# Linear Data Analysis
# Quiz 2 Review

Cain Susko

Queen's University
School of Computing

February 17, 2022

# Main Concepts

- Standardization of data
- Projecting a vector to a subspace
- Linear regression (essentially vector projection)
- Validation of linear regression
- cross-validation by $k$-fold training and testing
- Singular-value decomposition (SVT)
- How the SVD can be used to describe matrix subspaces

# Vector Standatdization

1. Transform vector $\vec{a} \in \mathbb{R}^m$ to zero-mean $\vec{m}$
2. Transform $\vec{m}$ to unit variance $\vec{z}$

This is done by:

$$\overline{a} = \frac{\vec{1}^\top \vec{a}}{m} \qquad \text{Find the mean of a vector in a matrix}$$

$$\vec{m} = \vec{a} - \vec{1}\overline{a} \qquad \text{find the zero-mean from the mean}$$

$$\sigma^2 = \frac{||\vec{m}||^2}{m-1} \qquad \text{find the unit variance of } \vec{a} \text{ from the mean}$$

$$\vec{z} = \frac{\vec{m}}{\sigma} \qquad \text{derive the z-score/z-vector for } \vec{a}.$$

Note: $m$ is the size of the vector; $\vec{m}$ is the zero mean of a vector. Additionally, note that it is $\sigma$ used in the final equation, not $\sigma^2$

# Projection

Projection is the process of taking a vector $\vec{c}$ to subspace $\mathbb{U}$, spanned by the vectors $\vec{a}_j$. The result of this process is a vector $\vec{p}$.

$$\begin{aligned} &\vec{e} =^{defined} \vec{c} - \vec{p} && \text{error vector} \\ &\vec{w} && \text{the weight vector} \\ &[A^\top A]\vec{w} = A^\top \vec{c} && \text{find } \vec{w} \text{ using the normal equation} \end{aligned}$$

.

Thus, projection is:

$$\vec{p} = A\vec{w} = A[A^\top A]^{-1} A^\top \vec{c} = P\vec{c}.$$

Note: there is a special case where: if $A$ is singular, use a basis of $\mathbb{U}$ rather than an ordinary span of $\mathbb{U}$ like $\vec{a}_j$

# Linear Regression

Linear Regression is the process of projecting observations $\vec{c}$ to the column space of $A$.

$$\begin{aligned} &\vec{w} && \text{weight vector} \\ &\vec{e}(\vec{w}) = \vec{c} - A\vec{w} && \text{error vector} \\ &RMS(A, \vec{c}; \vec{w}) = \frac{||\vec{e}(\vec{w})||}{\sqrt{m}} && \text{error of } \vec{e} \text{ using Root Mean Square} \end{aligned}$$

.

# Cross Validation

Validation is the process of confirming that the outputs of a model are acceptable using RMS error (see above). Cross validation is the process of dividing data into a training and validation sets and then validating the model $\vec{w}$ using both sets. In $k$-fold cross validation, divide the data into $k$ sets we then:

1. Train on $k-1$ sets and validate on 1 set, for all sets
2. accumulate RMS errors for analysis

# Singular Value Decomposition

for any matrix $A \in \mathbb{R}^{m \times n}$ with all real entries of rank $r$, the SVD of $A$ is:

$$A = U \Sigma V^{\top}.$$

where:

- $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ are orthogonal
- $\Sigma \in \mathbb{R}^{m \times n}$ is 'diagonal', the singular values (the values in this matrix) are:
$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0.$$
- $\vec{u}_j$ of $\sigma_j \neq 0$ is the basis for $\mathbb{U}$ which is the column space of $A$
- $\vec{v}_j$ of $\sigma_j = 0$ is the basis for the nullspace of $A$
- $\vec{u}_j$ of $\sigma_j$ is the orthogonal complement of $\mathbb{U}$, which is represented as $\perp\mathbb{U}$

# Preparation

One should have the data and working code for:

- Assignment 2 (regression and $k$-fold cross validation)
- Homework for week 1, especially **'your' data matrix**
- Homework for week 4, projections
- Homework for week 5, Cross Validation (CV) and SVD