Module 6: Dimensionality Reduction

Lecture 33: Embedding

Prev topic Next topic Prev page
Next page

The Lecture Contains:

Module 6: Dimensionality Reduction

Lecture 33: Embedding

Prev topic Next topic Prev page
Next page

Johnson-Lindenstrauss lemma

- Embedding (mapping f) of N points from (R^m, L_2) to (R^k, L_2) where

  $k = O(\lg N/\epsilon^2)$ with distortion at most $(1+\epsilon)$

- Randomized construction
- Construct k vectors $r_1, r_2, \ldots, r_k$ where each component r_{ij} is chosen from a standard normal distribution $N(0,1)$
- Object o is then mapped to $f(o)$ where

$$f(o) = (1/\sqrt{k})\{\langle o, r_1\rangle, \langle o, r_2\rangle, \ldots, \langle o, r_k\rangle\}$$

- Then, for two objects x and y, with very high probability,

$$(1-\epsilon)d(x,y) \leq d'(f(x), f(y)) \leq (1+\epsilon)d(x,y)$$

i.e., $(1-\epsilon)||x-y||^2 \leq ||f(x)-f(y)||^2 \leq (1+\epsilon)||x-y||^2$

Efficient construction

- Choose $k = \frac{4+2\beta}{\epsilon^2/2-\epsilon^3/3}\lg N = O(\lg N/\epsilon^2)$
- For data matrix A of size $N \times m$, construct embedding matrix R of size $m \times k$ where r_{ij} are independent random variables (iid) chosen from one of the following distributions:

$$r_{ij} = \begin{cases} +1 & \text{with probability } 1/2 \\ -1 & \text{with probability } 1/2 \end{cases}$$

$$r_{ij} = \sqrt{3} \times \begin{cases} +1 & \text{with probability } 1/6 \\ 0 & \text{with probability } 2/3 \\ -1 & \text{with probability } 1/6 \end{cases}$$

- Embedded data E is given by

$$E = (1/\sqrt{k})A.R$$

- With probability $1-(1/N^\beta)$,

$$(1-\epsilon)||x-y||^2 \leq ||f(x)-f(y)||^2 \leq (1+\epsilon)||x-y||^2$$

Embedding Quadratic Form distance

- QFDs and other statistical distances do not allow indexes
    - Aim is to design a low distortion low-dimensional embedding
- Matrix A in QFD allows Cholesky decomposition since A is symmetric and positive definite

$$QFD(x,y) = \sqrt{(x-y)^T A (x-y)}$$

$$= \sqrt{(x-y)^T C^T C (x-y)} = \sqrt{(Cx-Cy)^T (Cx-Cy)}$$

$$= L_2(Cx, Cy)$$

- Therefore, embedding is $(x, QFD) \rightarrow (Cx, L_2)$
- Allows further dimensionality reduction as well

Prev topic Next topic Prev page
Next page

Bounds on distortion

- If original distance d is not a metric, it is advisable to embed into a metric distance d'
- However, if d does not obey relaxed metric properties, then this embedding cannot be low distortion
- Relaxed metric properties:
  - γ-relaxed symmetry:

$$\forall p, q, \ |d(p,q) - d(q,p)| \leq \gamma \quad [\gamma \geq 0]$$

  - α-relaxed triangular inequality:

$$\forall p, q, r, \ d(p,r) + d(r,q) \geq \alpha . d(p,q) \quad [\alpha \leq 1]$$

Bounds on distortion

- If original distance d is not a metric, it is advisable to embed into a metric distance d'
- However, if d does not obey relaxed metric properties, then this embedding cannot be low distortion
- Relaxed metric properties:
  - γ-relaxed symmetry:

$$\forall p, q, \ |d(p,q) - d(q,p)| \leq \gamma \quad [\gamma \geq 0]$$

  - α-relaxed triangular inequality:

$$\forall p, q, r, \ d(p,r) + d(r,q) \geq \alpha . d(p,q) \quad [\alpha \leq 1]$$

- If d does not obey γ-relaxed symmetry, then any embedding into a metric space d' must have a distortion of at least than $1 + \gamma / M$ where $\forall p, q, \ d(p,q) \leq M$

Prev topic Next topic Prev page
Next page

Bounds on distortion

- If original distance d is not a metric, it is advisable to embed into a metric distance d'
- However, if d does not obey relaxed metric properties, then this embedding cannot be low distortion
- Relaxed metric properties:
  - γ-relaxed symmetry:

$$\forall p, q, \ |d(p,q) - d(q,p)| \leq \gamma \quad [\gamma \geq 0]$$

  - α-relaxed triangular inequality:

$$\forall p, q, r, \ d(p,r) + d(r,q) \geq \alpha . d(p,q) \quad [\alpha \leq 1]$$

- If d does not obey γ-relaxed symmetry, then any embedding into a metric space d' must have a distortion of at least than $1 + \gamma / M$ where $\forall p, q, \ d(p,q) \leq M$
- If d does not obey α-relaxed triangular inequality, then any embedding into a metric space d' must have a distortion of at least $1/\alpha$

Prev topic Next topic Prev page
Next page

Proof for γ-relaxed symmetry

- If distortion is D, then

$$\forall p, q,\ \ c.d(p,q) \leq d'(p',q') \leq c.D.d(p,q)$$

- If d does not follow γ-relaxed symmetry, then for some p, q,

$$d(p,q) > d(q,p) + \gamma$$

$$\therefore\ \ c.d(p,q) > c.d(q,p) + c.\gamma$$

$$\implies\ d'(p',q') > c.d(q,p) + c.\gamma$$

$$\implies\ \frac{d'(p',q')}{c.d(q,p)} > 1 + \frac{c.\gamma}{c.d(q,p)}$$

$$\implies\ \frac{d'(q',p')}{c.d(q,p)} > 1 + \frac{c.\gamma}{c.M} \quad [\because d(q,p) \leq M]$$

$$\implies\ D > 1 + \frac{\gamma}{M}$$

- γ/M is just an additive distortion

Proof for α-relaxed triangular inequality

- If distortion is D, then

$$\forall p, q,\ c.d(p,q) \leq d'(p',q') \leq c.D.d(p,q)$$

- If d does not follow α-relaxed triangular inequality, then

$$d(p,r) + d(r,q) < \alpha.d(p,q)$$

- Since d' is a metric,

$$d'(p',r') + d'(r',q') \geq d'(p',q')$$

$$\implies c.D.d(p,r) + c.D.d(r,q) \geq d'(p',q')$$

$$\implies c.D.(d(p,r) + d(r,q)) \ \geq c.d(p,q)$$

$$\implies c.D.(\alpha.d(p,q)) \geq c.d(p,q)$$

$$\implies D \geq \frac{1}{\alpha}$$

- $1/\alpha$ is a multiplicative distortion