Advanced Machine Learning Subsidary Notes

Lecture 8: Singular Value Decomposition (SVD)

Adam Prügel-Bennett

April 21, 2020

1 Keywords

- Singular Valued Decomposition, SVD, general linear maps

2 Main Points

2.1 Singular Value Decomposition

- Any $n \times m$ matrix, $\mathbf{X}$ can be decomposed as $\mathbf{X} = \mathbf{U}\,\mathbf{S}\,\mathbf{V}^\mathsf{T}$
  - $\mathbf{U}$ is an $n \times n$ orthogonal matrix
  - $\mathbf{S}$ is an $n \times m$ matrix with zeros everywhere except the diagonal where $S_{ii} = s_i \geq 0$
  - $\mathbf{V}$ is an $m \times m$ orthogonal matrix
- The values s_i are known as the *singular values* of $\mathbf{X}$
- The SVD of a symmetric matrix is just the eigen-decomposition
- **Economical SVD**
  - If $n > m$ some algorithms won't bother outputting the last $n - m$ columns of $\mathbf{U}$
  - If $m < m$ some algorithms won't bother outputting the last $m - n$ columns of $\mathbf{V}$
  - In this case it will output a square $\mathbf{S}$ matrix

2.2 General Linear Mapping

- Recall that matrices are the most general linear operators
- Since any matrix $\mathbf{M}$ can be written as $\mathbf{U}\,\mathbf{S}\,\mathbf{V}^\mathsf{T}$ we can interpret any linear mapping as doing three operations
  - A rotation (with possibly a reflection) defined by $\mathbf{V}^\mathsf{T}$
  - A rescaling of each coordinate by s_i
  - A rotation (with possibly a reflection) defined by $\mathbf{U}$
- **Duality**
  - Using $\mathbf{X} = \mathbf{U}\,\mathbf{S}\,\mathbf{V}^\mathsf{T}$ then
    * $\mathbf{C} = \mathbf{X}\,\mathbf{X}^\mathsf{T} = \mathbf{U}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{U}$
    * $\mathbf{D} = \mathbf{X}^\mathsf{T}\mathbf{X} = \mathbf{V}\mathbf{S}^\mathsf{T}\mathbf{S}\mathbf{V}$
  - $\mathbf{S}\mathbf{S}^\mathsf{T}$ and $\mathbf{S}^\mathsf{T}\mathbf{S}$ are diagonal elements with non-zero diagonal elements s_i^2

2.3 Ridge Regression

- Ridge regression is linear regression with an L_2 regulariser
- Adding a regulariser $\nu \|\boldsymbol{w}\|^2$ the weights, $\boldsymbol{w}^*$, that minimise the loss function are given by $\boldsymbol{w}^* = (\mathbf{X}^\mathsf{T}\mathbf{X} + \nu\,\mathbf{I})^{-1}\mathbf{X}^\mathsf{T}\boldsymbol{y}$
- Using $\mathbf{X} = \mathbf{U}\,\mathbf{S}\,\mathbf{V}^\mathsf{T}$ then

$$\boldsymbol{w}^* = \mathbf{V}\,\bar{\mathbf{S}}^+\mathbf{U}^\mathsf{T}\boldsymbol{y}$$

  where $\bar{\mathbf{S}}^+$ is a regularised pseudo-inverse of $\mathbf{S}$ given by

$$\bar{\mathbf{S}}^+ = (\mathbf{S}^\mathsf{T}\mathbf{S} + \nu\,\mathbf{I})^{-1}\mathbf{S}$$

  - If $\nu = 0$ this is equal to the pseudo-inverse of $\mathbf{S}$
- $\bar{\mathbf{S}}^+$ is and $n \times m$ matrix which is zero everywhere except on the diagonal, where $\bar{S}^+_{ii} = \frac{s_i}{s_i^2+\nu}$
  - Note if $s_i = 0$ linear regression has an infinity of solutions and the pseudo-inverse of $\mathbf{X}$ does not exist (setting $\nu = 0$ we get $S^+_{ii} = 1/s_i$ which is not define when $s_i = 0$)
  - In the regularised case $\bar{S}^+_{ii} = 0$ (we have selected one of the solutions that minimise the squared error)
  - If $s_i \ll \nu$ then without the regularisation term the inverse is very ill-conditions while with the regularisation term $\bar{S}^+_{ii}$ will be small
  - If $s_i \gg \nu$ then $\bar{S}^+_{ii} \approx \frac{1}{s_i} = S^+_{ii}$
- Adding a L_2 regulariser means that the optimum weights, $\boldsymbol{w}^*$, will be less sensitive to the training data reducing the variance in the bias-variance dilemma

3 Exercises

3.1 Ridge regression

- Ridge regression is just linear regression with an L_2 regularier
  1. Derive the optimal weights in ridge regression
  2. Show that using $\mathbf{X} = \mathbf{U}\,\mathbf{S}\,\mathbf{V}^\mathsf{T}$ then $\boldsymbol{w}^* = \mathbf{V}\,(\mathbf{S}^\mathsf{T}\mathbf{S} + \nu\,\mathbf{I})^{-1}\mathbf{S}\,\mathbf{U}\boldsymbol{y}$
- See answers

4 Experiments

4.1 SVD

- Using Matlab/Octave or python have a play with svd

```
X = randn(3,4)        % construct a random matrix
[U,S,V] = svd(X)      % compute singular value decomposition
U*S*V'                % should be the same as X
U*U'                  % should be the identity up to round error
U'*U                  % should be the identity up to round error
V*V'                  % should be the identity up to round error
V'*V                  % should be the identity up to round error
[Ue,L1] = eig(X*X') % Ue should be the same as U up to permutation
S*S'                  % same as L1 up to permutation
[Ve,L2] = eig(X'*X) % Ve should be the same as V up to permutation
```

```
S'*S                  % same as L2 up to permutation


inv(X'*X + 0.1*eye(4))       % check identity
V*inv(S'*S + 0.1*eye(4))*V'  % should be the same
```

4.2 Verify Identity

- Again use Matlab/Octave or python
- For a random 4×5 matrix $\mathbf{X}$
  - Check that using $\mathbf{X} = \mathbf{U}\,\mathbf{S}\,\mathbf{V}^\mathsf{T}$ that

    $$(\mathbf{X}^\mathsf{T}\mathbf{X} + \eta\,\mathbf{I})^{-1} = \mathbf{V}\,(\mathbf{S}^\mathsf{T}\mathbf{S} + \nu\,\mathbf{I})^{-1}\mathbf{V}^\mathsf{T}$$

    holds for some random matrix using Matlab/Octave or python
  - Examine $\mathbf{S}^\mathsf{T}\mathbf{S}$, $\mathbf{S}^\mathsf{T}\mathbf{S} + 0.1\,\mathbf{I}$. $(\mathbf{S}^\mathsf{T}\mathbf{S} + 0.1\,\mathbf{I})^{-1}$ and $(\mathbf{S}^\mathsf{T}\mathbf{S} + 0.1\,\mathbf{I})^{-1}\mathbf{S}^\mathsf{T}$
  - See if you can invert $\mathbf{X}^\mathsf{T}\mathbf{X}$: it is singular, but due to rounding errors it might be inverted (it was a scary matrix when I tried it)

```
X = randn(4,5)               % construct a random matrix
[U,S,V] = svd(X)             % compute singular value decomposition

inv(X'*X + 0.1*eye(5))       % check identity
V*inv(S'*S + 0.1*eye(5))*V'  % should be the same

S'*S                         % singular
S'*S + 0.1*eye(5)            % now invertible
inv(S'*S + 0.1*eye(5))
inv(S'*S + 0.1*eye(5))*S'    % 4x5 diagonal matrix

inv(X'*X)                    % shouldn't be able to do this
```

5 Answers

5.1 Ridge regression

1. It is straightforward to show

   $$\boldsymbol{w}^* = (\mathbf{X}^\mathsf{T}\,\mathbf{X} + \nu\,\mathbf{I})^{-1}\mathbf{X}^{-1}\boldsymbol{y}$$

2. The only hard part is to show is that

   $$(\mathbf{X}^\mathsf{T}\mathbf{X} + \nu\,\mathbf{I})^{-1} = \mathbf{V}\,(\mathbf{S}^\mathsf{T}\mathbf{S} + \nu\,\mathbf{I})^{-1}\mathbf{V}^\mathsf{T}$$

   - It is easy to show that $\mathbf{X}^\mathsf{T}\mathbf{X} = \mathbf{V}\,\mathbf{S}^\mathsf{T}\mathbf{S}\,\mathbf{V}^\mathsf{T}$
   - But we also have $\mathbf{I} = \mathbf{V}\,\mathbf{V}^\mathsf{T}$ as $\mathbf{V}$ is an orthogonal matrix
   - Thus $\mathbf{M} = \mathbf{X}^\mathsf{T}\mathbf{X} + \nu\,\mathbf{I} = \mathbf{V}\,(\mathbf{S}^\mathsf{T}\mathbf{S} + \nu\mathbf{I})\mathbf{V}^\mathsf{T} = \mathbf{V}\,\mathbf{W}\,\mathbf{V}^\mathsf{T}$ where $\mathbf{W} = \mathbf{S}^\mathsf{T}\mathbf{S} + \nu\mathbf{I}$
   - But $(\mathbf{A}\,\mathbf{B}\,\mathbf{C})^{-1} = \mathbf{C}^{-1}\mathbf{B}^{-1}\mathbf{A}^{-1}$ (which we can verify by multiplying $\mathbf{C}^{-1}\mathbf{B}^{-1}\mathbf{A}^{-1}$ on either the left or right by $\mathbf{A}\,\mathbf{B}\,\mathbf{C}$)
   - Thus $\mathbf{M}^{-1} = (\mathbf{V}\,\mathbf{W}\,\mathbf{V}^\mathsf{T})^{-1} = (\mathbf{V})^{\mathsf{T}\,-1}\,\mathbf{W}^{-1}\mathbf{V}^{-1} = \mathbf{V}\,\mathbf{W}\mathbf{V}^\mathsf{T}$ where we use $\mathbf{V}^{-1} = \mathbf{V}^\mathsf{T}$ as $\mathbf{V}$ is an orthogonal matrix