsymbol{\mu}_{\boldsymbol{d}} - z_{\boldsymbol{d}})^2}{2}}$$

$$= (2\pi)^{-\frac{\boldsymbol{d}}{2}} e^{-\frac{||\boldsymbol{\mu} - \boldsymbol{z}||_2^2}{2}}$$

$$\Pr[\boldsymbol{x_1} = \boldsymbol{z_1}, \boldsymbol{x_2} = \boldsymbol{z_2}, \dots, \boldsymbol{x_n} = \boldsymbol{z_n}]$$

$$= (2\pi)^{-\frac{\boldsymbol{d}n}{2}} e^{-\frac{||\boldsymbol{\mu} - \boldsymbol{z_1}||_2^2 + ||\boldsymbol{\mu} - \boldsymbol{z_2}||_2^2 + \cdots + ||\boldsymbol{\mu} - \boldsymbol{z_d}||_2^2}{2\sigma^2}}$$

Maximum Likelihood Estimator

- PDF: $(2\pi)^{-\frac{dn}{2}} e^{-\frac{||\mu - z_1||_2^2 + ||\mu - z_2||_2^2 + \cdots + ||\mu - z_d||_2^2}{2\sigma^2}}$
- MLE for $\boldsymbol{\mu}$ is $\boldsymbol{\mu} = \frac{1}{n}(\boldsymbol{x_1} + \boldsymbol{x_2} + \cdots + \boldsymbol{x_n})$
- Take gradient w.r.t $\boldsymbol{\mu}$ and make it $= 0$
- $\nabla_{\mu} \left|\left|\boldsymbol{\mu} - \boldsymbol{x}\right|\right|_2^2 = 2(\boldsymbol{\mu} - \boldsymbol{x})$

$$2(\boldsymbol{\mu} - \boldsymbol{x_1}) + 2(\boldsymbol{\mu} - \boldsymbol{x_2}) + \cdots + 2(\boldsymbol{\mu} - \boldsymbol{x_n}) = 0$$

MLE for Variance

- $a = ||\boldsymbol{\mu} - \boldsymbol{x_1}||_2^2 + ||\boldsymbol{\mu} - \boldsymbol{x_2}||_2^2 + \cdots + ||\boldsymbol{\mu} - \boldsymbol{x_d}||_2^2$
- $\nu = 1/2\sigma^2$
- PDF: $\dfrac{e^{-a\nu}}{\left[\int_{x \in R^{d}} e^{-\nu||\boldsymbol{x}||_2^2} d\boldsymbol{x}\right]^n}$
- Log(PDF): $-a\nu - n \ln\left[\int_{x \in R^{d}} e^{-\nu||\boldsymbol{x}||_2^2} d\boldsymbol{x}\right]$
- Differentiate w.r.t. ν and set derivative $= 0$

MLE for Variance

- Log(PDF): $-a\nu - n\ln\left[\int_{\boldsymbol{x}\in R^{\boldsymbol{d}}} e^{-\nu||\boldsymbol{x}||_2^2}\, d\boldsymbol{x}\right]$
- $\frac{d}{d\nu}$ Log(PDF):

$$-a + n\frac{\int_{\boldsymbol{x}\in R^{\boldsymbol{d}}} ||\boldsymbol{x}||_2^2\, e^{-\nu||\boldsymbol{x}||_2^2} d\boldsymbol{x}}{\int_{\boldsymbol{x}\in R^{\boldsymbol{d}}} e^{-\nu||\boldsymbol{x}||_2^2} d\boldsymbol{x}}$$

- $y = ||\nu\boldsymbol{x}||_2^2$:

$$-a + \frac{n}{\nu}\frac{\int_{\boldsymbol{x}\in R^{\boldsymbol{d}}} y^2\, e^{-y^2}\, d\boldsymbol{x}}{\int_{\boldsymbol{x}\in R^{\boldsymbol{d}}} e^{-y^2}\, d\boldsymbol{x}} = -a + \frac{n}{\nu}\times\frac{\boldsymbol{d}}{2} = 0$$

- $\nu = \frac{1}{2\sigma^2} \Rightarrow$ MLE(σ)= $\sqrt{\frac{a}{n\boldsymbol{d}}}$ = sample standard deviation