ML@B NMEP - HW 1

Arvind Rajaraman

February 11, 2021

1 Basic Computations

a) $\begin{bmatrix}1\\2\end{bmatrix}\begin{bmatrix}3\\2\end{bmatrix}=\begin{bmatrix}4\\2\end{bmatrix}$.

b) $\begin{bmatrix}1\\2\\3\end{bmatrix}\begin{bmatrix}3\\2\\1\end{bmatrix}=\begin{bmatrix}4\\2\\0\end{bmatrix}=\begin{bmatrix}8\\6\\4\end{bmatrix}$.

c) $\begin{bmatrix}0&1\\1&0\end{bmatrix}\begin{bmatrix}1\\2\end{bmatrix}=\begin{bmatrix}0(1)+1(2)\\1(1)+0(2)\end{bmatrix}=\begin{bmatrix}2\\1\end{bmatrix}$.

2 Linear Transformations

a) $\mathrm{C}(A)=\mathrm{span}(\begin{bmatrix}1\\-1\\0\end{bmatrix},\begin{bmatrix}2\\4\\6\end{bmatrix},\begin{bmatrix}3\\2\\5\end{bmatrix})$. $\mathrm{N}(A)=\mathrm{span}(\begin{bmatrix}-\frac{4}{3}\\-\frac{5}{6}\\1\end{bmatrix})$.

b) Linear transformations on a space can be fully represented by the transformed basis vectors of that space. If A_1 is the matrix that describes T_1's transformed basis vectors and A_2 for T_2, we know that matrix multiplication is not necessarily commutative. In other words, $A_1A_2 \neq A_2A_1$ in general. Thus, performing linear transformations in a different order on $\mathbf{x}$ can lead to different results.

c) Not necessarily. Here is a counterexample: $A_1=\begin{bmatrix}1&1\\1&1\end{bmatrix}$ and $A_2=\begin{bmatrix}1&2\\-1&-2\end{bmatrix}$. For a vector $\mathbf{x}=\begin{bmatrix}x_1\\x_2\end{bmatrix}$, $A_1A_2\mathbf{x}=\begin{bmatrix}1&1\\1&1\end{bmatrix}\begin{bmatrix}1&2\\-1&-2\end{bmatrix}\begin{bmatrix}v_1\\v_2\end{bmatrix}=\begin{bmatrix}0&0\\0&0\end{bmatrix}\begin{bmatrix}v_1\\v_2\end{bmatrix}=\begin{bmatrix}0\\0\end{bmatrix}$.

3 Least Squares, Projection

a) $A^TA\mathbf{x}=A^T\mathbf{b}$:

$$\begin{bmatrix}3&0&-6\\0&24&24\\-6&24&36\end{bmatrix}\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix}=\begin{bmatrix}9\\12\\6\end{bmatrix}.$$

$$\mathrm{rref}(A^TA)=\begin{bmatrix}1&0&-2&0\\0&1&1&0\\0&0&0&1\end{bmatrix}.$$

$x_1-2x_3=3 \implies x_1=2x_3+3$.
$x_2+x_3=\frac{1}{2} \implies x_2=-x_3+\frac{1}{2}$.

$$\mathbf{x}=\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix}\in\boxed{\{\begin{bmatrix}2\alpha+3\\-\alpha+\frac{1}{2}\\\alpha\end{bmatrix}|\alpha\in\mathbb{R}\}}.$$

b) $\text{proj}_V(\mathbf{b}) = A[(A^TA)^{-1}A^T\mathbf{b}] = \begin{bmatrix} 1 & 2 & 0 \\ -1 & 4 & 6 \\ 1 & 2 & 0 \end{bmatrix} \begin{bmatrix} 2\alpha + 3 \\ -\alpha + \frac{1}{2} \\ \alpha \end{bmatrix} = \begin{bmatrix} 2\alpha + 3 - 2\alpha + 1 + 0\alpha \\ -2\alpha - 3 - 4\alpha + 2 + 6\alpha \\ 2\alpha + 3 - 2\alpha + 1 + 0\alpha \end{bmatrix} = \boxed{\begin{bmatrix} 4 \\ -1 \\ 4 \end{bmatrix}}.$

c) $\text{dist}(\mathbf{b}, \text{span}(\mathbf{v_1}, \mathbf{v_2})) = \|\mathbf{b} - \text{proj}_V(\mathbf{b})\| = \left\| \begin{bmatrix} 3 \\ -1 \\ 5 \end{bmatrix} - \begin{bmatrix} 4 \\ -1 \\ 4 \end{bmatrix} \right\| = \left\| \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} \right\| = \sqrt{1^2 + 0^2 + 1^2} = \boxed{\sqrt{2}}.$

4 Ridge Regression Derivation

Here, we compute the gradient of the loss function:

$$\begin{aligned} &\nabla(\|X\mathbf{w} - Y\|_2^2 + \lambda \|\mathbf{w}\|_2^2) \\ =&\nabla \|X\mathbf{w} - Y\|_2^2 + \nabla\lambda \|\mathbf{w}\|_2^2 \\ =&\nabla(X\mathbf{w} - Y)^T(X\mathbf{w} - Y) + \lambda\nabla\mathbf{w}^T\mathbf{w} \\ =&2X^T(X\mathbf{w} - Y) + \lambda(2\mathbf{w}) = 0. \end{aligned}$$

Here, we isolate $\mathbf{w}$ to find the optimal solution:

$$\begin{aligned} 2\mathbf{w}X^TX - 2X^TY + 2\lambda\mathbf{w} &= 0 \\ \mathbf{w}X^TX - X^TY + \lambda\mathbf{w} &= 0 \\ \mathbf{w}(X^TX + \lambda I) &= X^TY \\ \mathbf{w} &= \boxed{(X^TX + \lambda I)^{-1}X^TY}. \end{aligned}$$

By choosing different values for λ, we can penalize the parameters in $\mathbf{w}$ for having wildly different values (which contribute to high variance). By increasing λ, we penalize high values more, which reduces the variance of the model. By decreasing λ, we keep the model more intact and allow for a more complex decision boundary.