DATA 100: Vitamin 7 Solutions

October 18, 2019

1 Dimensionality

We say that a dataset has 10 dimensions if it has

- ☑ 10 unique variables
- ☐ 10 variables, though 2 are redundant
- ☑ 11 variables, though 2 are redundant
- ☐ 10 observations

Explanation: A dataset has 10 dimensions if its columnspace has rank 10. Therefore, options 1 and 3 are correct since these datasets have a columnspace of rank 10.

2 Singular Values

Let X be an $n \times p$ matrix of rank r, $r < p$. Suppose we take the SVD of X such that $X = U\Sigma V^\top$. Which of the following statements are true?

- ☑ The diagonal entries of Σ are non-negative
- ☑ U and V are orthogonal matrices
- ☑ The r^{th} singular value is positive
- ☑ The off-diagonal entries of Σ are all 0

Explanation: See Lecture 10 slides for the properties of SVD.

3 Rank 1 Approximation

The SVD of our data matrix X equals $U\Sigma V^\top$. How can we compute the rank 1 approximation of X?

- ☐ $U\Sigma \times V^\top$
- ☐ $U\Sigma[:, 0:1] \times V^\top[:, 0:1]$
- ☑ $U\Sigma[:, 0:1] \times V^\top[0:1, :]$
- ☐ $U\Sigma[:, 0:1] \times V[0:1, :]$

Explanation: $U\Sigma[:, 0:1] \times V^\top[0:1, :]$ is the correct answer since it's the definition of the rank 1 approximation of X given its SVD.

4 SVD Error Minimization

Fill in the blank:

The rank 1 approximation of a dataset produced with the help of its SVD minimizes the error ____ to the subspace onto which we're projecting.

- ☑ perpendicular
- ☐ parallel
- ☐ adjacent

Explanation: The SVD minimizes the error perpendicular to the subspace onto which it's projecting since it minimizes the error between the projection and the original data. See Lecture 11 slides for a sketch.

5 Why Use PCA?

PCA is appropriate for EDA when:

- ☐ The data we're exploring is already low-dimensional
- ☑ The data we're exploring is inherently low rank
- ☑ We wish to identify clusters of similar observations in a high-dimensional dataset

Explanation: PCA is appropriate for exploring the data if it is high-dimensional, but we believe it to be inherently low rank. This will often facilitate the discovery of groups of similar observations in the dataset, though only if we believe that this dataset has actually contains groups. If the data is already low-rank, there's no need to reduce the rank any further through an approximation.