ORIE 4741: Learning with Big Messy Data

Underdetermined Least Squares

Professor Udell

Operations Research and Information Engineering
Cornell

November 11, 2016
Definition

The **null space** of a matrix $X : \mathbb{R}^{n \times d}$ is

$$X = \{ w \in \mathbb{R}^d : Xw = 0 \}$$

(The all-zero vector 0 is always in the null space.)

The following conditions are equivalent:

- $(X) = \{0\}$
- If $Xw = 0$, then $w = 0$
- The columns of $X$ are linearly independent
- $\forall z \in \mathbb{R}^n$, if $Xw = z$ and $Xw' = z$, then $w = w'$
- $X$ has a left inverse
What if the Gram matrix is not invertible?

\[ X^T X w = X^T y \]
\[ w = (X^T X)^{-1} X^T y \]

1. **Q:** if \( X^T X \) is not invertible, do the normal equations still define the solution?
What if the Gram matrix is not invertible?

\[ X^T X w = X^T y \]
\[ w = (X^T X)^{-1} X^T y \]

1. **Q:** if \( X^T X \) is not invertible, do the normal equations still define the solution?
   **A:** yes! we derived them with no assumptions.

2. **Q:** if they have a solution, is the solution unique?
   **A:** no, it's not unique! if \( \text{Rank}(X) < d \), then for some \( v \neq 0 \), \( Xv = 0 \). so if \( w \) is a solution, so is \( w + \alpha v \) for any \( \alpha \in \mathbb{R} \).

3. **Q:** is non-uniqueness a problem for a predictive model?
   **A:** yes.
What if the Gram matrix is not invertible?

\[ X^T X w = X^T y \]
\[ w = (X^T X)^{-1} X^T y \]

1. **Q:** If \( X^T X \) is not invertible, do the normal equations still define the solution?
   **A:** yes! we derived them with no assumptions.

2. **Q:** If they have a solution, is the solution unique?
   **A:** no, it's not unique!

   If \( \text{Rank}(X) < d \), then for some \( v \neq 0 \), \( Xv = 0 \).

   If \( \mathbf{w} \) is a solution, so is \( \mathbf{w} + \alpha \mathbf{v} \) for any \( \alpha \in \mathbb{R} \).
What if the Gram matrix is not invertible?

\[ X^T X w = X^T y \]
\[ w = (X^T X)^{-1} X^T y \]

1. **Q:** if \( X^T X \) is not invertible, do the normal equations still define the solution?
   **A:** yes! we derived them with no assumptions.

2. **Q:** if they have a solution, is the solution unique?
   **A:** no, it’s not unique!
   if \( \text{Rank}(X) < d \), then for some \( v \neq 0 \), \( Xv = 0 \).
   so if \( w \) is a solution, so is \( w + \alpha v \) for any \( \alpha \in \mathbb{R} \).
What if the Gram matrix is not invertible?

\[ X^T X w = X^T y \]
\[ w = (X^T X)^{-1} X^T y \]

1. **Q:** if \( X^T X \) is not invertible, do the normal equations still define the solution?
   **A:** yes! we derived them with no assumptions.

2. **Q:** if they have a solution, is the solution unique?
   **A:** no, it’s not unique!
   if \( \text{Rank}(X) < d \), then for some \( v \neq 0 \), \( Xv = 0 \).
   so if \( w \) is a solution, so is \( w + \alpha v \) for any \( \alpha \in \mathbb{R} \).

3. **Q:** is non-uniqueness a problem for a predictive model?
What if the Gram matrix is not invertible?

\[ X^T X w = X^T y \]
\[ w = (X^T X)^{-1} X^T y \]

1. \textbf{Q:} if \( X^T X \) is not invertible, do the normal equations still define the solution?  
\textbf{A:} yes! we derived them with no assumptions.

2. \textbf{Q:} if they have a solution, is the solution unique?  
\textbf{A:} no, it’s not unique!  
if \( \text{Rank}(X) \) \(< d \), then for some \( v \neq 0 \), \( Xv = 0 \).  
so if \( w \) is a solution, so is \( w + \alpha v \) for any \( \alpha \in \mathbb{R} \).

3. \textbf{Q:} is non-uniqueness a problem for a predictive model?  
\textbf{A:} yes.
Example: non-uniqueness

- goal: predict cancer risk from mutations in genes
- $X_{ij}$ is 1 if person $i$ has a mutation in gene $j$
- genes 1 and 2 vary together: every person with a mutation in gene 1 has one in gene 2, too, and vice versa
- so the first and second column of $X$ are identical: $X_{1:} = X_{2:}$
- suppose our least squares solution is $w$
- $w' = w + \alpha e_1 - \alpha e_2^1$, for $\alpha \in \mathbb{R}$, makes the same predictions:

\[
Xw' = X(w + \alpha e_1 - \alpha e_2) = Xw + \alpha X(e_1 - e_2) \\
= Xw + \alpha (X_{1:} - X_{2:}) = Xw
\]

- now suppose a new person $x$ arrives with a mutation in gene 1 ($x_1 = 1$) but not in gene 2 ($x_2 = 0$).
Example: non-uniqueness

- **goal**: predict cancer risk from mutations in genes
- $X_{ij}$ is 1 if person $i$ has a mutation in gene $j$
- genes 1 and 2 vary together: every person with a mutation in gene 1 has one in gene 2, too, and vice versa
- so the first and second column of $X$ are identical: $X_{1\cdot} = X_{2\cdot}$
- suppose our least squares solution is $w$
- $w' = w + \alpha e_1 - \alpha e_2^1$, for $\alpha \in \mathbb{R}$, makes the same predictions:

\[
Xw' = X(w + \alpha e_1 - \alpha e_2^1) = Xw + \alpha X(e_1 - e_2)
\]

\[
= Xw + \alpha (X_{1\cdot} - X_{2\cdot}) = Xw
\]

- now suppose a new person $x$ arrives with a mutation in gene 1 ($x_1 = 1$) but not in gene 2 ($x_2 = 0$).
- **Q**: do $w$ and $w'$ make the same prediction?
Example: non-uniqueness

- goal: predict cancer risk from mutations in genes
- \( X_{ij} \) is 1 if person \( i \) has a mutation in gene \( j \)
- genes 1 and 2 vary together: every person with a mutation in gene 1 has one in gene 2, too, and vice versa
- so the first and second column of \( X \) are identical: \( X_{1:} = X_{2:} \)
- suppose our least squares solution is \( w \)
- \( w' = w + \alpha e_1 - \alpha e_2 \), for \( \alpha \in \mathbb{R} \), makes the same predictions:

\[
Xw' = X(w + \alpha e_1 - \alpha e_2) = Xw + \alpha X(e_1 - e_2) \\
= Xw + \alpha(X_{1:} - X_{2:}) = Xw
\]

- now suppose a new person \( x \) arrives with a mutation in gene 1 \( (x_1 = 1) \) but not in gene 2 \( (x_2 = 0) \).
- \textbf{Q:} do \( w \) and \( w' \) make the same prediction? \textbf{A: no!}
Example: non-uniqueness

- goal: predict cancer risk from mutations in genes
- $X_{ij}$ is 1 if person $i$ has a mutation in gene $j$
- genes 1 and 2 vary together: every person with a mutation in gene 1 has one in gene 2, too, and vice versa
- so the first and second column of $X$ are identical: $X_1 = X_2$
- suppose our least squares solution is $w$
- $w' = w + \alpha e_1 - \alpha e_2$, for $\alpha \in \mathbb{R}$, makes the same predictions:
  \[
  Xw' = X(w + \alpha e_1 - \alpha e_2) = Xw + \alpha X(e_1 - e_2)
  = Xw + \alpha (X_1 - X_2) = Xw
  \]
- now suppose a new person $x$ arrives with a mutation in gene 1 ($x_1 = 1$) but not in gene 2 ($x_2 = 0$).
- Q: do $w$ and $w'$ make the same prediction?
  A: no!
- Q: what criterion might you pick to choose a good $w$?

---

$e_1$ is the first standard basis vector $[1, 0, \ldots, 0]$; $e_2$ is the second standard basis vector $[0, 1, \ldots, 0]$; etc.
Example: non-uniqueness

- goal: predict cancer risk from mutations in genes
- \(X_{ij}\) is 1 if person \(i\) has a mutation in gene \(j\)
- genes 1 and 2 vary together: every person with a mutation in gene 1 has one in gene 2, too, and vice versa
- so the first and second column of \(X\) are identical: \(X_{1:} = X_{2:}\)
- suppose our least squares solution is \(w\)
- \(w' = w + \alpha e_1 - \alpha e_2\), for \(\alpha \in \mathbb{R}\), makes the same predictions:

\[
Xw' = X(w + \alpha e_1 - \alpha e_2) = Xw + \alpha X(e_1 - e_2)
\]

\[
= Xw + \alpha(X_{1:} - X_{2:}) = Xw
\]

- now suppose a new person \(x\) arrives with a mutation in gene 1 \((x_1 = 1)\) but not in gene 2 \((x_2 = 0)\).
- \(Q:\) do \(w\) and \(w'\) make the same prediction?
  - \(A:\) no!
- \(Q:\) what criterion might you pick to choose a good \(w\)?
  - \(A:\) pick a \(w\) that’s small; it will make less crazy predictions
The Singular Value Decomposition (SVD)

rewrite $X \in \mathbb{R}^{n \times d}$ in terms of easier matrices

$\begin{align*}
X &= U \Sigma V^T \\
U &\in \mathbb{R}^{n \times r} \text{ has orthogonal columns: } U^T U = I_r \\
V &\in \mathbb{R}^{d \times r} \text{ has orthogonal columns: } V^T V = I_r \\
\Sigma &\in \mathbb{R}^{r \times r} \text{ is diagonal and positive:} \\
&\quad \Sigma_{ii} > 0 \text{ for } i = 1, \ldots, r \\
&\quad \Sigma_{ij} = 0 \text{ for } i \neq j
\end{align*}$

can compute SVD factorization of $X$ in $O(nd^2)$ flops
The Singular Value Decomposition (SVD)

rewrite $X \in \mathbb{R}^{n \times d}$ in terms of easier matrices

- $X = U \Sigma V^T$
- $U \in \mathbb{R}^{n \times r}$ has orthogonal columns: $U^T U = I_r$
- $V \in \mathbb{R}^{d \times r}$ has orthogonal columns: $V^T V = I_r$
- $\Sigma \in \mathbb{R}^{r \times r}$ is diagonal and positive:
  - $\Sigma_{ii} > 0$ for $i = 1, \ldots, r$
  - $\Sigma_{ij} = 0$ for $i \neq j$

can compute SVD factorization of $X$ in $O(nd^2)$ flops

in julia (or matlab), use the SVD function

$U, S, V = \text{svd}(X)$
previous version sometimes called thin SVD. to make full SVD, augment $\Sigma$ with zeros.

suppose $d \leq n$. full SVD is

- $X = U\Sigma V^T$
- $U \in \mathbb{R}^{n \times d}$ has orthogonal columns: $U^T U = I_d$
- $V \in \mathbb{R}^{d \times d}$ has orthogonal columns: $V^T V = I_d$
- $\Sigma \in \mathbb{R}^{d \times d}$ is diagonal and nonnegative:
  - $\Sigma_{ii} \geq 0$ for $i = 1, \ldots, d$
  - $\Sigma_{ij} = 0$ for $i \neq j$

$V$ is square and orthogonal, so $V^T V = VV^T = I_d$
SVD for least squares

if \( X = U\Sigma V^T = \sum_{i=1}^{r} \sigma_i u_i v_i^T \), then

\[
X^T X = V\Sigma^T U^T U\Sigma V^T = V\Sigma^2 V^T
\]

normal equations are

\[
\begin{align*}
X^T X w &= X^T y \\
V\Sigma^2 V^T w &= V\Sigma U^T y \\
\Sigma^{-2} V^T V\Sigma^2 V^T w &= \Sigma^{-2} V^T V\Sigma U^T y \\
V^T w &= \Sigma^{-1} U^T y
\end{align*}
\]

try \( w = V\Sigma^{-1} U^T y = \sum_{i=1}^{d} v_i \frac{1}{\sigma_i} u_i^T y \):

\[
V^T w = V^T V\Sigma^{-1} U^T y = \Sigma^{-1} U^T y
\]

so we’ve found a solution (without assuming invertibility)!
Demo: SVD

https://github.com/ORIE4741/demos/SVD.ipynb
Review: three methods for least squares

- gradient descent (most flexible, $O(nd)$ flops per iteration)
- QR factorization (most efficient exact solution method, $O(nd^2)$ flops)
- SVD factorization (exact solution method for underdetermined problems, $O(nd^2)$ flops)
Quadratic regularization

add a small penalty for large coefficients

$$\text{minimize} \quad ||y - Xw||^2 + \lambda ||w||^2$$

where $\lambda > 0$ is the **regularization parameter**

(also called “regularized least squares”, “ridge regression”, “Tikhonov regularization”, or “weight decay”)

why regularize?

- prevent overfitting
- stabilize estimate
- solution is always unique
Solving regularized regression

minimize \[ ||y - Xw||^2 + \lambda ||w||^2 \]

- solve by setting the derivative to 0: optimal \( w^{\text{ridge}} \) satisfies

\[
0 = \nabla_w ||y - Xw^{\text{ridge}}||^2 + \lambda ||w^{\text{ridge}}||^2 \\
= -2X^Ty + 2X^TXw^{\text{ridge}} + 2\lambda w^{\text{ridge}} \\
(X^TX + \lambda I)w^{\text{ridge}} = X^Ty
\]

- \( X^TX + \lambda I \) is always invertible, so

\[
w^{\text{ridge}} = (X^TX + \lambda I)^{-1}X^Ty
\]
Quadratic regularization and the SVD

suppose $X = U\Sigma V^T$ is the (full) SVD of $X$. regularized solution is

$$w^{\text{ridge}} = (X^TX + \lambda I)^{-1}V\Sigma U^Ty$$
$$= (V\Sigma^2V^T + V(\lambda I)V^T)^{-1}V\Sigma U^Ty$$
$$= V(\Sigma^2 + \lambda I)^{-1}V^T V\Sigma U^Ty$$
$$= V(\Sigma^2 + \lambda I)^{-1}\Sigma U^Ty$$

$$= \sum_{i=1}^{d} v_i \frac{\sigma_i}{\sigma_i^2 + \lambda} u_i^T y$$

ridge regression shrinks $\sigma_i^{-1} = \frac{\sigma_i}{\sigma_i^2}$ to $\frac{\sigma_i}{\sigma_i^2 + \lambda}$
Bias variance tradeoff for regression

- Suppose $y = Xw^\dagger + \epsilon$
- $X = U\Sigma V^T$ is the SVD of $X$

\[ w^{\text{ridge}} = \sum_{i=1}^{d} v_i \frac{\sigma_i}{\sigma_i^2 + \lambda} u_i^T y, \quad w^{\text{lsq}} = \sum_{i=1}^{d} v_i \frac{1}{\sigma_i} u_i^T y \]
Bias variance tradeoff: least squares regression

- suppose $y = Xw^\dagger + \varepsilon$, $\varepsilon_i \sim \mathcal{N}(0, 1)$ for $i = 1, \ldots, n$
- different samples of datasets $\mathcal{D}$ have same $X$, different $\varepsilon$
- $X = U\Sigma V^T$ is the SVD of $X$

$$f(x) = x^T w^\dagger$$
$$g_{\mathcal{D}}(x) = x^T (X^T X)^{-1} X^T y = x^T (X^T X)^{-1} X^T (Xw^\dagger + \varepsilon)$$
$$= x^T w^\dagger + x^T (X^T X)^{-1} X^T \varepsilon$$
$$\bar{g}(x) = \mathbb{E}_\mathcal{D} [g_{\mathcal{D}}(x)] = x^T w^\dagger$$
Bias variance tradeoff: least squares regression

so

\[\text{bias}(x) = f(x) - \bar{g}(x) = 0\]
\[\text{var}(x) = \mathbb{E}_D [(g_D(x) - \bar{g}(x))^2]\]
\[= \mathbb{E}_D \left[ x^T (X^T X)^{-1} X^T \mathbb{E}_D [\epsilon \epsilon^T] X (X^T X)^{-1} x \right]\]
\[= x^T (X^T X)^{-1} X^T \mathbb{E}_D [\epsilon \epsilon^T] X (X^T X)^{-1} x\]
\[= x^T (X^T X)^{-1} X^T I X (X^T X)^{-1} x\]
\[= x^T (X^T X)^{-1} X^T X (X^T X)^{-1} x\]
\[= x^T (X^T X)^{-1} x\]
\[= x^T \left( \sum_{i=1}^{d} \frac{1}{\sigma_i^2} v_i v_i^T \right) x\]
Bias variance tradeoff: ridge regression

- Suppose \( y = Xw^\dagger + \varepsilon, \varepsilon_i \sim \mathcal{N}(0, 1) \) for \( i = 1, \ldots, n \)
- Different samples of datasets \( \mathcal{D} \) have same \( X \), different \( \varepsilon \)
- \( X = U\Sigma V^T \) is the SVD of \( X \)

\[
\begin{align*}
  f(x) & = x^T w^\dagger \\
  g_D(x) & = x^T w^{\text{ridge}} = x^T (X^T X + \lambda I)^{-1} X^T y \\
           & = x^T (X^T X + \lambda I)^{-1} X^T (Xw^\dagger + \varepsilon) \\
  \bar{g}(x) & = \mathbb{E}_D [g_D(x)] = x^T (X^T X + \lambda I)^{-1} X^T Xw^\dagger
\end{align*}
\]
Bias variance tradeoff: ridge regression

so

\[
\text{bias}(x) = f(x) - \bar{g}(x) = x^T((X^TX + \lambda I)^{-1}X^TX - I)w^\dagger
\]

\[
\text{var}(x) = E_D[(g_D(x) - \bar{g}(x))^2]
\]

= \[E_D\left[x^T(X^TX + \lambda I)^{-1}X^T\varepsilon\varepsilon^TX(X^TX + \lambda I)^{-1}x\right]
\]

= \[x^T(X^TX + \lambda I)^{-1}X^TE_D[\varepsilon\varepsilon^T]X(X^TX + \lambda I)^{-1}x\]

= \[x^T(X^TX + \lambda I)^{-1}X^TIX(X^TX + \lambda I)^{-1}x\]

= \[x^T(X^TX + \lambda I)^{-1}X^TX(X^TX + \lambda I)^{-1}x\]

= \[x^T \left( \sum_{i=1}^{d} v_i \frac{\sigma_i^2}{(\sigma_i^2 + \lambda)^2} v_i^T \right) x\]