Data 100, Spring 2024

Exam Prep Section #11

PCA

1. Recall that principal component analysis (PCA) works by figuring out the important "directions" of the covariance matrix using singular value decomposition (SVD) applied to the design matrix. Suppose we have the SVD of a particular (centered) data matrix X with 3 features and 5 observations; that is, we have U, S, and V^T.

   If we break down these components, recall that S is a diagonal matrix that contains $\sigma_1, ..., \sigma_r$ in its diagonal entries. Both U and V are orthonormal matrices with column vectors $u_1, ..., u_r$ and $v_1, ..., v_r$ respectively.

   (a) Which of the following are true mathematical equivalencies that always hold for any data matrix X?

   ☐ A. $\sigma_1 \geq \sigma_2$
   ☐ B. $||u_1|| = ||u_2||$
   ☐ C. $\text{Var}[u_1] = \text{Var}[u_2]$
   ☐ D. $\sigma_1^2 \text{Var}[u_1] \geq \sigma_2^2 \text{Var}[u_2]$
   ☐ E. $US = XV$

   (b) Suppose that the variance captured by the first principal component is 5. Calculate σ_1.

   (c) Given that the variance captured by the first principal component is 5, what is the largest that the total variance of all the columns in X can be? Justify your answer.

(d) Suppose that we forgot to center the columns of X. Which of the following may happen?

- ☐ A. SVD will always fail; it will be unable to produce a correct decomposition $X = USV^T$.
- ☐ B. PCA will always fail; it will be unable to produce valid principal components.
- ☐ C. SVD may fail; it may be unable to produce a correct decomposition $X = USV^T$.
- ☐ D. PCA may always fail; it may be unable to produce valid principal components.

Clustering

2. Stephanie has the following question: what kinds of clusters or groups of student preferences exist within Data 100 based on their location and preference of media (e.g., movies)? To answer this, she uses survey data collected from Data 100 students!

   She one-hot encodes all the qualitative features such as country, city, etc. Then, she standardizes all the one-hot encoded "dummy" features along with all the numerical features into a matrix X. A sample of the rows and columns of X are shown below.

	pref_cartoon_character	n_movies_per_yr	lives_in_united_states
0	Mickey Mouse	6	1
1	Homer Simpson	2	0
2	Fred Flintstone	4	1
3	Mickey Mouse	2	0
4	Franklin	9	0

   The dimensionality of X is 200×190. Stephanie uses the power of `numpy` to find that the **rank of the dataset is 150**.

   (a) Stephanie wants to cluster the standardized data received in the survey as shown above. However, DataHub cannot hold all the features in memory, so she decides to use the first k principal components! Which of the following is true if Stephanie applies PCA to the standardized data shown and trains a multiple linear regression model on the first k principal components?

   - ☐ A. As k increases, the training loss decreases and then is constant.
   - ☐ B. As k increases, the training loss decreases and then increases.
   - ☐ C. As k increases, the test loss decreases and then is constant.
   - ☐ D. As k increases, the test loss decreases and then increases.

   (b) Recall that SVD returns a matrix decomposition of a standardized data matrix $X = USV^T$ that is used to generate principal components. Which of the following is true regarding this matrix decomposition?

   - ☐ A. The decomposition is unique (i.e., U, S, and V^T are unique).
   - ☐ B. It is possible that σ_1 is less than the variance of any of the individual features (that is, column vectors) x_i in X.
   - ☐ C. When using a compact form of SVD, it is possible that $UU^T \neq I$
   - ☐ D. All principal component directions are orthogonal to one another.

K-Fails

3. Recall that K-Means alternates between cluster reassignments to points (by assigning the closest cluster center for each point) and point reassignments to clusters (by assigning the cluster center to the average of all points in the cluster). By iterating between these for a period of time, we hope to converge to a reasonable clustering such that all the cluster assignments are as close together as possible, with each cluster as distinct as possible. Unlike least squares, this problem cannot be solved with a closed-form solution, and in fact, even this solution is possibly exponential in time complexity with no guarantees about convergence! Let's see how well it does on a simple example with 5 points and with 2 clusters (i.e. k = 2):

   $$(0, 1), (0, 3), (1, 2), (9, 0), (10, -1)$$

   Our initialization for the cluster centers is (0, 0) and (10, 10).

   (a) What is the perfect or ideal clustering that minimizes the distortion?

   (b) Apply an iteration of K-Means by applying the two updates mentioned above. How many points are in the wrong cluster as per the "perfect" clustering?

   (c) Apply another iteration of K-Means by applying the two updates mentioned above. How many points are in the wrong cluster as per the "perfect" clustering?

(d) Suppose that finally, we let K-Means run updates on clustering for a few thousand iterations (i.e. enough to achieve convergence in theory). Does it achieve the perfect clustering? If not, what is the issue?