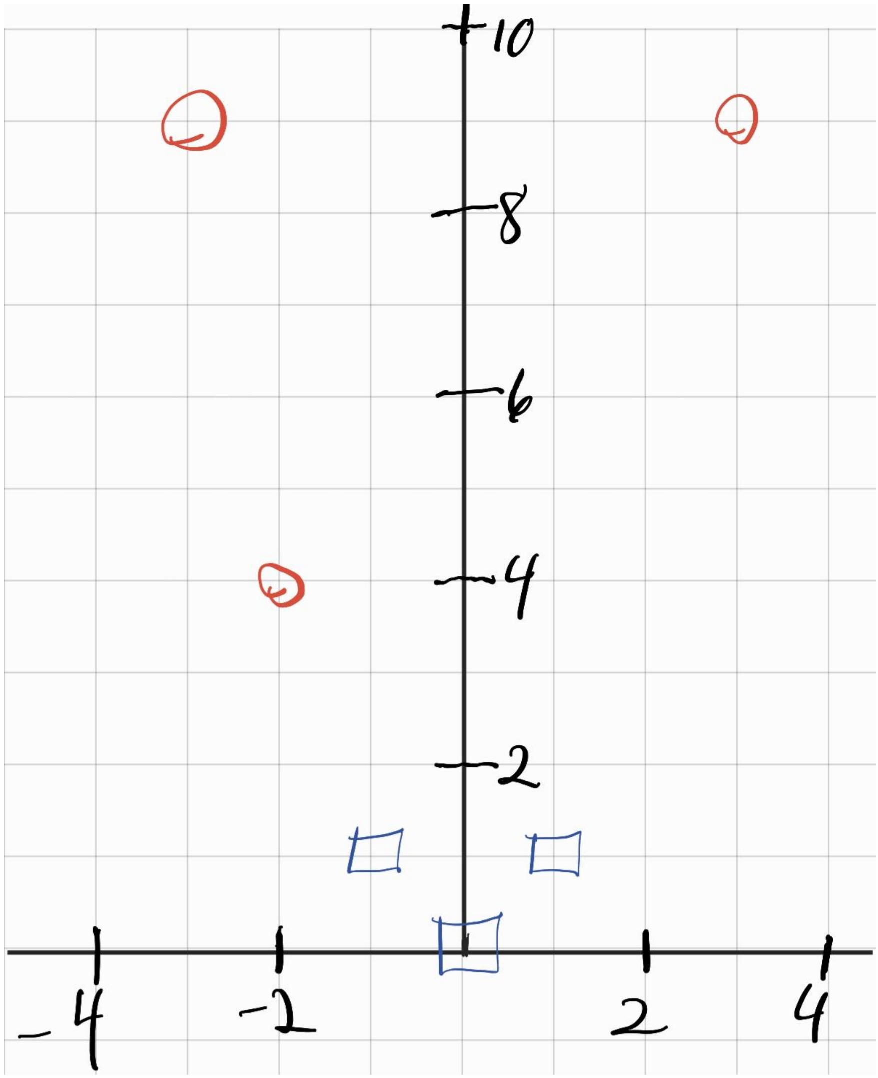
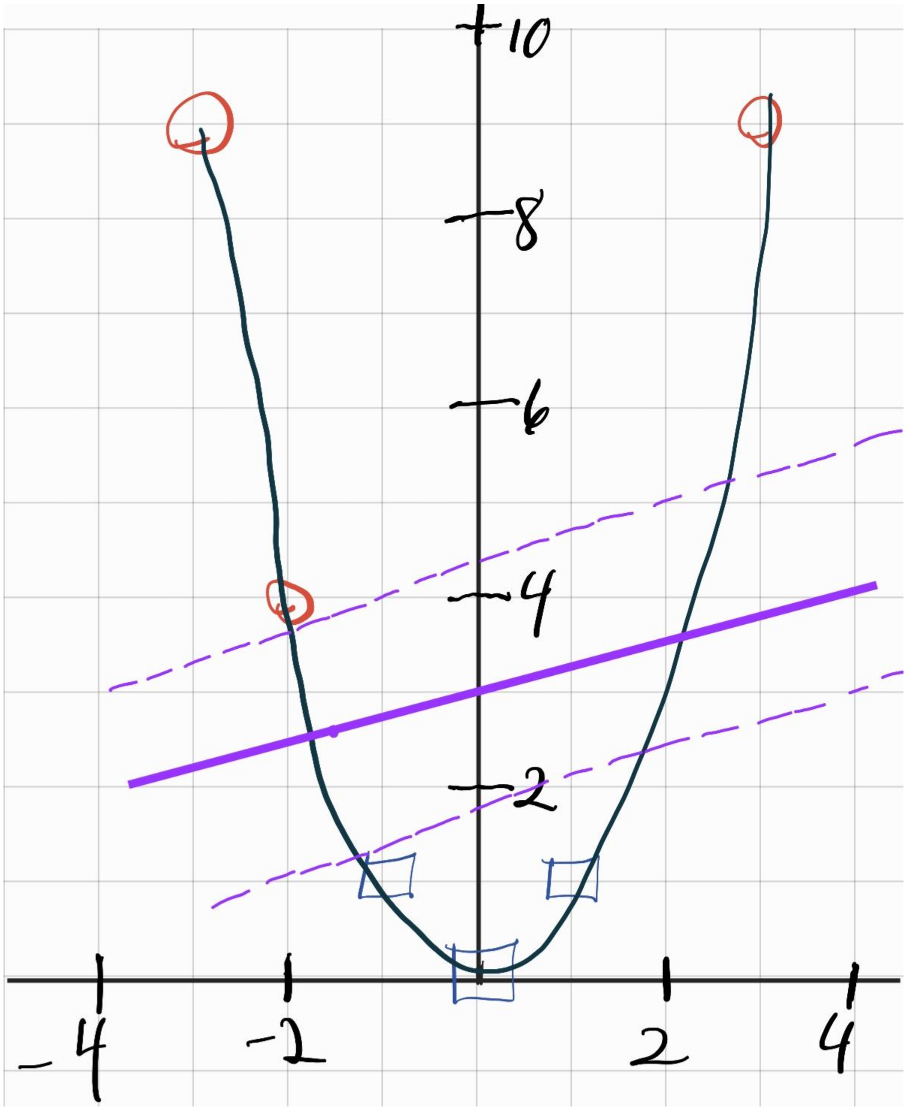
1.1 (Finite Features and SVMs)

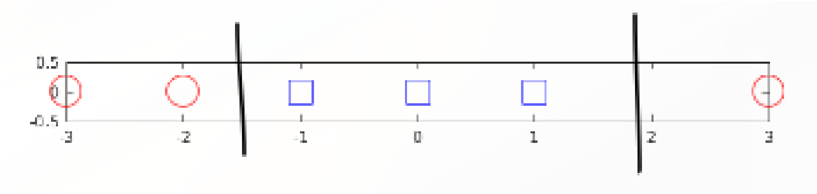
1. In its current feature space, the dataset can not be perfectly separated by a linear separator. The data's class changes twice along the one dimension and a linear classifier in a single dimension can only separate a single class change.
2. 

3. Yes, after being translated into 2 dimensional space, a linear separator can perfectly separate the points because there is now a clear path for a line to be drawn between the positive and negative samples.
4. In order to fully separate the data with a maximized margin, the line must pass through the midpoint between the two closest opposing points located at (-2,4) and (-1,1). The midpoint between the two points would be the point (-3/5, 5/2) and the hyperplane would be perpendicular to the line formed if one were to connect the points. Because of that, the line would have a slope of 1/3. Using the point and slope you can find that w_1=-1/3, w_2=1, and c=-3. The margin would be the The margin would be half of the distance between the two points which is $\sqrt{10}/2$.

5.



6.



1.2 (Infinite Features Spaces and Kernel Magic)

1. We cannot explicitly construct $\phi_{\infty}(x)$ because it is infinite-dimensional. This would mean writing down an infinite number of basis functions which is not possible.
2. Yes, $k(a, b) \; = \; \sum_{i=1}^{\infty} \frac{e^{-a^2/2} a^i}{\sqrt{i!}} \frac{e^{-b^2/2} b^i}{\sqrt{i!}}$

   $= exp[- \frac{a^2+b^2}{2}] \sum_{i=1}^{\infty} [\frac{(ab)^i}{i!}]$

   $= exp[- \frac{a^2+b^2}{2}] \, exp[ab]$

   $= exp[- \frac{1}{2} (a^2 - 2ab + b^2)]$

   $= exp[- \frac{1}{2}(a - b)^2]$
3. No, since model complexity in SVMs is not managed by feature space dimensions and instead by the number of support vectors.
4. $k(x_i + x_0, x_j + x_0) = e^{\frac{-((x_i+x_0)-(x_j+x_0))^2}{2}} = e^{\frac{-(x_i-x_j)^2}{2}} = k(x_i, x_j)$

2 (Naïve Bayes Classifier)

1. $$P(Y = 1 | Z = missing) = \frac{0.08 * \theta_{y=1}}{0.02 * (1 - \theta_{y=1}) + 0.08 * \theta_{y=1}}$$
2.
3. **Log-joint probability for a single example:**

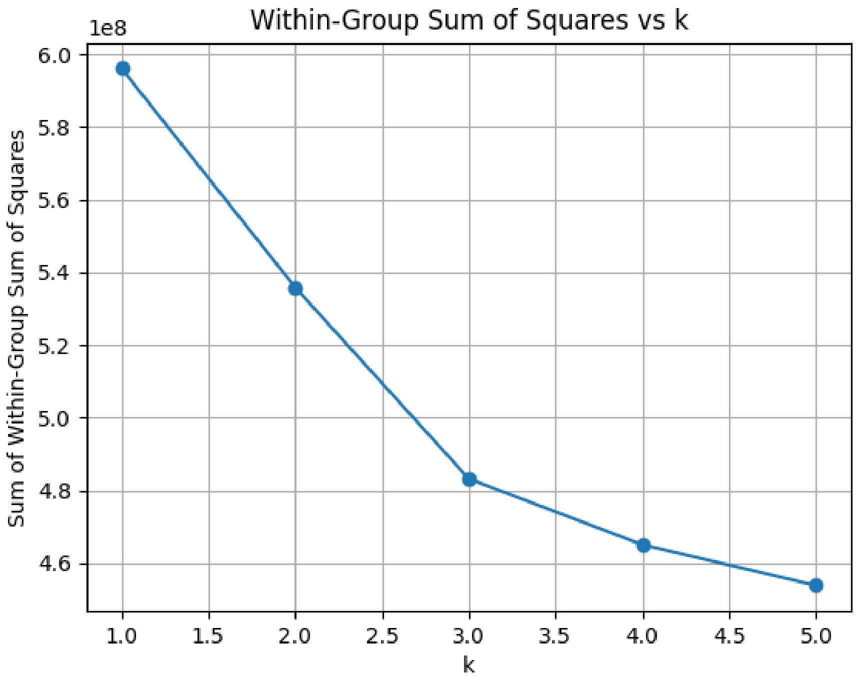
   $\log P(X_1 = x_1, X_2 = x_2, X_3 = x_3, Z = z, Y = y) = \log[p(y) \cdot p(x_1|y) \cdot p(x_2|y) \cdot p(x_3|y) \cdot P(Z = z|Y = y)]$

   For n i.i.d. examples:

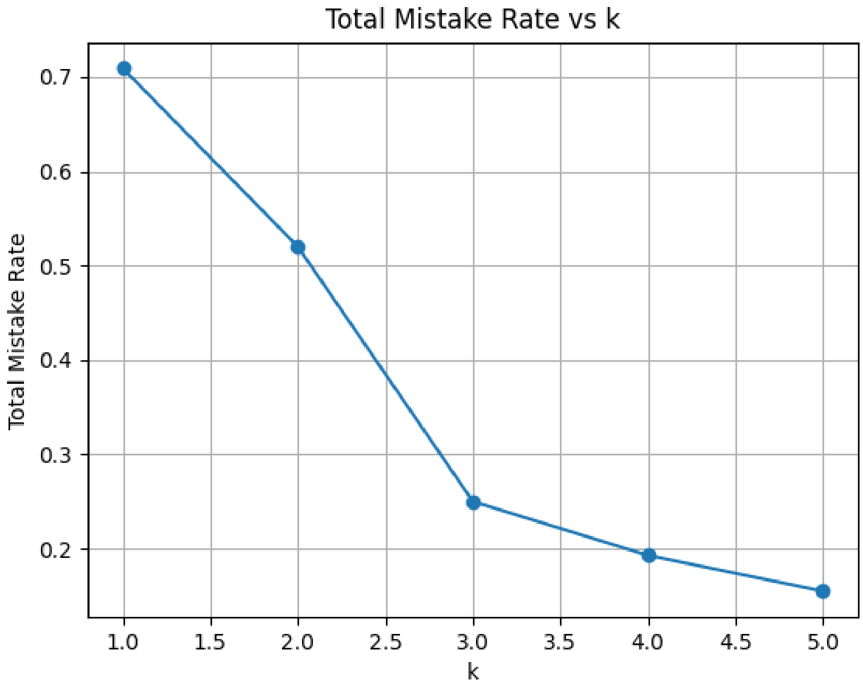
   $$\log P(X, Y, Z) = \sum_{i=1}^{n} \log[p(y_i) \cdot p(x_{i1}|y_i) \cdot p(x_{i2}|y_i) \cdot p(x_{i3}|y_i) \cdot P(Z_i = z_i|Y_i = y_i)]$$

3 (K-means Clustering)

1.

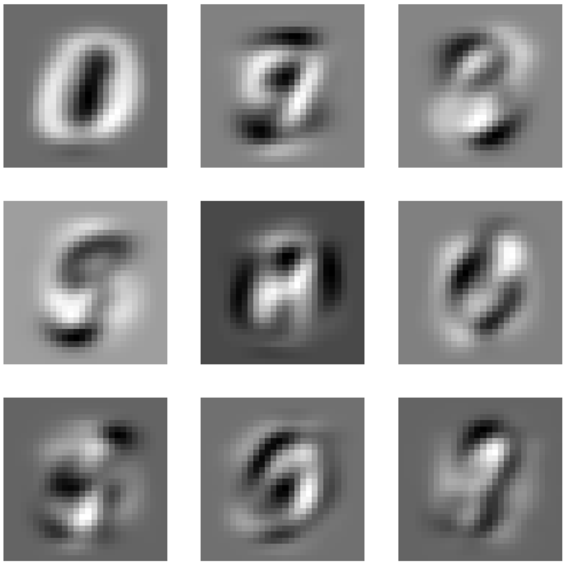


2.

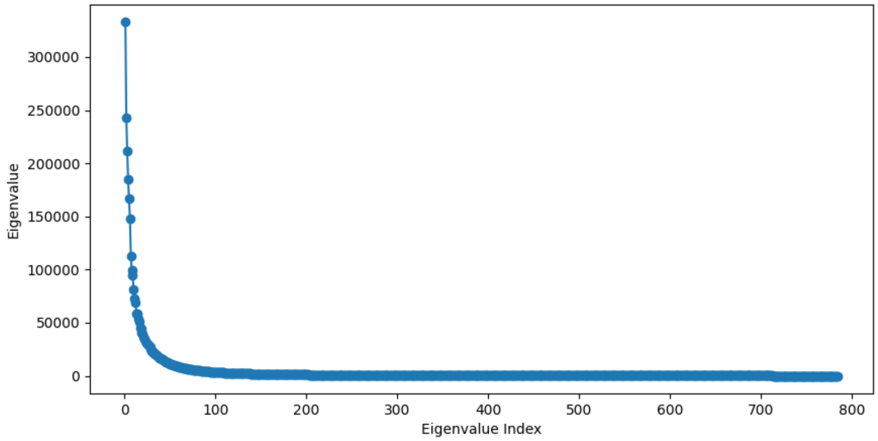


4 (Principal Components Analysis)

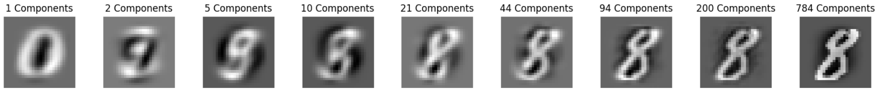
1.



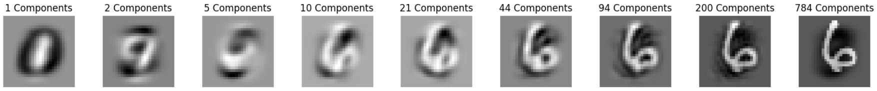
2. ~30 eigenvectors.



3. Yes, much better.



1 Components
2 Components
5 Components
10 Components
21 Components
44 Components
94 Components
200 Components
784 Components



1 Components
2 Components
5 Components
10 Components
21 Components
44 Components
94 Components
200 Components
784 Components