Subspace Embeddings for Regression

Lecture 12
October 1, 2020
Subspace Embedding

**Question:** Suppose we have linear subspace $E$ of $\mathbb{R}^n$ of dimension $d$. Can we find a projection $\Pi : \mathbb{R}^d \to \mathbb{R}^k$ such that for every $x \in E$, $\|\Pi x\|_2 = (1 \pm \epsilon)\|x\|_2$?

- Not possible if $k < d$.
- Possible if $k = \ell$. Pick $\Pi$ to be an orthonormal basis for $E$.

**Disadvantage:** This requires knowing $E$ and computing orthonormal basis which is slow.

**What we really want:** *Oblivious* subspace embedding ala JL based on random projections
Oblivious Subspace Embedding

**Theorem**

Suppose $E$ is a linear subspace of $\mathbb{R}^n$ of dimension $d$. Let $\Pi$ be a DJL matrix $\Pi \in \mathbb{R}^{k \times d}$ with $k = O\left(\frac{d}{\epsilon^2 \log(1/\delta)}\right)$ rows. Then with probability $(1 - \delta)$ for every $x \in E$,

$$\frac{1}{\sqrt{k}} \|\Pi x\|_2 = (1 \pm \epsilon) \|x\|_2.$$ 

In other words JL Lemma extends from one dimension to arbitrary number of dimensions in a graceful way.
Part I

Faster algorithms via subspace embeddings
Linear model fitting

An important problem in data analysis

- $n$ data points
- Each data point $\mathbf{a}_i \in \mathbb{R}^d$ and real value $b_i$. We think of $\mathbf{a}_i = (a_{i,1}, a_{i,2}, \ldots, a_{i,d})$. Interesting special case is when $d = 1$.
- What model should one use to explain the data?
Linear model fitting

An important problem in data analysis

- **n** data points
- Each data point $a_i \in \mathbb{R}^d$ and real value $b_i$. We think of $a_i = (a_{i,1}, a_{i,2}, \ldots, a_{i,d})$. Interesting special case is when $d = 1$.
- What model should one use to explain the data?

Simplest model? Affine fitting. $b_i = \alpha_0 + \sum_{j=1}^{d} \alpha_j a_{i,j}$ for some real numbers $\alpha_0, \alpha_1, \ldots, \alpha_d$. Can restrict to $\alpha_0 = 0$ by lifting to $d + 1$ dimensions and hence linear model.
Linear model fitting

An important problem in data analysis

- \( n \) data points
- Each data point \( a_i \in \mathbb{R}^d \) and real value \( b_i \). We think of \( a_i = (a_{i,1}, a_{i,2}, \ldots, a_{i,d}) \). Interesting special case is when \( d = 1 \).
- What model should one use to explain the data?

Simplest model? Affine fitting. \( b_i = \alpha_0 + \sum_{j=1}^{d} \alpha_j a_{i,j} \) for some real numbers \( \alpha_0, \alpha_1, \ldots, \alpha_d \). Can restrict to \( \alpha_0 = 0 \) by lifting to \( d + 1 \) dimensions and hence linear model.

But data is noisy so we won’t be able to satisfy all data points even if true model is a linear model. How do we find a good linear model?
Regression

- \( n \) data points
- Each data point \( a_i \in \mathbb{R}^d \) and real value \( b_i \). We think of \( a_i = (a_{i,1}, a_{i,2}, \ldots, a_{i,d}) \).

Linear model fitting: Find real numbers \( \alpha_1, \ldots, \alpha_d \) such that 
\[
b_i \approx \sum_{j=1}^{d} \alpha_j a_{i,j} \quad \text{for all points.}
\]

Let \( A \) be matrix with one row per data point \( a_i \). We write \( x_1, x_2, \ldots, x_d \) as variables for finding \( \alpha_1, \ldots, \alpha_d \).

**Ideally:** Find \( x \in \mathbb{R}^d \) such that \( Ax = b \)

**Best fit:** Find \( x \in \mathbb{R}^d \) to minimize \( Ax - b \) under some norm.

- \( \|Ax - b\|_\infty, \|Ax - b\|_2, \|Ax - b\|_1 \)
Linear least squares: Given $A \in \mathbb{R}^{n \times d}$ and $b \in \mathbb{R}^d$ find $x$ to minimize $\|Ax - b\|_2$. Optimal estimator for certain noise models

Interesting when $n \gg d$ the over constrained case when there is no solution to $Ax = b$ and want to find best fit.
Linear least squares/Regression

Linear least squares: Given $A \in \mathbb{R}^{n \times d}$ and $b \in \mathbb{R}^d$ find $x$ to minimize $\|Ax - b\|_2$.

Interesting when $n \gg d$ the over constrained case when there is no solution to $Ax = b$ and want to find best fit.

Geometrically $Ax$ is a linear combination of columns of $A$. Hence we are asking what is the vector $z$ in the column space of $A$ that is closest to vector $b$ in $\ell_2$ norm.

Closest vector to $b$ is the projection of $b$ into the column space of $A$ so it is “obvious” geometrically. How do we find it?
Fix $x \in \mathbb{R}^n$.

If $Ax = b$, then $x = x_1 A_1 + x_2 A_2 + \ldots + x_k A_k$.

Suppose $b$ is not in the column span of $A$. What is the answer?
Linear least squares/Regression

Linear least squares: Given $A \in \mathbb{R}^{n \times d}$ and $b \in \mathbb{R}^d$ find $x$ to minimize $\|Ax - b\|_2$.

Geometrically $Ax$ is a linear combination of columns of $A$. Hence we are asking what is the vector $z$ in the column space of $A$ that is closest to vector $b$ in $\ell_2$ norm.

Closest vector to $b$ is the projection of $b$ into the column space of $A$ so it is “obvious” geometrically. How do we find it?

- Find an orthonormal basis $z_1, z_2, \ldots, z_r$ for the columns of $A$.
- Compute projection $c$ of $b$ to column space of $A$ as $c = \sum_{j=1}^{r} \langle b, z_j \rangle z_j$ and output answer as $\|b - c\|_2$.
- What is $x$? $x$ is obtained by expressing $c$ as $Ax = c$.
Linear least squares/Regression

Linear least squares: Given $A \in \mathbb{R}^{n \times d}$ and $b \in \mathbb{R}^d$ find $x$ to minimize $\|Ax - b\|_2$.

Geometrically $Ax$ is a linear combination of columns of $A$. Hence we are asking what is the vector $z$ in the column space of $A$ that is closest to vector $b$ in $\ell_2$ norm.

Closest vector to $b$ is the projection of $b$ into the column space of $A$ so it is “obvious” geometrically. How do we find it?

- Find an orthonormal basis $z_1, z_2, \ldots, z_r$ for the columns of $A$.
- Compute projection $c$ of $b$ to column space of $A$ as $c = \sum_{j=1}^r \langle b, z_j \rangle z_j$ and output answer as $\|b - c\|_2$.
- What is $x$? We know that $Ax = c$. Solve linear system. Can combine both steps via SVD and other methods.
Linear least square: Optimization perspective

Linear least squares: Given $A \in \mathbb{R}^{n \times d}$ and $b \in \mathbb{R}^d$ find $x$ to minimize $\|Ax - b\|_2$.

Optimization: Find $x \in \mathbb{R}^d$ to minimize $\|Ax - b\|_2^2$

$$\|Ax - b\|_2^2 = x^T A^T A x - 2 b^T A x + b^T b$$

The quadratic function $f(x) = x^T A^T A x - 2 b^T A x + b^T b$ is a convex function since the matrix $A^T A$ is positive semi-definite. $\nabla f(x) = 2 A^T A x - 2 b^T A$ and hence optimum solution $x^*$ is given by $x^* = (A^T A)^{-1} b^T A$. 

Computational perspective

$n$ large (number of data points), $d$ smaller so $A$ is tall and skinny.

Exact solution requires SVD or other methods. Worst case time $nd^2$.

Can we speed up computation with some potential approximation?

\[
\frac{d}{\varepsilon^2} \quad \frac{1}{nd^2} \quad \approx \frac{d^3}{\varepsilon^2} + nd
\]
Let $A^{(1)}, A^{(2)}, \ldots, A^{(d)}$ be the columns of $A$ and let $E$ be the subspace spanned by $\{A^{(1)}, A^{(2)}, \ldots, A^{(d)}, b\}$.

Note columns are in $\mathbb{R}^n$ corresponding to $n$ data points $E$ has dimension at most $d + 1$.

Use subspace embedding on $E$. Applying JL matrix $\Pi$ with $k = O\left(\frac{d^2}{\epsilon^2}\right)$ rows we reduce $\{A^{(1)}, A^{(2)}, \ldots, A^{(d)}, b\}$ to $\{A^{(1)}', A^{(2)}', \ldots, A^{(d)}', b'\}$ which are vectors in $\mathbb{R}^k$.

Solve $\min_{x' \in \mathbb{R}^d} \|A'x' - b'\|_2$
\[
\begin{pmatrix}
A^{(r)} & \cdots & A^{d} \\
\vdots & \ddots & \vdots \\
A^{d} & \cdots & A^{r}
\end{pmatrix}
\begin{bmatrix}
\mathbf{b} \\
\mathbf{c} \\
\vdots \\
\mathbf{b}
\end{bmatrix}
= 
\begin{bmatrix}
d+1 \\
k+1 \\
1
\end{bmatrix}
\]

\( T \subset \mathbb{R}^{k \times n} \quad k = O\left(\frac{d}{\varepsilon^2} \frac{\ln \frac{1}{\delta}}{\delta}\right) \)

\[
\Pi \mathbf{A} = \Pi \mathbf{A} \quad \Pi \mathbf{b} = \mathbf{b}^1
\]

\[
K_{\mathbf{A}} \mathbf{A} = \begin{bmatrix}
\mathbf{w}^{(1)}(\mathbf{A}) & \mathbf{w}^{(2)}(\mathbf{A}) & \cdots & \mathbf{w}^{(k)}(\mathbf{A})
\end{bmatrix}
\begin{bmatrix}
\mathbf{U} \\
\mathbf{V} \\
\mathbf{W}
\end{bmatrix}
\]

\( K = \frac{d}{\varepsilon^2} \)

Analysis

Lemma

With probability \((1 - \delta)\),

\[
(1 - \epsilon) \min_{x \in \mathbb{R}^d} \|Ax - b\| \leq \min_{x' \in \mathbb{R}^d} \|A'x' - b'\|_2 \leq (1 + \epsilon) \min_{x \in \mathbb{R}^d} \|Ax - b\|
\]
Analysis

Lemma

With probability \((1 - \delta)\),

\[
(1 - \epsilon) \min_{x \in \mathbb{R}^d} \|Ax - b\| \leq \min_{x' \in \mathbb{R}^d} \|A'x' - b'\|_2 \leq (1 + \epsilon) \min_{x \in \mathbb{R}^d} \|Ax - b\|
\]

With probability \((1 - \delta)\) via the subspace embedding guarantee, for all \(z \in E\),

\[
(1 - \epsilon) \|z\|_2 \leq \|\Pi z\|_2 \leq (1 + \epsilon) \|z\|_2
\]

Now prove two inequalities in lemma separately using above.
Suppose $x^*$ is an optimum solution to $\min_x \|Ax - b\|_2$.

Let $z = Ax^* - b$. We have $\|\Pi z\|_2 \leq (1 + \epsilon)\|z\|_2$ since $z \in E$. 
Suppose \( x^* \) is an optimum solution to \( \min_x \|Ax - b\|_2 \).

Let \( z = Ax^* - b \). We have \( \|\Pi z\|_2 \leq (1 + \epsilon)\|z\|_2 \) since \( z \in E \).

Since \( x^* \) is a feasible solution to \( \min_{x'} \|A'x' - b'\| \),

\[
\min_{x'} \|A'x' - b'\|_2 \leq \|A'x^* - b'\|_2 = \|\Pi(Ax^* - b)\|_2 \leq (1 + \epsilon)\|Ax^* - b\|_2
\]
Analysis

For any $y \in \mathbb{R}^d$, $\|\Pi Ay - \Pi b\|_2 \geq (1 - \epsilon)\|Ay - b\|_2$ because $Ay - b$ is a vector in $E$ and $\Pi$ preserves all of them.

$$\|\Pi (Ay - b)\|_2 \leq \|\Pi Ay - \Pi b\|_2$$

$$\|\Pi\|_2 \geq (1 - \epsilon) \|\|$$
For any $y \in \mathbb{R}^d$, $\|\Pi Ay - \Pi b\|_2 \geq (1 - \epsilon)\|Ay - b\|_2$ because $Ay - b$ is a vector in $E$ and $\Pi$ preserves all of them.

Let $y^*$ be optimum solution to $\min_{x'} \|A'x' - b'\|_2$. Then

$$\|\Pi (Ay^* - b)\|_2 \geq (1 - \epsilon)\|Ay^* - b\|_2 \geq (1 - \epsilon)\|Ax^* - b\|_2$$
Running time

Reduce problem for $d$ vectors in $\mathbb{R}^n$ to $d$ vectors in $\mathbb{R}^k$ where $k = O(d/\epsilon^2)$.

Computing $\Pi A, \Pi b$ can be done in $\text{nnz}(A)$ via sparse/fast JL (input sparsity time).

Need to solve least squares on $A', b'$ which can be done in $O(d^3/\epsilon^2)$ time.

Essentially reduce $n$ to $d/\epsilon^2$. Useful when $n \gg d/\epsilon^2$ (for this $\epsilon$ should not be too small).
Further improvement

Reduced dimension of vectors from $\mathbb{R}^n$ to $\mathbb{R}^k$ where $k = O(d/\varepsilon^2)$.

For small $\varepsilon$ a dependence of $1/\varepsilon^2$ is not so good. Can we improve?

Can use $\Pi$ with $k = O(d/\varepsilon)$.

- Suffices if $\Pi$ has $1/10$-approximate subspace embedding property and property of preserving matrix multiplication
- $(\Pi A)^T (\Pi A)$ has small condition number
- Use $\Pi$ that has $1/10$-approximate subspace embedding property and then use gradient descent whose convergence depends on condition number of $A$. 
Other uses of JL/subspace embeddings in numerical linear algebra

- Approximate matrix multiplication
- Low rank approximation and SVD
- Compressed Sensing