Homework #3

Spring 2020, CSE 446/546: Machine Learning
Richy Yun
Due: Thursday 5/28/2020 11:59 PM

Conceptual questions

A.1

a. *[2 points]* True or False: Given a data matrix $X \in \mathbb{R}^{n \times d}$ where d is much smaller than n, if we project our data onto a k dimensional subspace using PCA where $k = \text{rank}(X)$, our projection will have 0 reconstruction error (we find a perfect representation of our data, with no information loss). True. Because $k = \text{rank}(X)$ we're not actually doing any dimensionality reduction, just a linear transformation.

b. *[2 points]* True or False: The maximum margin decision boundaries that support vector machines construct have the lowest generalization error among all linear classifiers.

False. SVM is not the most optimal linear method in every situation.

c. *[2 points]* True or False: An individual observation x_i can occur multiple times in a single bootstrap sample from a dataset X, even if x_i only occurs once in X.

True. Bootstrapping is sampling with replacement, so there is a chance to have the same observation multiple times.

d. *[2 points]* True or False: Suppose that the SVD of a square $n \times n$ matrix X is $USV^\top$, where S is a diagonal $n \times n$ matrix. Then the rows of V are equal to the eigenvectors of $X^\top X$.

True. We can see that the eigenvalue decomposition is: $X^\top X = VS^\top U^\top USV^\top = V(SS^\top)V^\top$. Thus V is the eigenvectors of $X^\top X$.

e. *[2 points]* True or False: Performing PCA to reduce the feature dimensionality and then applying the Lasso results in an interpretable linear model.

False. The new feature dimension is often not easily interpretable as each feature is now a weighted combination of all the original features.

f. *[2 points]* True or False: choosing k to minimize the k-means objective (see Equation (1) below) is a good way to find meaningful clusters.

False. The higher the k the lower the objective becomes due to the points being closer to centroids if there are simply more clusters. The k that absolutely minimizes the objective is to have as many clusters as data points so that each point is a cluster, which would give an objective of 0.

g. *[2 points]* Say you trained an SVM classifier with an RBF kernel ($K(u, v) = \exp(-\frac{||u,v||_2^2}{2\sigma^2})$). It seems to underfit the training set: should you increase or decrease σ?

You should decrease σ. If the classifier was underfit it means the classification was not as specific to the train data. A larger σ amplifies the distances resulting in a more specific (more overfit) classification.

Kernels and Bootstrap

A.2 *[5 points]* We can write the dot product as:

$$\begin{aligned}
\phi(x) \cdot \phi(x') &= \sum_{i=0}^{\infty} \frac{1}{\sqrt{i!}} e^{-x^2/2} x^i \frac{1}{\sqrt{i!}} e^{-x'^2/2} x'^i \\
&= e^{-(x^2+x'^2)/2} \sum_{i=0}^{\infty} \frac{1}{i!} (xx')^i \\
&= e^{-(x^2+x'^2)/2} e^{xx'} \\
&= e^{-(x^2-2xx'+x'^2)/2} \\
&= e^{-(x-x')^2/2}
\end{aligned}$$

The step that removes the sum is due to the Taylor series expansion of e^x. Thus, $K(x, x')$ is a kernal function of this feature map.

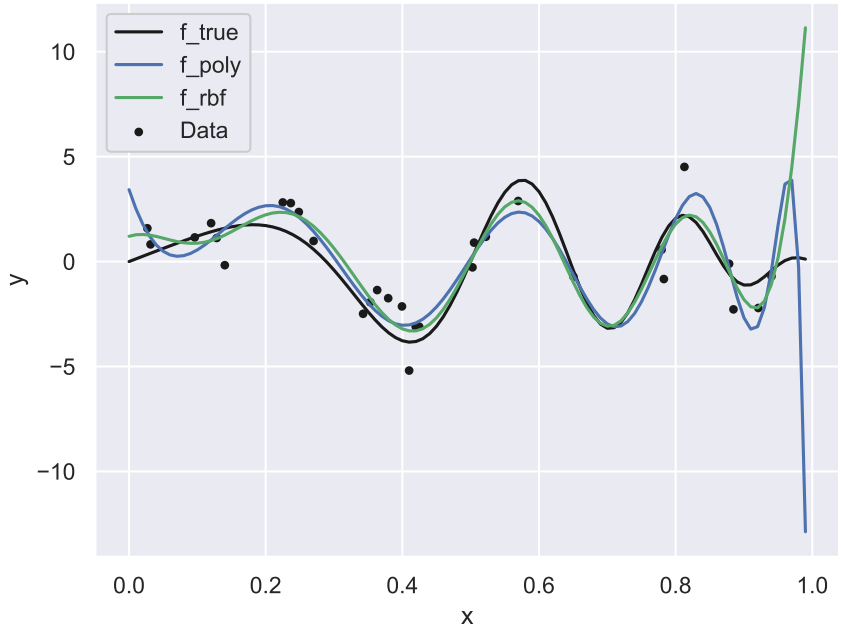
A.3

a. *[10 points]* To implement we first find the gradient of the objective with respect to α:

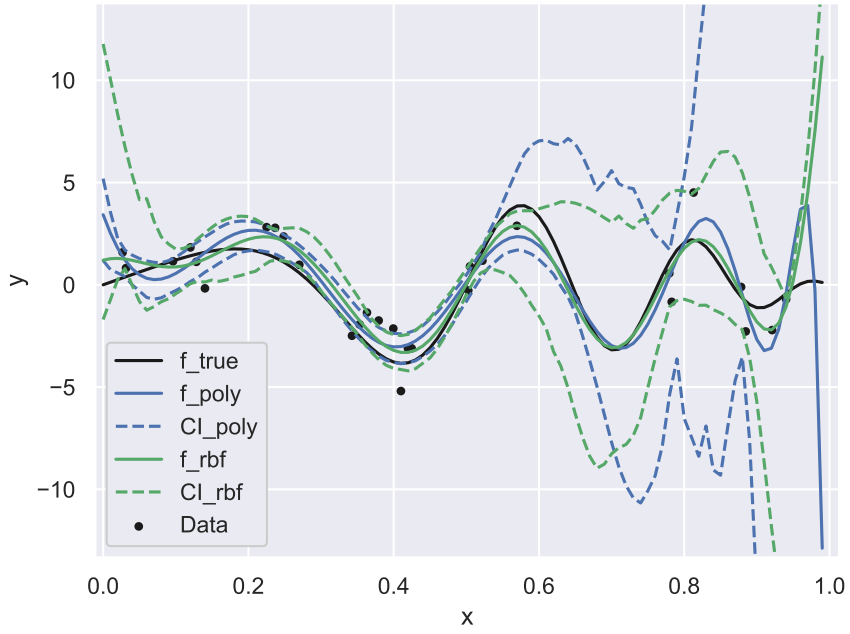
$$\begin{aligned}
\nabla_\alpha(||K\alpha - y||^2 + \lambda \alpha^\top K\alpha) &= 2K(K\alpha - y) + 2\lambda K\alpha \\
0 &= 2K(K + \lambda I)\alpha - 2Ky \\
\alpha &= (K + \lambda I)^{-1} y
\end{aligned}$$

Due to the small sample size of $n = 30$ the optimal values vary a lot trial by trial. Various trials of cross validation was performed with grid search over 50 values each of $\lambda = [1e-8, 1]$, $\doteq [1, 50]$, and $\gamma = [1, 25]$ For k_{poly}. Although it is difficult to get "the" optimal parameters, for the purposes of the problem I used what seemed to be the approximate median values that avoided overfitting: $\lambda = 1\text{e}{-4}$, $d = 25$, and $\gamma = 20$.

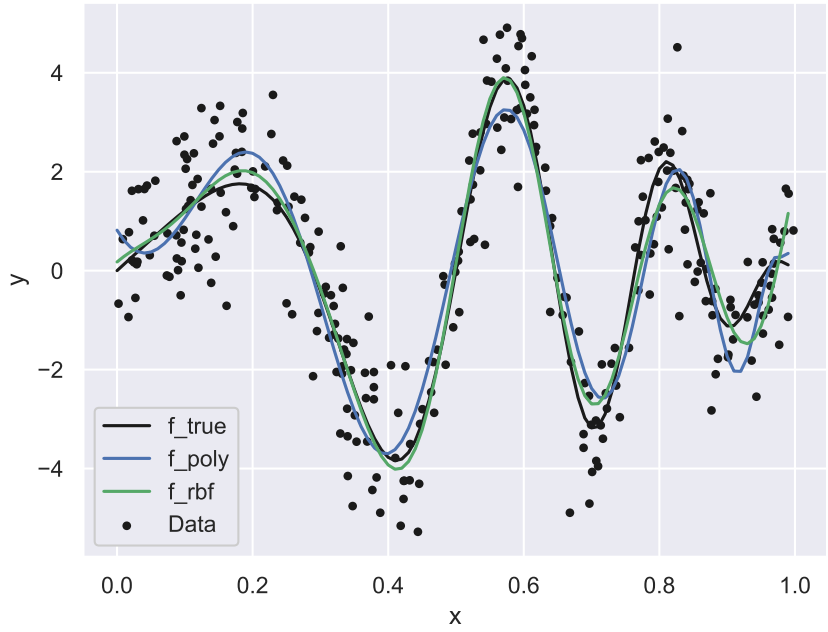
b. *[10 points]* The curve predicts the true function fairly well. As expected the extremes of the function vary a lot due to both lack of data at the extremes and the slight overfitting that occurs.

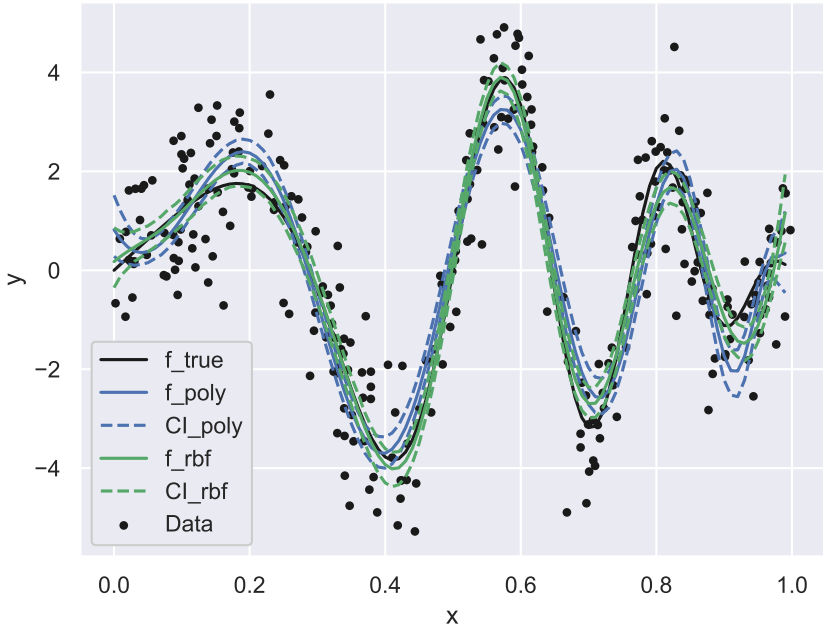


c. *[5 points]* Although the curves seem to fit fairly well, the 5 and 95% confidence intervals tell a different story. The interval is very tight where there is a lot of data points, suggesting the curve was fit very well at those spots, but begin diverging very quickly. The confidence intervals at $x = 1$ (cut off because of the zoom) go to ± 100.



d. *[5 points]* For $n = 300$ the optimal hyperparameters were: $\lambda = 5\text{e}{-3}$, $d = 30$, and $\gamma = 25$. Clearly the predictor does a much better job as it was trained with 10 times the amount of data. The confidence intervals are very tight across all x as well.





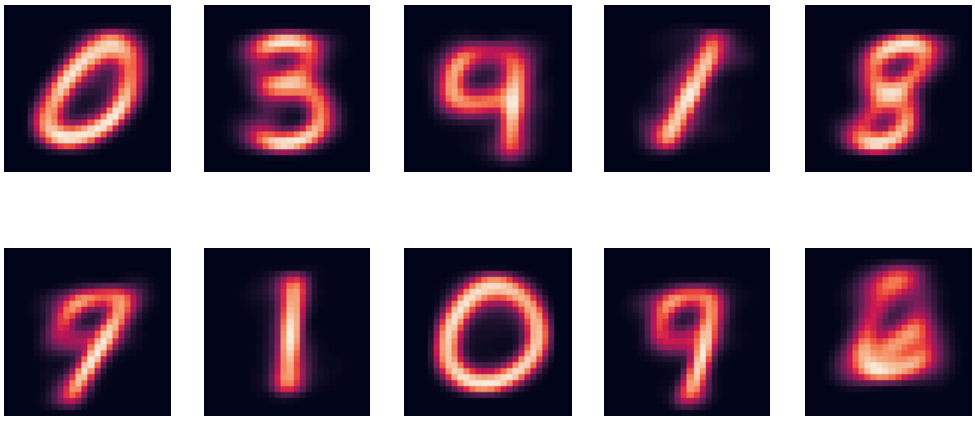
e. *[5 points]* The 5 and 95% confidence interval obtained were -0.4372 and 0.2371 respectively for this specific trial. Since the confidence interval contains 0 (which would be when the squared error is the same for both techniques on average), there is not significant statistical evidence to suggest one of $\hat{f}_{poly}$ and $\hat{f}_{rbf}$ is the better predictor.

k-means clustering

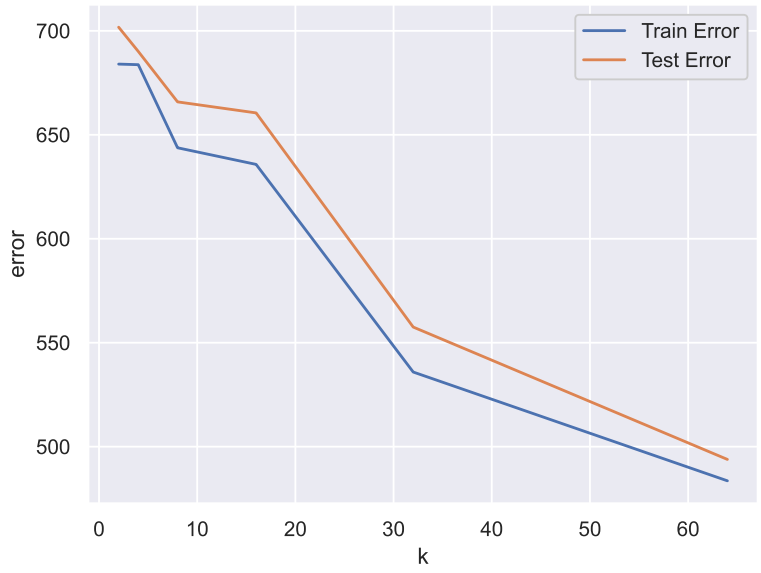
A.4

a. *[5 points]* See code at the end of problem for implementation.

b. *[5 points]* The 10 centroids look like:



c. *[5 points]* The errors with respect to k:



B.1

a. *[2 points]*

b. *[2 points]*

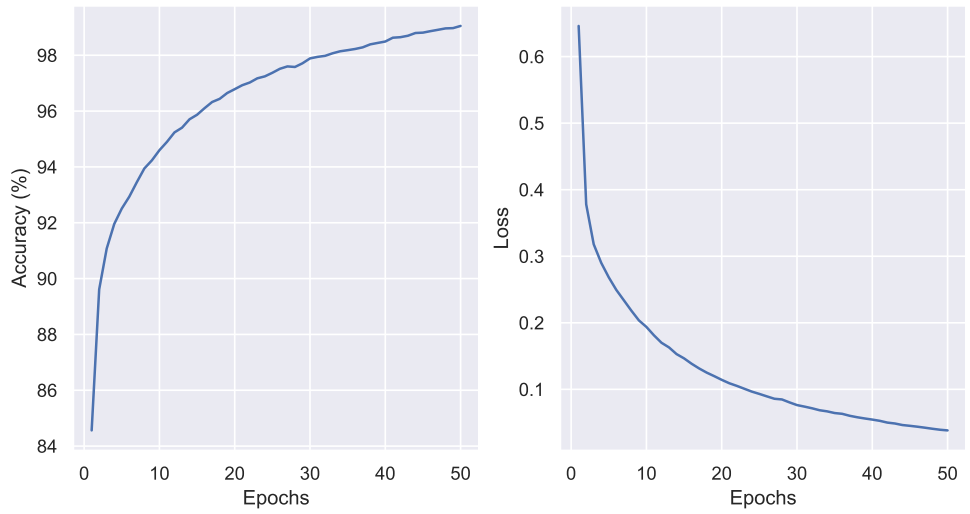
c. *[2 points]*

d. *[4 points]*

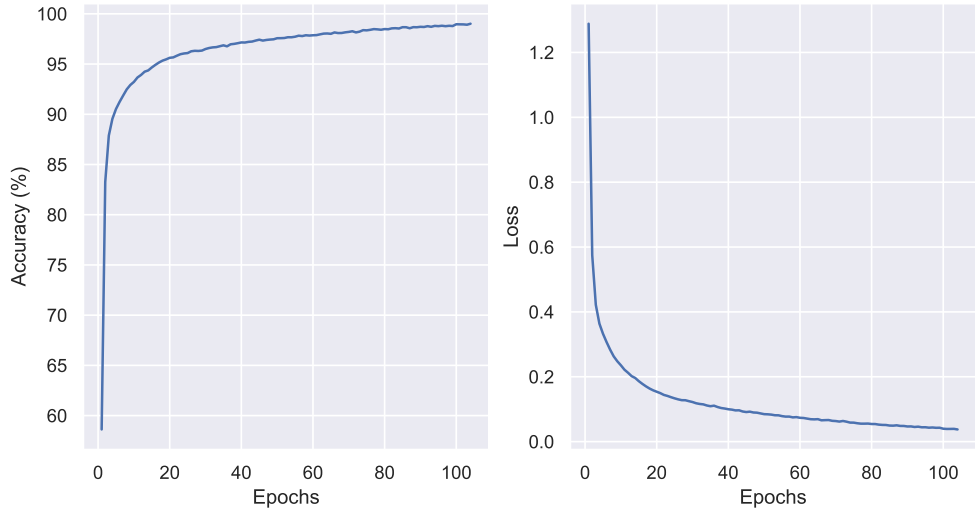
Neural Networks for MNIST

A.5

a. *[10 points]* For both a and b I used a learning rate of 1e-3 and mini-batch of 1000 samples.



b. *[10 points]*



c. *[5 points]* For a there are: $64 \times 784 + 64 + 10 \times 64 + 10 = 50890$ parameters. For b there are: $32 \times 784 + 32 + 32 \times 32 + 32 + 10 \times 32 + 10 = 26506$ parameters, around half of a. It's difficult to determine which method is explicitly "better" than the other. Although the wide shallow (a) method reaches 99% accuracy about twice as fast as the narrow deep (b) method, they do both ultimately reach it. On the other hand, the narrow deep method uses about half of the parameters the wide shallow method does meaning it doesn't require as much memory. Fewer parameters also often mean less overfitting. Thus the "better" method depends on your prioritization between speed and memory/generalization.

PCA

A.6

a. *[2 points]* Note: the images were all normalized by dividing by 255 for all parts of this problem.

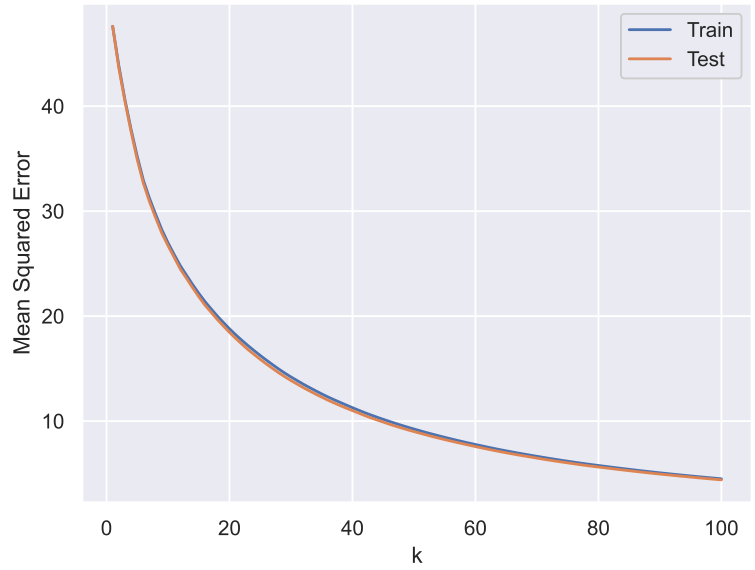
$$\begin{aligned} \lambda_1 &= 5.1168 \\ \lambda_2 &= 3.7413 \\ \lambda_{10} &= 1.2427 \\ \lambda_{30} &= 0.3643 \\ \lambda_{50} &= 0.1697 \\ \sum_{i=1}^{d} &= 52.7250 \end{aligned}$$

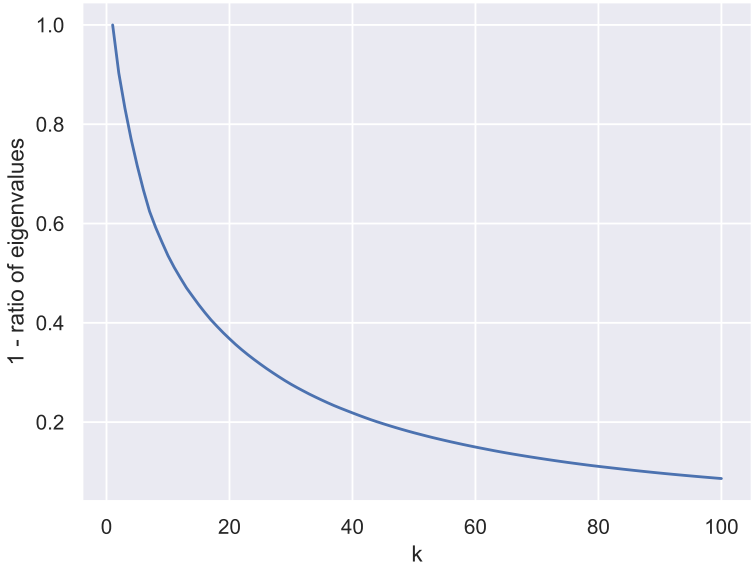
b. *[5 points]* $x \approx \mu + U_k D_k$ where U_k is the first k eigenvectors and D_k is a diagonal matrix with the first k eigenvalues. Since we know $D_k = U_K^\top (x - \mu)$ we have:

$$x \approx \mu + U_k U_k^\top (x - \mu)$$

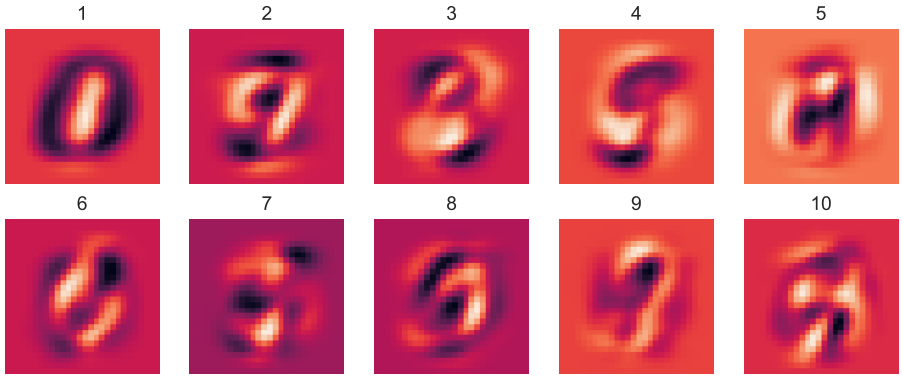
We are essentially transforming $x - \mu$ into the k dimensional eigenvector space, transforming it back, then correcting the mean subtraction by adding in μ.

c. *[5 points]*





d. *[3 points]* The eigenvectors are shown below. Some of them look like numbers (the first eigenvector looks like a 0, the third an 8, etc.) but most of them seem to be focusing on edges and the differences in brightness, rather than specific lines, to capture the characteristics of each digit.



e. *[3 points]* The digits are already fairly distinguishable at $k = 15$. By $k = 40$ they are distinct and $k = 100$ simply makes them clearer. Thus we need very few dimensions (less than 10%) to accurately portray all digits.