Problem 1

Suppose that we have a dataset in d dimensions and we would like to do k-means clustering for some k that is much smaller than d. Instead of doing k-means on the data set directly, it is much cheaper to first reduce the dimension and then do k-means clustering since most algorithms for k-means clustering (e.g. Lloyds algorithm) take time proportional to the di- mension in every step. However, if we do dimension reducion then we may potentially get a solution whose cost is low for the projected dataset but very large with respect to the original dataset. We will show that this is not the case is we project to the best fit k-dimensional subspace.

Since the explained variance in a lower-dimensional space X is bound to be $\leq$ the complete variance in the original number of dimensions, the cost of an optimal solution in X, OPT_X, is bound to be $\leq$ the cost of an optimal solution in the original number of dimensions OPT.

$$C_X^{proj} \leq C_{OPT}$$

The distance from the projection in X to the original space is sum of square distances from the points in the original space to the corresponding points in subspace X:

$$C_{proj} = \sum_{i=1}^{n} (p_i - p_{ix})^2 = E(X)$$

Since the cost of the solution is determined by summing the square distances from the cluster centers to the points pertaining to the cluster, the cost of the optimal solution in X when calculated in the original space can be written as:

$$C_X^{orig} = C_X^{proj} + C_{proj} = C_X^{proj} + E_X$$

C_{OPT} is the cost of the optimal solution in the original space by definition. Another way to write C_{OPT} is C_T^{orig} where T is the subspace defined by the k cluster centers in OPT - the optimal solution in the original space, or OPT_T.

$$C_T^{orig} = C_{OPT}$$

Note that by definition, $E_T = C_T^{proj}$ and since $C_X^{orig} = C_X^{proj} + E_X$, for T this results in $C_T^{orig} = 2C_T^{proj} = 2C_{OPT}$. Define S as the best-fit k-dimensional subspace for the original space computed using SVD. Since the orthonormal basis v_i of S are linear combinations of the orthonormal basis vectors of T such that $\sum_{i=1}^{k} |X \cdot v_i|^2$ is maximized, the error in projection for S E_S is $< E_T$, and $E_S = E_T$ when $S = T$. In general then,

$$E_S \leq E_T$$

Using the above result,

$$C_S^{orig} \leq 2C_{OPT}$$

Problem 2

In the greedy algorithm we presented in class for computing best fit subspaces, we assumed that the best fit (k +1)-dimensional subspace for a data set contains the best fit k-dimensional subspace for the data set. Prove that this assumption is valid.

Trivial for k = 1. For some k-dimensional best-fit subspace V, let $v_1, ..., v_k$ be an orthonormal basis. Consider a (k+1)-dimensional best-fit subspace W with its orthonormal basis $w_1, ..., w_{k+1}$, such that w_{k+1} is perpendicular to $v_1, ..., v_k$ by taking a vector perpendicular to the projections of $v_1, ..., v_k$ into W. If $v_1, ..., v_{k+1}$ maximize $\sum_{i=1}^{k} |X \cdot v_i|^2$ then,

$$\sum_{j=1}^{k+1} |X \cdot w_j|^2 \geq \sum_{i=1}^{k} |X \cdot v_i|^2$$

Hence, W_{k+1} is at least as good as V_k.

Problem 3

MNIST SVD decomposition.

Figures 1 through 5.

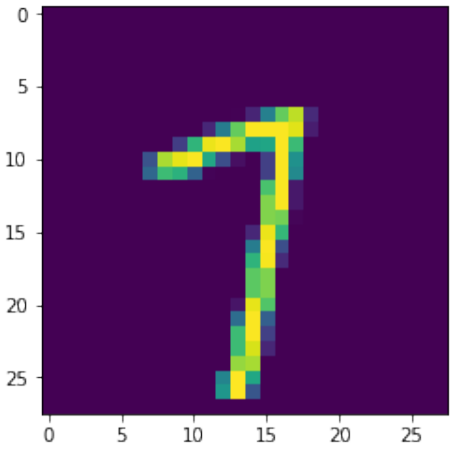

Figure 1: Example of an image (number '7' in this case) from the MNIST dataset.

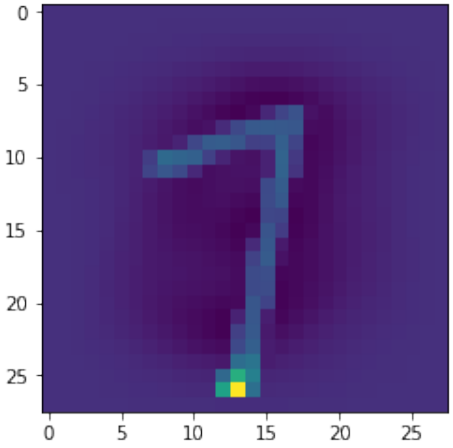

Figure 2: Example of an image (number '7' in this case) from the MNIST dataset when normalized.

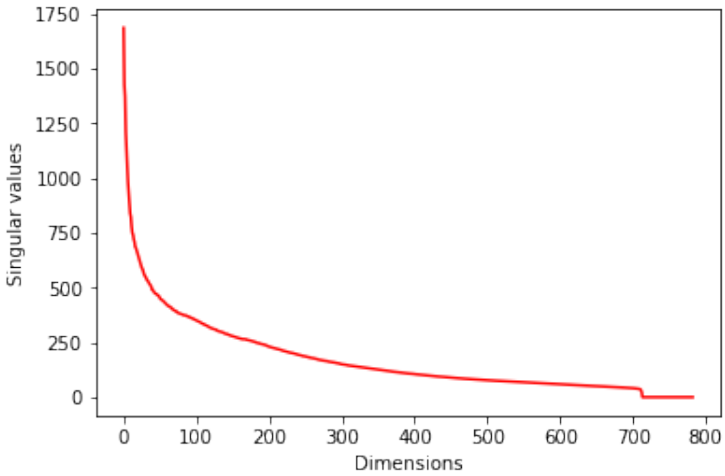

Figure 3: Singular values of the near-full (783/784 dimensions) SVD of the MNIST dataset.

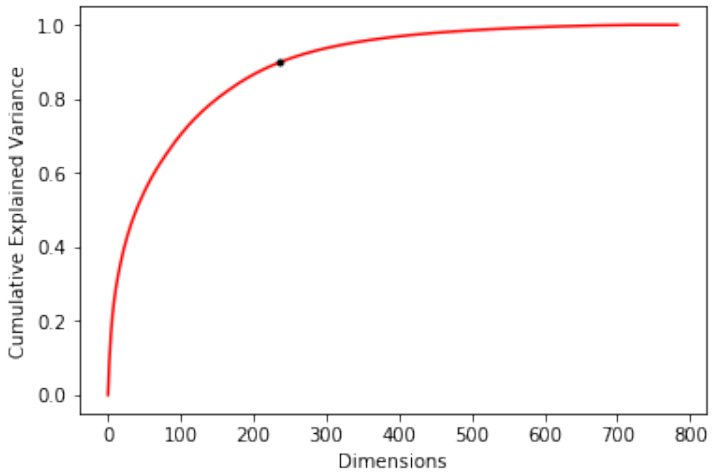

Figure 4: The top 237 most important dimensions in the SVD of the MNIST dataset explain 90% of the variance. The explained variance ratio for each singular value in decreasing order was added until the cumulative explained variance added up to 0.9, which happened to be at the 237^{th} singular value.

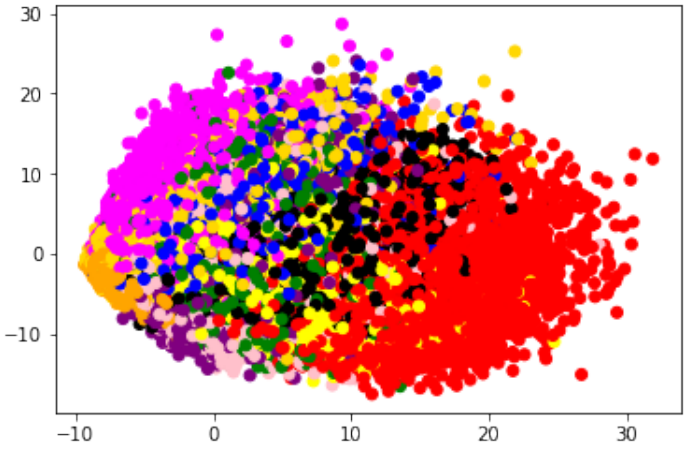

Figure 5: 2-dimensional (9.6% explained variance) SVD decomposed representation of the MNIST dataset color coded by class.

Problem 4

Consider a set of data points in the plane where each point is labelled either 0 or 1. Suppose that the two classes are separable by a circle i.e. there exists a circle so that all points with one of the labels lie inside the circle and all the points with the other label lie outside the circle. Prove that if we map each point (x, y) in the dataset to the point (x, y, $x^2 + y^2$) in three dimensions then the two classes become linearly separable in three dimensions.

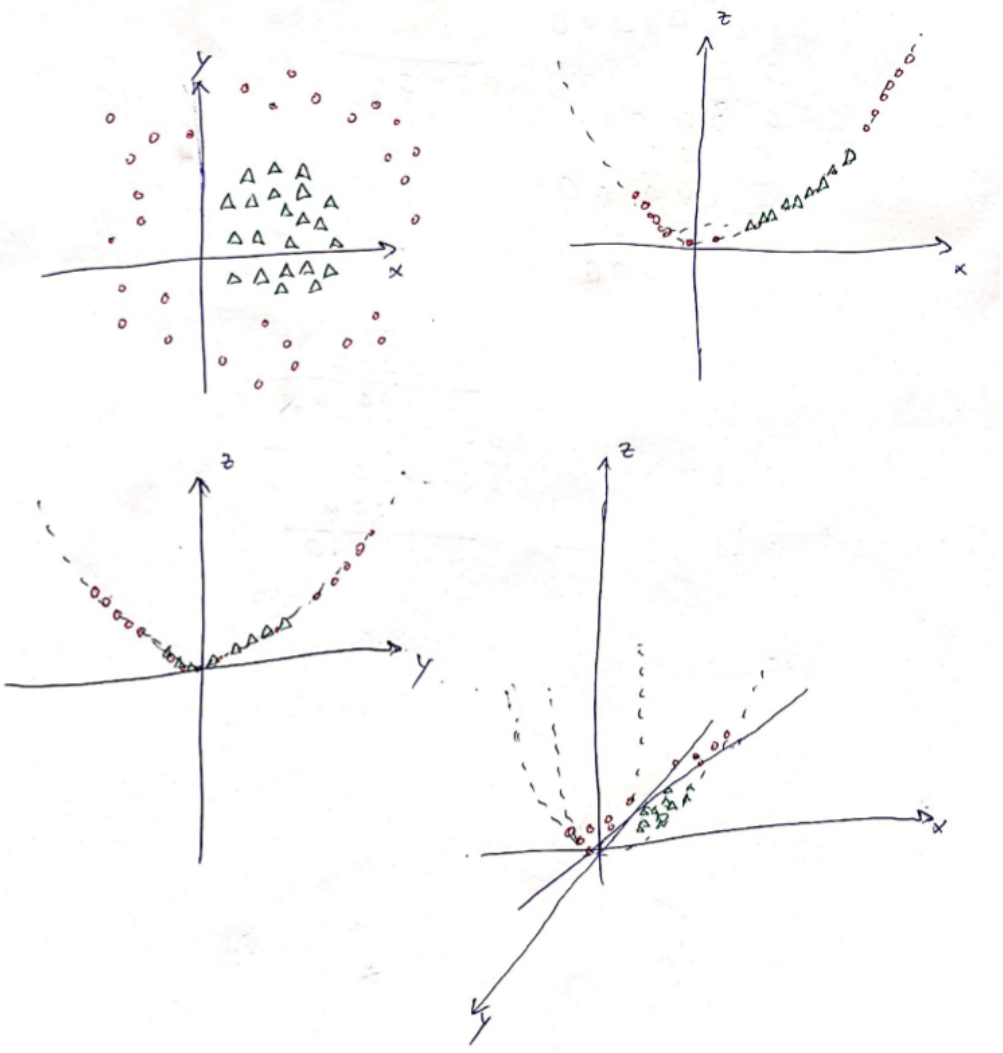

Figure 6: Two classes separable by a circle in 2D or a plane 3D.

Since any ellipse centered at the origin can be defined by some parameters a and b:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

And specifically for circles $a^2 = b^2$:

$$x^2 + y^2 = a^2$$

where a is the radius of the circle.

It can be then seen that if the each point in the dataset is mapped on the z axis to a value equal to $x^2 + y^2$ where x, and y are its coordinates, the height of each point will then reflect the magnitude of the radius of the circle centered at the origin with the point on its circumference. Since z is monotonically increasing in x and y,

Since $x^2 + y^2 - a^2 = 0$ has a scalar coefficient for term $x^2 + y^2$ with a scalar intercept $-a^2$, a linear discriminator with an appropriate weight for terms $x, y, x^2 + y^2$ will be able to separate the two classes with a plane, provided they are separable by a circle in x, y

Problem 6

> *Let **K(x, x')** be a kernel function defined on pairs of vectors in R^d. Let $x_1, ..., x_n$ be n vectors in R^d. Consider the $n \times n$ matrix A in which $A_{ij} = K(x_i, x_j)$. Prove that the matrix A is symmetric and positive semidefinite.*

A is symmetric - $A_{ji} = A_{ij}$ since $K(x_j, x_i) = K(x_i, x_j)$ because dot product (between $\phi(x)$ and $\phi(x')$ in $\mathbf{K}$) is commutative.

The property of positive semi-definite for matrix A is defined as:

$$vAv^T \geq 0$$

Consider the following:

$$\begin{aligned} vAv^T &= \sum_i \sum_j v_i A_{ij} v_j && (1) \\ &= \sum_i \sum_j v_i K(x_i, x_j) v_j && (2) \\ &= \sum_i \sum_j v_i \phi(x_i) \phi(x_j) v_j && (3) \\ &= \sum_i v_i \phi(x_i) \sum_j \phi(x_j) v_j && (4) \\ &= \sum_i \Big(v_i \phi(x_i) \Big)^2 \geq 0 && (5) \end{aligned}$$