Math HW

February 10, 2020

1 Basic Computations

These are boring, but its good to know how vectors and matrix vector products work by just the numbers.

- Compute the sum $\begin{bmatrix} 1 \\ 2 \end{bmatrix} + \begin{bmatrix} 3 \\ 2 \end{bmatrix}$.
- Compute the sum $\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} + \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix} + \begin{bmatrix} 4 \\ 2 \\ 0 \end{bmatrix}$.
- Compute the value of $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix}$.

2 Linear Transformations

- Compute the column space and null space of the linear transformation $\begin{bmatrix} 1 & 2 & 3 \\ -1 & 4 & 2 \\ 0 & 6 & 5 \end{bmatrix}$. Express your answer as the span of some vectors.
- For two linear transformations T_1 and T_2, is $T_1(T_2(\mathbf{v})) = T_2(T_1(\mathbf{v}))$ always true for all $\mathbf{v}$? Explain why, and assume there are no issues with domain/range stuff.
- If two linear transformations T_1 and T_2 satisfy $T_1(T_2(\mathbf{v})) = \mathbf{0}$ for all $\mathbf{v}$, does one of T_1 or T_2 have to be the linear transformation that maps all vectors to $\mathbf{0}$? Assume there are no issues with domain/range stuff.

3 Least Squares, Projection

- Compute $\mathbf{x}$ such that $||\mathbf{Ax} - \mathbf{b}||$ is minimized, where $A = \begin{bmatrix} 1 & 2 & 0 \\ -1 & 4 & 6 \\ 1 & 2 & 0 \end{bmatrix}$, $b = \begin{bmatrix} 3 \\ -1 \\ 5 \end{bmatrix}$, and the norm is the L2 norm.
- Using the previous question, compute the projection of $\mathbf{b}$ onto the the plane spanned by $\mathbf{v_1}$ and $\mathbf{v_2}$, where $\mathbf{v_2} = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}$ and $\mathbf{v_2} = \begin{bmatrix} 2 \\ 4 \\ 2 \end{bmatrix}$.
- Using the previous parts, what is the distance from b to $\text{span}\{\mathbf{v_1}, \mathbf{v_2}\}$?

4 Ridge Regression Derivation

We mentioned during lecture that one of the caveats of OLS was the assumption that our input matrix, X, is full rank. However, when the features of our data are close to collinear, X might lose rank or have singular values very close to 0. This means $(X^T X)^{-1}$ will have extremely large singular values resulting in abnormally high values in the optimal w solution (our parameters).

However, there is a very simple solution for this!

$$\min_{w} \|\mathbf{Xw} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{w}\|_2^2$$

By adding a penalty term with a fixed small scalar $\lambda > 0$ (this is a hyperparameter!), we can prevent w from becoming too large. Make sure you understand why this is the case.

In lecture we defined our OLS loss function to be:

$$L(w) = \|\mathbf{Xw} - \mathbf{y}\|_2^2$$

Our new loss function with the penalty term is:

$$L(w) = \|\mathbf{Xw} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{w}\|_2^2$$

Using vector calculus, derive the optimal solution w for the ridge regression loss function. (hint: calculate the gradient!). Also, explain how we might tune the λ hyperparameter to find the best solution.