CS181-S19

Assignment #3

Due: 11:59pm March 29, 2019

Homework 3: Max-Margin, Ethics, Clustering

Introduction

This homework assignment will have you work with max-margin methods and clustering, as well as an ethics assignment. The aim of the assignment is (1) to further develop your geometrical intuition behind margin-based classification and decision boundaries, (2) try coding a simple K-means classifier, and (3) to have you reflect on the ethics lecture and to address the scenario discussed in class in more depth by considering the labor market dynamically.

We encourage you to first read the Bishop textbook coverage of these topics, particularly: Section 7.1 (Max-Margin and SVMs) and Section 9.1 (Clustering). Chapters 5 and 6 of the student textbook are also relevant.

There is a mathematical component and a programming component to this homework. Please submit your PDF, tex, and Python files to Canvas, and push all of your work to your GitHub repository. If a question requires you to make any plots, like Problem 2, please include those in the writeup.

Problem 1 (Fitting an SVM by hand, 7pts)

For this problem you will solve an SVM without the help of a computer, relying instead on principled rules and properties of these classifiers.

Consider a dataset with the following 7 data points each with $x \in \mathbb{R}$:

$$\{(x_i, y_i)\}_i = \{(-3, +1), (-2, +1), (-1, -1), (0, +1), (1, -1), (2, +1), (3, +1)\}$$

Consider mapping these points to 2 dimensions using the feature vector $\boldsymbol{\phi}(x) = (x, -\frac{8}{3}x^2 + \frac{2}{3}x^4)$. The hard margin classifier training problem is:

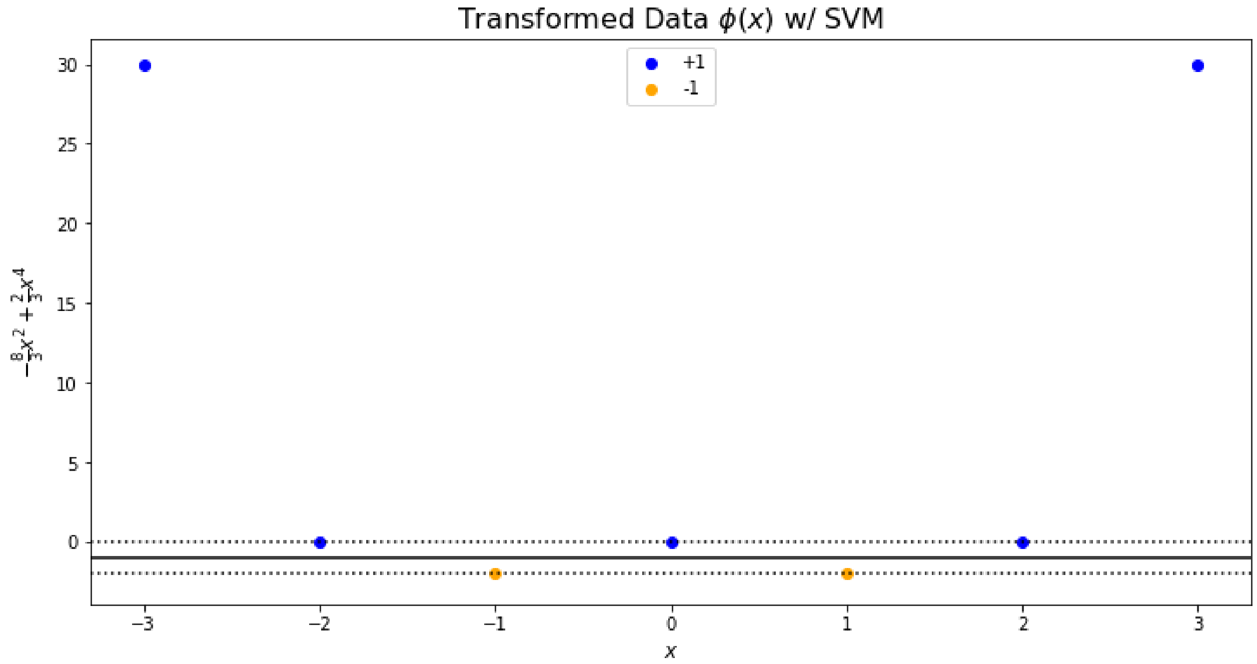
$$\begin{aligned} &\min_{\mathbf{w}, w_0} \|\mathbf{w}\|_2^2 \\ \text{s.t.} \quad & y_i(\mathbf{w}^\top \boldsymbol{\phi}(x_i) + w_0) \geq 1, \ \forall i \in \{1, \dots, n\} \end{aligned}$$

The exercise has been broken down into a series of questions, each providing a part of the solution. Make sure to follow the logical structure of the exercise when composing your answer and to justify each step.

1. Plot the transformed training data in $\mathbb{R}^2$ and draw the decision boundary of the max margin classifer.
2. What is the value of the margin achieved by the optimal decision boundary?
3. What is a vector that is orthogonal to the decision boundary?
4. Considering discriminant $h(\boldsymbol{\phi}(x); \mathbf{w}, w_0) = \mathbf{w}^\top \boldsymbol{\phi}(x) + w_0$, give an expression for *all possible* $(\mathbf{w}, w_0)$ that define the optimal decision boundary. Justify your answer.
5. Consider now the training problem. Using your answers so far, what particular solution to $\mathbf{w}$ will be optimal for this optimization problem?
6. Now solve for the corresponding value of w_0, using your general expression from part (4.) for the optimal decision boundary. Write down the discriminant function $h(\boldsymbol{\phi}(x); \mathbf{w}, w_0)$.
7. What are the support vectors of the classifier? Confirm that the solution in part (6.) makes the constraints above binding for support vectors.

Solution

1.



2.

The value of the max margin achieved is 1; this is the perpendicular distance between the support vector and the decision boundary.

3.

Any vector which has a vertical component in the second dimension of the new basis and a zero component in the first dimension of the new basis is orthogonal to the decision boundary. Thus, a vector $\mathbf{w} = \alpha[0, 1] \mid \alpha \in \mathbf{R}$ will be orthogonal to the decision boundary.

4.

The decision boundary is defined by $\mathbf{w}^T\phi(x)+w_0 = 0$, but the parameters can be rescaled so that $\alpha(\mathbf{w}^T\phi(x)+w_0) = 0$, or $\alpha\mathbf{w}^T\phi(x) + \alpha w_0 = 0$. We found from above that a possible value for the vector orthogonal to the decision boundary is $[0, 1]^T$ which has a corresponding bias of 1. Thus, we have that the set of possible $(\mathbf{w}, w_0)$ that define the decision boundary as:

$$\alpha[0, 1]^T\phi(x) + \alpha \cdot 1 = 0 \quad \forall\alpha \in \mathbf{R}$$

5. and 6.

Now, if we set $\alpha = 1$, we have $[0, 1]^T \phi(x) + 1 = 0$, and if we find

$$y_i[0, 1]^T \phi(x_i) + 1$$

for all i, we get the following output vector:

$$[31., 1., 1., 1., 1., 1., 31.]$$

and we see that the constraint in the optimization problem, that

$$y_i(\mathbf{w}^\top \boldsymbol{\phi}(x_i) + w_0) \geq 1,\ \forall i \in \{1, \ldots, n\}$$

is met with $\alpha = 1$.

Thus, $\mathbf{w} = [0, 1]$ and $w_0 = 1$, giving us $h(\boldsymbol{\phi}(x); \mathbf{w}, w_0) = [0, 1]^T \phi(x) + 1$.

7.

From the plot, we see that the support vectors are the second, third, fourth, and fifth points. This matches the output vector we obtain from above when applying the discriminant function as the second, third, fourht, and fifth elements of the output vector are 1 i.e. the constraint is binding.

Problem 2 (K-Means, 10pts)

For this problem you will implement K-Means clustering from scratch. Using `numpy` is fine, but don't use a third-party machine learning implementation like `scikit-learn`. You will then apply this approach to clustering of image data.

We have provided you with the MNIST dataset, a collection of handwritten digits used as a benchmark of image recogntion (you can learn more about the data set at `http://yann.lecun.com/exdb/mnist/`). The MNIST task is widely used in supervised learning, and modern algorithms with neural networks do very well on this task.

Here we will apply unsupervised learning to MNIST. You have been given representations of 6000 MNIST images, each of which are 28×28 greyscale handwritten digits. Your job is to implement K-means clustering on MNIST, and to test whether this relatively simple algorithm can cluster similar-looking images together.

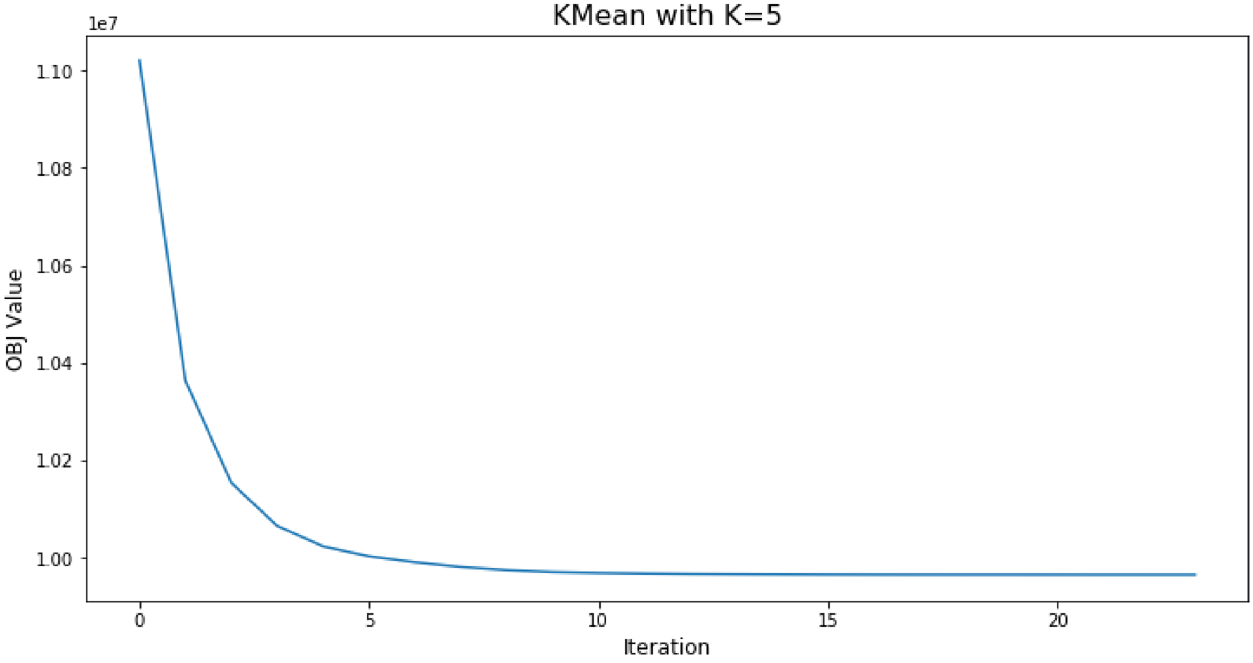
The given code loads the images into your environment as a 6000x28x28 array. In your code, you may use the ℓ_2 norm as your distance metric. (You should feel free to explore other metrics than the ℓ_2 norm, but this is strictly optional.)

- Starting at a random initialization and some choice of K, plot the K-means objective function as a function of iteration and verify that it never increases.
- Run the K-means algorithm from several different restarts for different values of K. Plot the final K-means objective as a function of K with errorbars over the random restarts. How does the objective and the variance of the objective change with K?
- For $K = 10$, for a couple of random restarts, show the mean images for each cluster. To render an image, use the pyplot `imshow` function.
- Now, before running K-means, standardize the data. That is, center the data first so that each pixel has mean 0 and variance 1 (except for any pixels that have zero variance). For $K = 10$, for a couple of random restarts, show the mean images for each cluster. Compare them to the previous part.

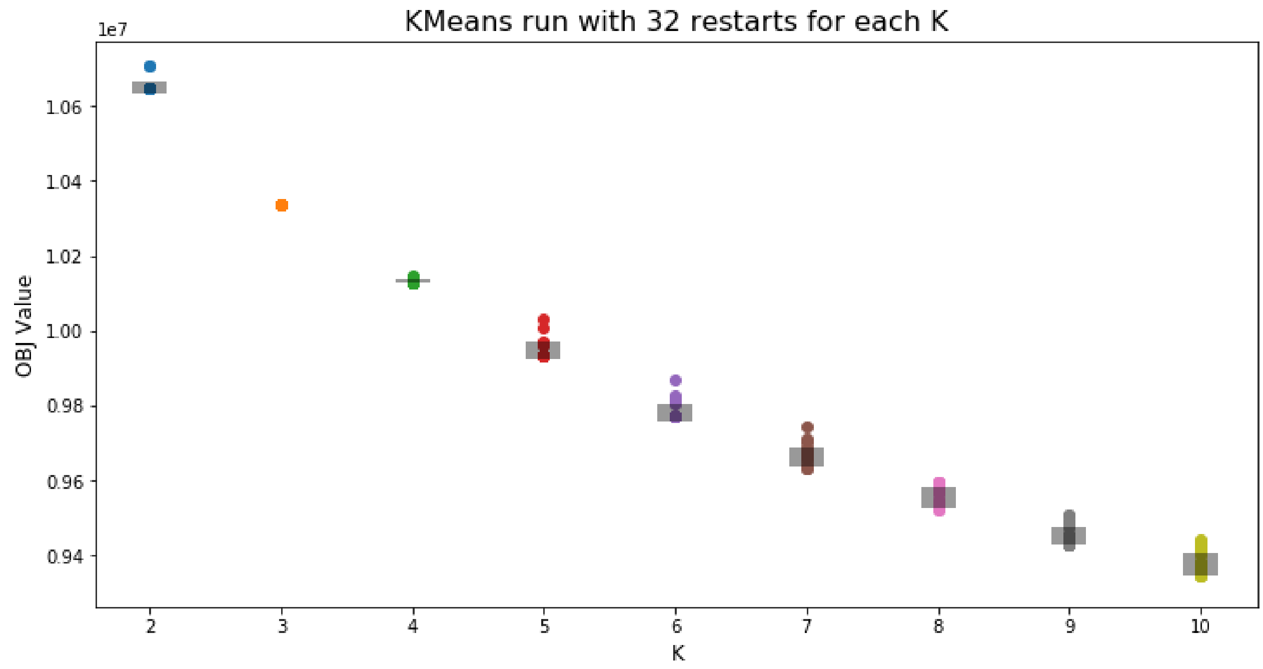
As in past problem sets, please include your plots in this document. (There may be several plots for this problem, so feel free to take up multiple pages.)

Solution

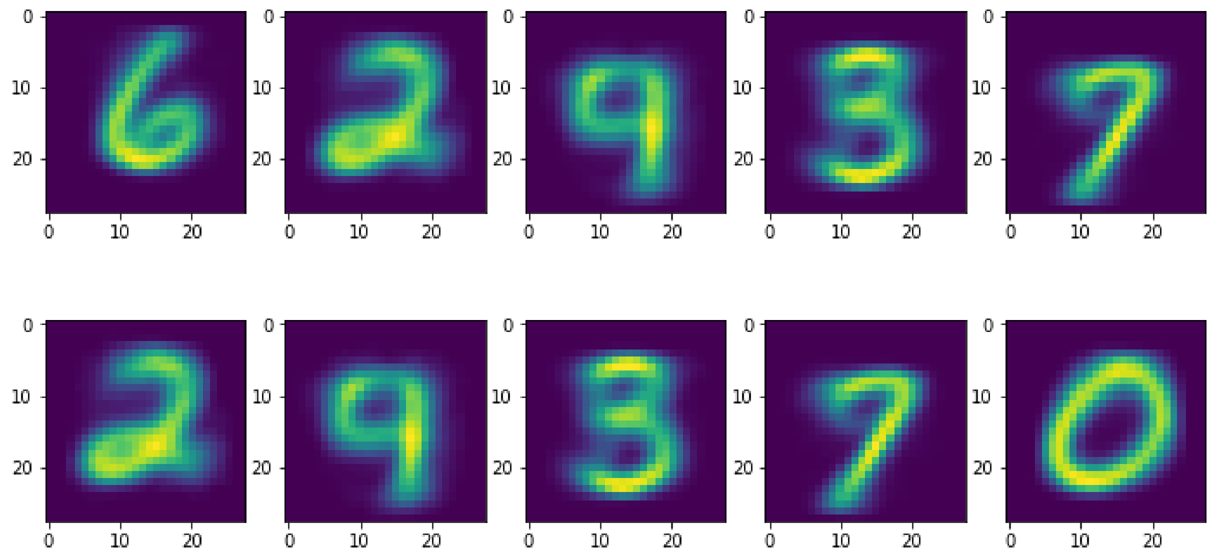
We see from the below plot that the K-means objective function never increases with the number of iterations.



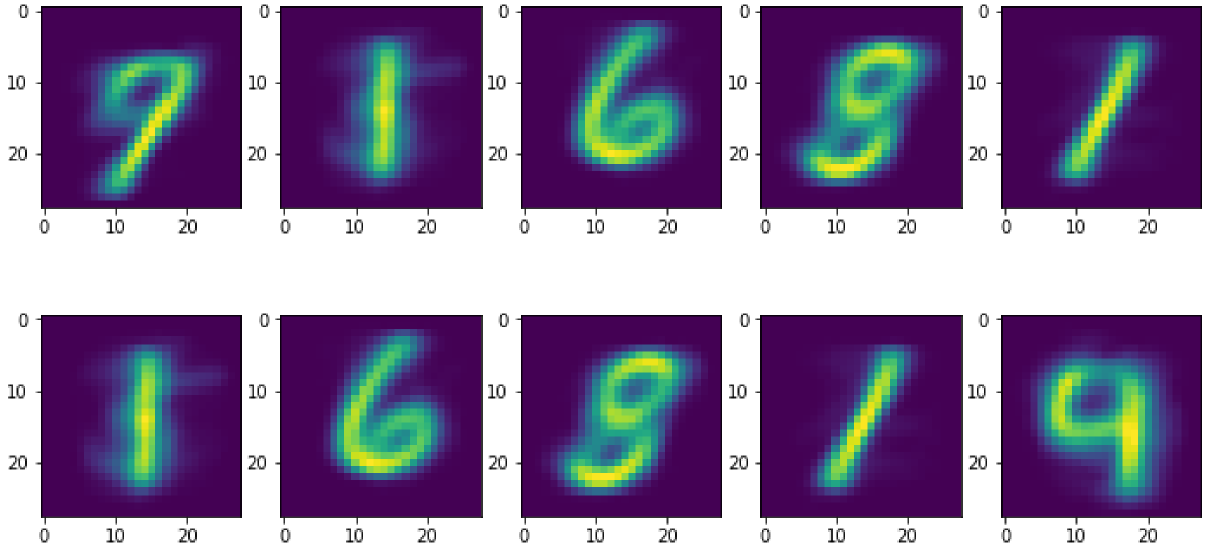
We see from the below plot that the objective clearly decreases as the number of clusters increases. From the plot too, we see that the standard deviation (and variance) does not have a concise pattern and appears to be the same across different number of clusters (except for discrepancies for small values of K which are likely the result of a small sample size).


KMeans run with 32 restarts for each K

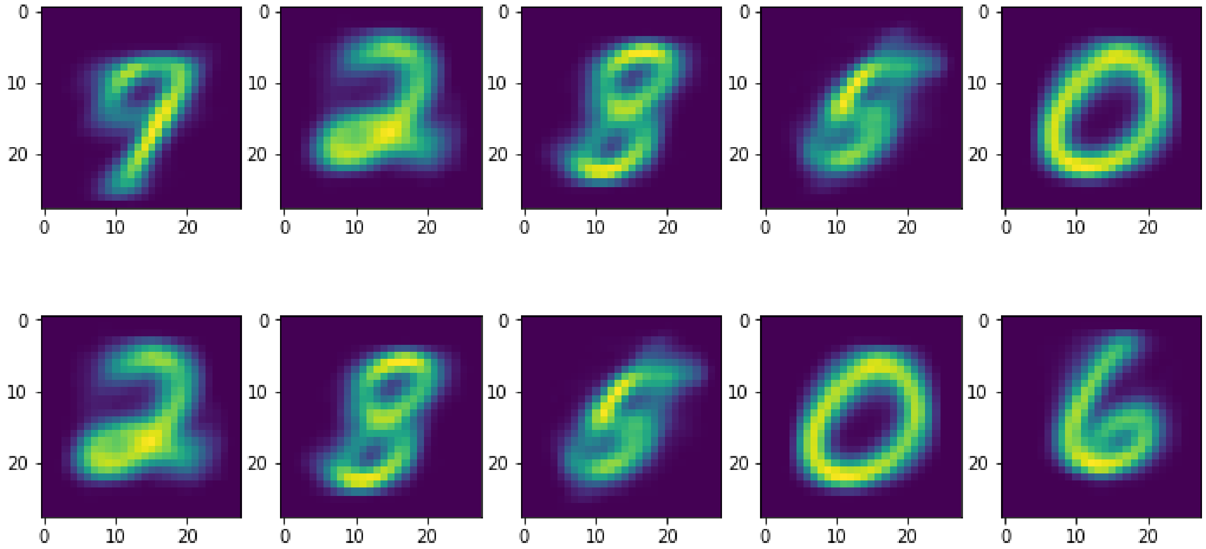
Below we plot for K = 10, for a couple of random restarts, the mean images for each cluster.

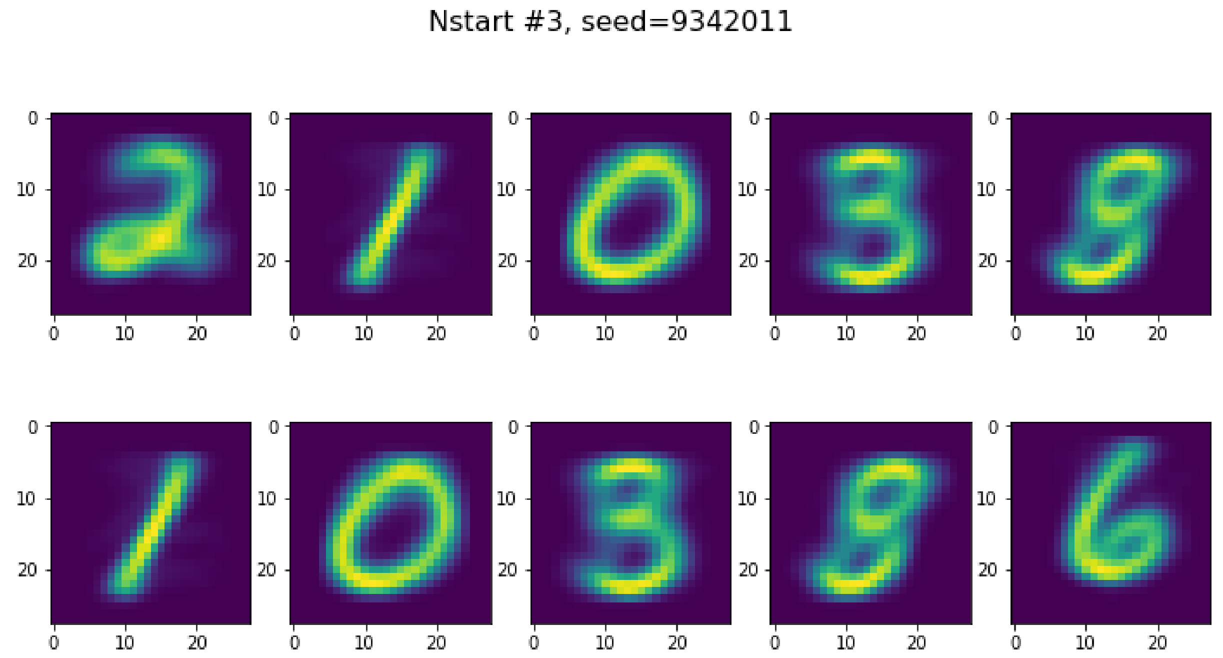

Nstart #0, seed=7890469

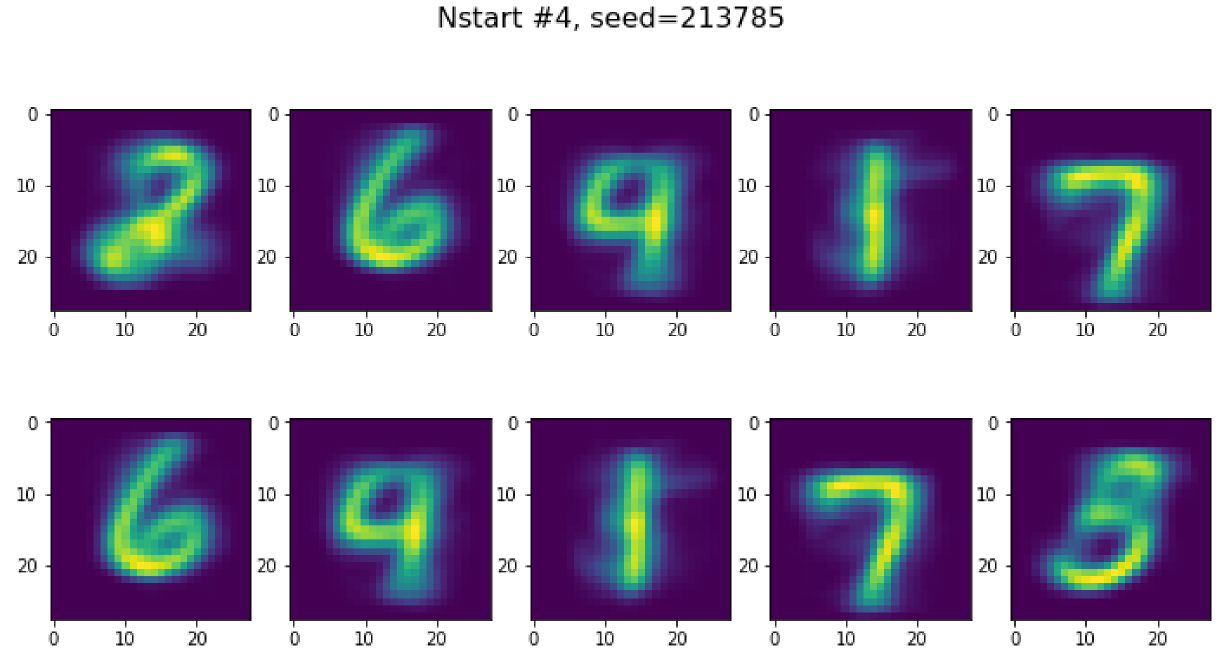
Nstart #1, seed=6117910



Nstart #2, seed=5830461

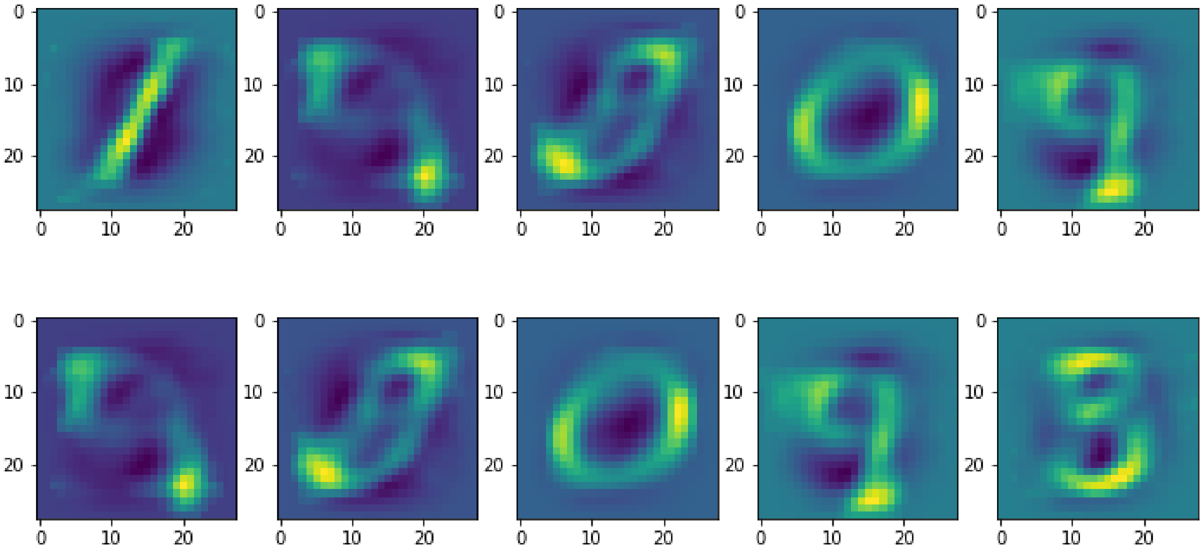




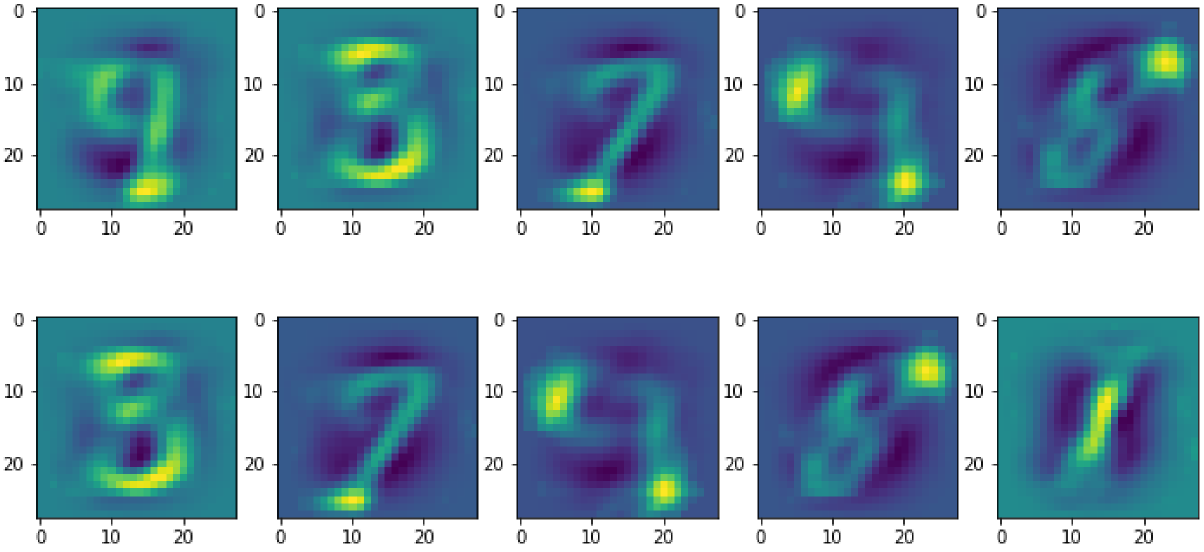


After standardizing the data such that the mean and standard deviation of each pixel across the data are 0 and 1, we plot for K = 10, for a couple of random restarts, the mean images for each cluster.

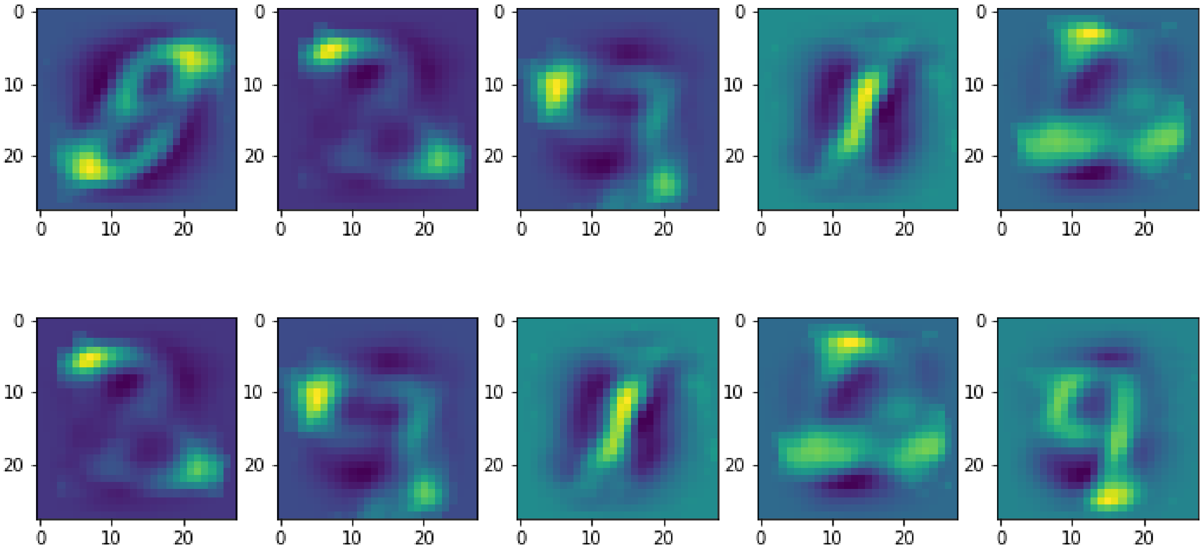
Nstart #0, seed=7528212



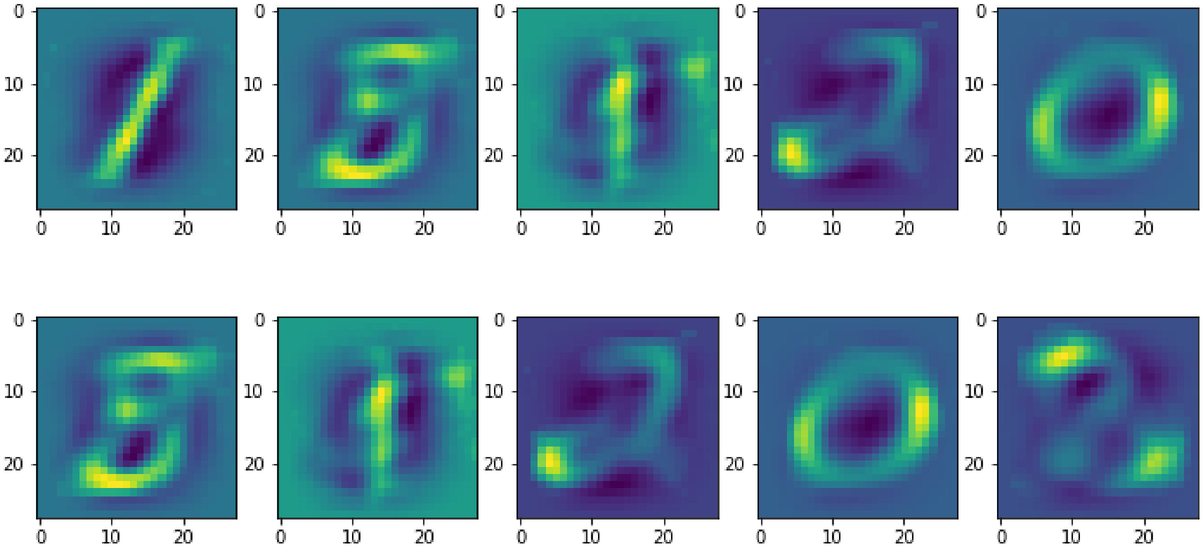
Nstart #1, seed=2453989

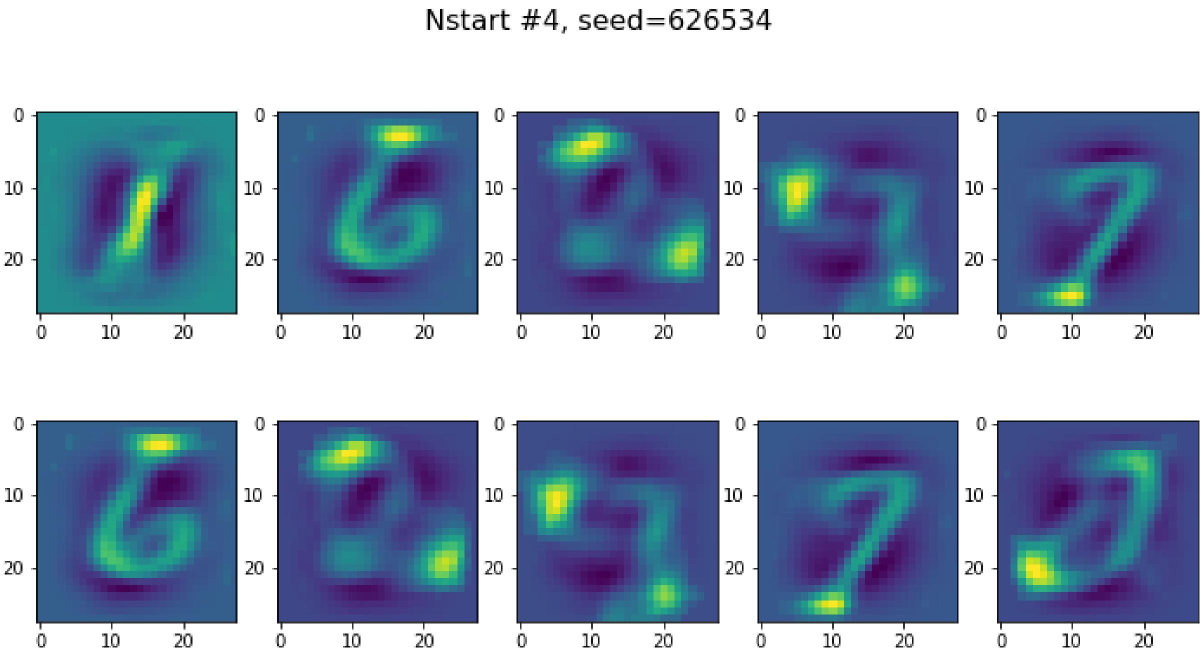


Nstart #2, seed=1084313

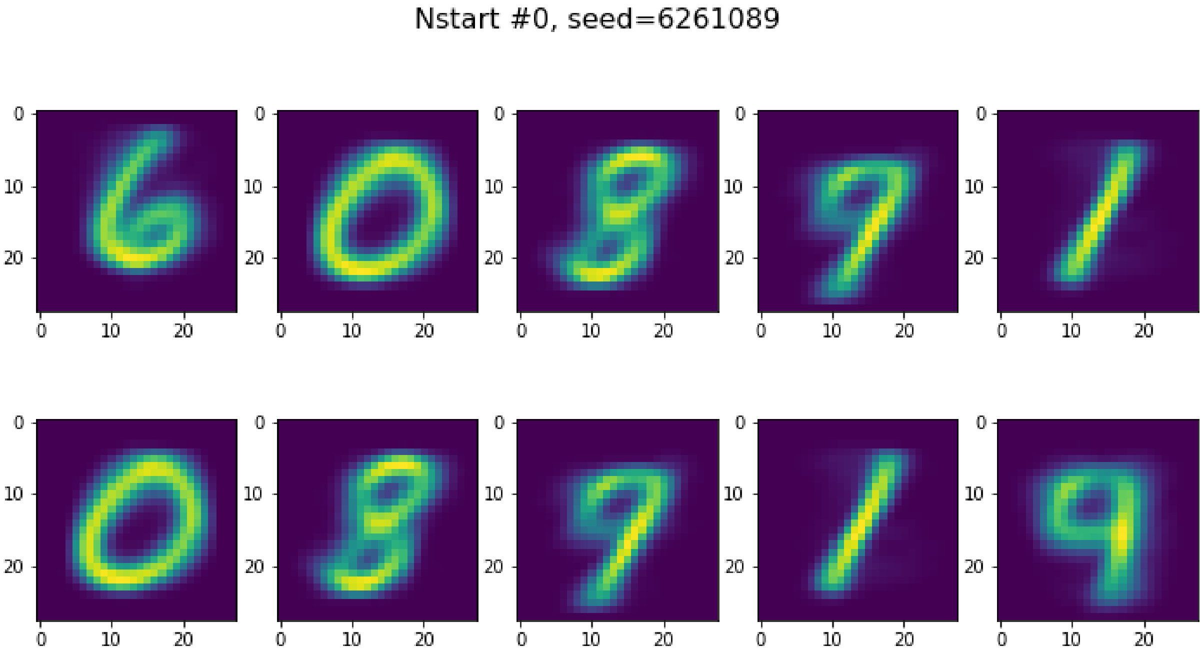


Nstart #3, seed=470556

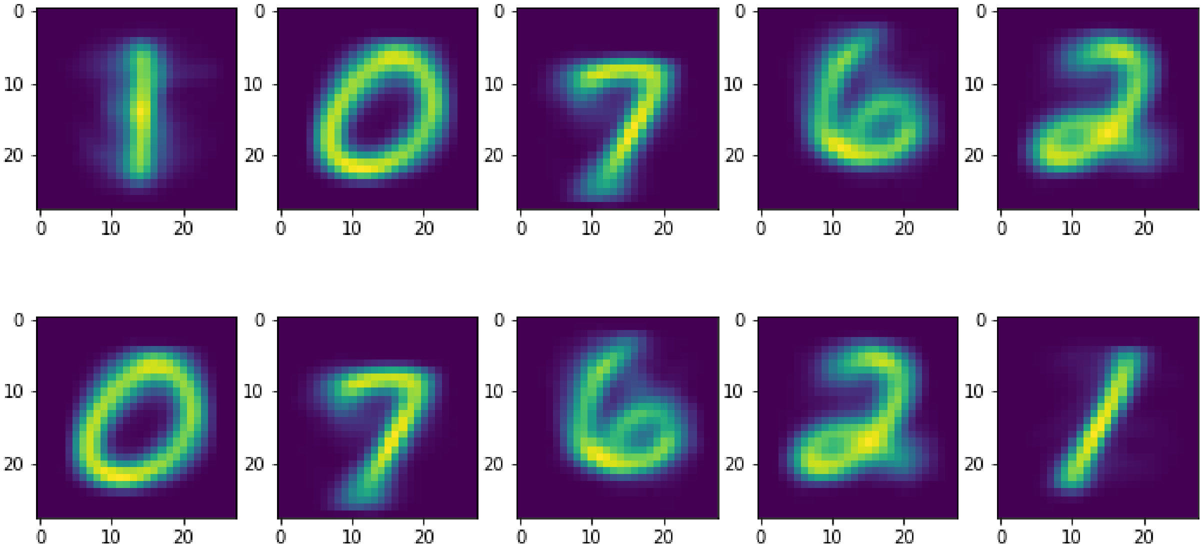



Nstart #4, seed=626534

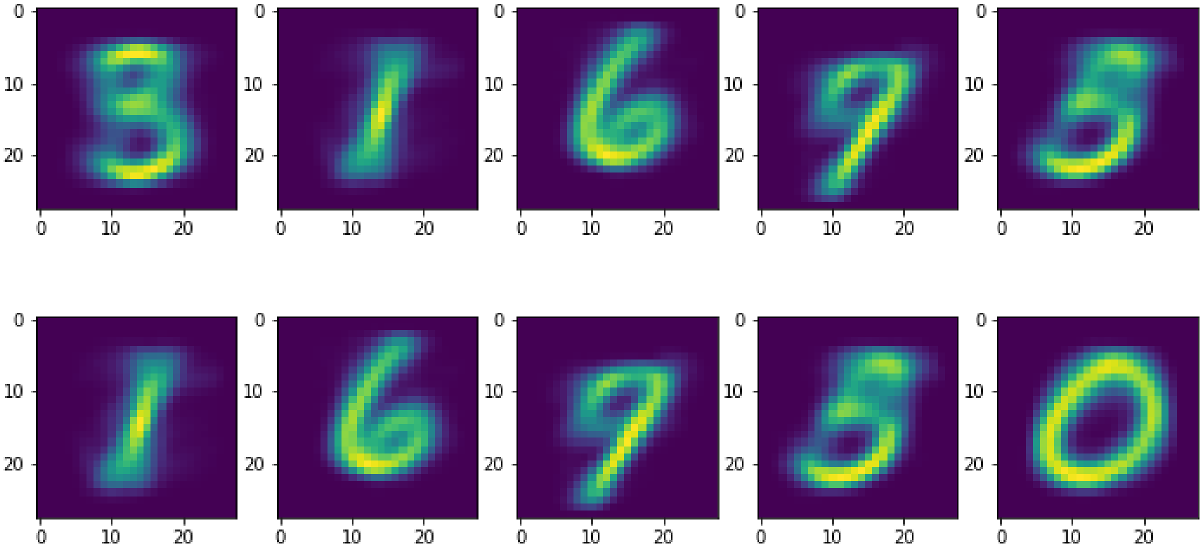
After re-standardizing the data with min-max scaling such that the values of the pixels range from 0 to 255, we plot for K = 10, for a couple of random restarts, the mean images for each cluster.

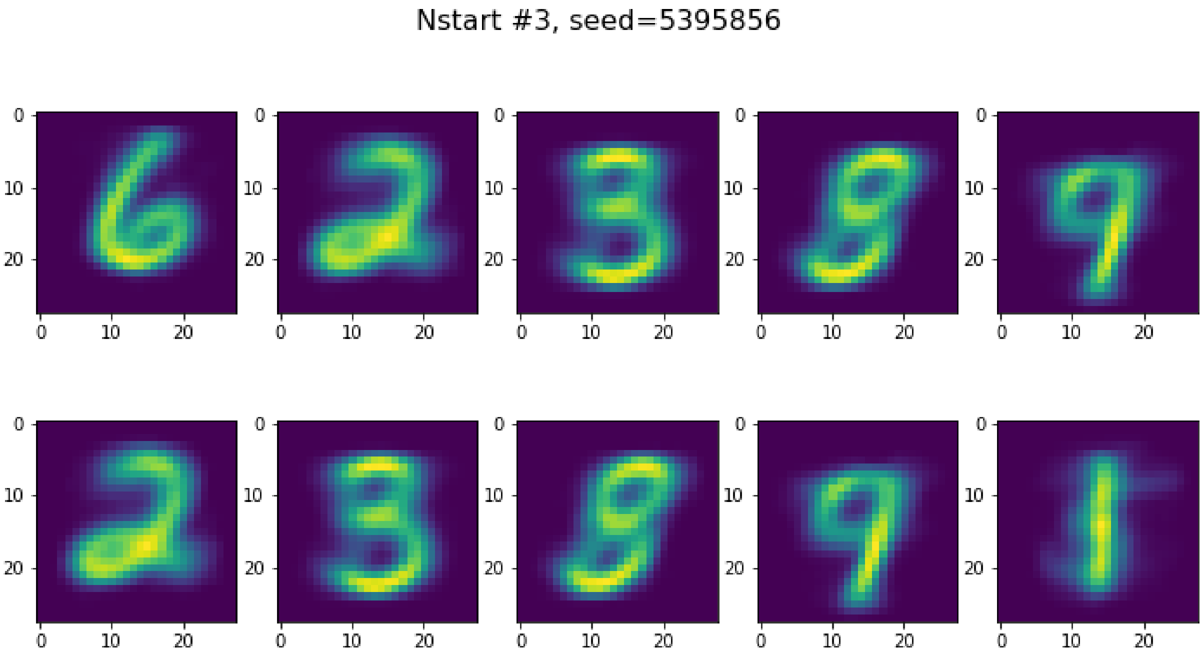

Nstart #0, seed=6261089

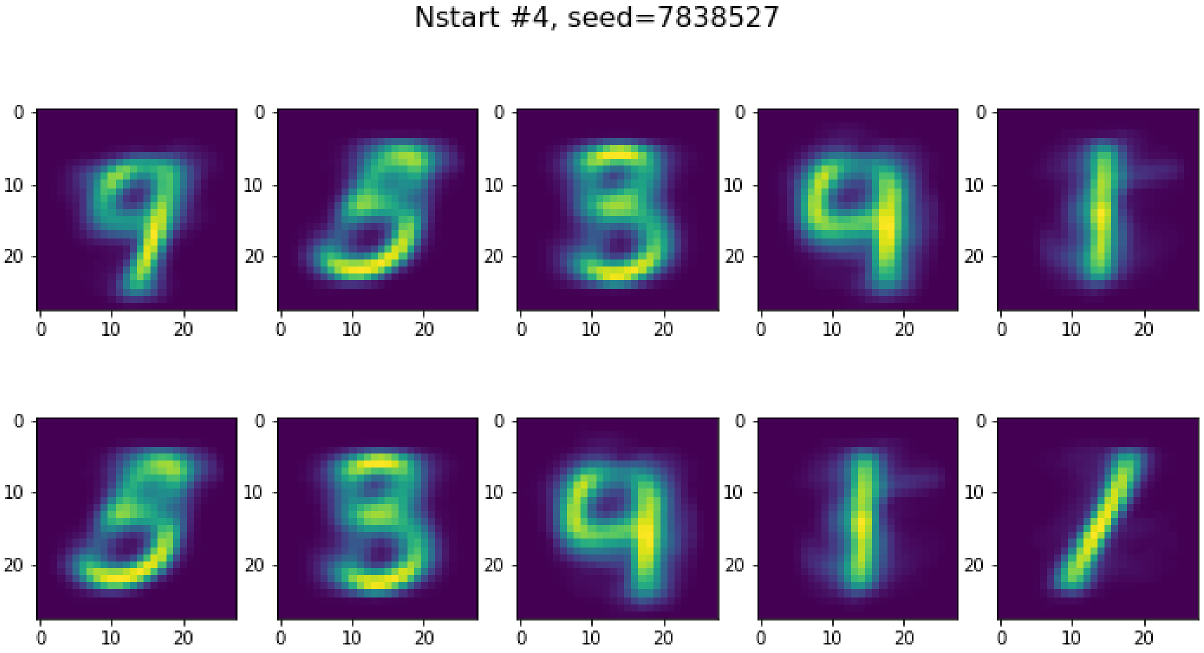
Nstart #1, seed=4796450



Nstart #2, seed=3732241




Nstart #3, seed=5395856


Nstart #4, seed=7838527

We see from the above that standardization does not necessarily improve the KMeans algorithm in this example; in fact we see that normalization - standardizing with mean 0 and standard deviation one - produces some questionable plots of mean images for the centers. Intuitively, this makes sense as it does not make sense to normalize pixel data as these values have intrinsic and underlying meaning that defines the image that should have keep its variation across the image. Scaling down the data via a min-max scaler makes more intuitive sense, in particular if we find that it helps improve convergence of our algorithm. Noticeably, across all three sets of plots, we see that the KMenas on various iterations of setting 10 clusters fails to perfectly achieve all 10 digits. We see that when the algorithm recognizes the digit '1' that it may not recognize the digits '4' - i.e. it fails to distinguish the '1' from the '4' by only picking up on the vertical line segment in the '4' and failing to pick up the left hand side of the digit. Further, we see that when the algorithm picks up what we might call as 'a straight line segment' that it conflates the '1' and '7' producing two mean images that are slanted vertical line segments characteristic of the '7' digit. The failures or areas of improvement for the current algorithm represent the limit of the KMeans algorithm using the L2 norm and the high dimensionality of the problem - a 28x28 image has essentially a dimentionality of 784 when we flatten the variables. It is because of these latter problem that digits with a high degree of similarity - '1' and '7', '5' and '8' - are conflated as the distance metric out-weights the similarities from the subtle differences - differences which are not subtle to us but which are subtle in the math,

Problem 3 (Ethics Assignment, 10pts)

Our class activity:

> Hiring at Forever 28. Forever 28 has hired a new computer science team to design an algorithm to predict the success of various job applicants to sales positions at Forever 28. As you go through the data and design the algorithm, you notice that African-American sales representatives have significantly fewer average sales than white sales representatives, and the most profit comes from interactions between white female sales representatives and white male customers. The algorithm's output recommends hiring far fewer African-Americans than white applicants, even after the percentage of applications from people of various races are adjusted for, and having those sales representatives target white male customers.

In class, we assumed that the problem was *static*: given historical data, such as data about sales performance, who should Forever 28 hire right now? In this follow-up assignment, think about consumer behavior and firm hiring practice dynamically. Looking at features of the labor market dynamically allows you different degrees of freedom in your model. For example, in class, you probably took consumer preference about the race of their sales representative as given. What would happen if you assumed that consumer preference could vary over time (say, on the basis of changing racial demographics in the sales force)?

Your new case:

> The US Secretary of Labor has heard about your team's success with Forever 28 and comes to you with a request. The Department of Labor wants to reduce disparate impact discrimination in hiring. They want you to come up with a model of fair hiring practices in the labor market that will reduce disparate impact while also producing good outcomes for companies.

Write two to three paragraphs that address the following:

- What is disparate impact, and how does it differ from disparate treatment?
- What are the relevant outcomes, for both workers and companies? Are these outcomes measurable?
- What are some properties of your algorithm that might produce those socially good results? Think about constraints that you might build in, such as the fairness constraints that we discussed in class, or how you might specify the prediction task that we are asking the machine to optimize. [Optional] What trade-offs might your algorithm have to balance?
- Recommend a deployment strategy. Are there any features of the data collection, algorithm implementation, or broader context that make you wary? Describe how your algorithm might fit into the broader context of a company (e.g. hiring, training, marketing, sales).

We expect clear, concise, and thoughtful engagement with this question, which includes providing your reasoning for your answers. In your response, depth is more important than breadth. For example, in question 1, you could simply choose profit for the outcome that companies may be interested in; you can run with a single fairness criterion. We do *not* expect you to do any outside research, though we encourage you to connect to lecture materials where relevant.

Solution

Disparate Treatment versus Disparate Impact:
Disparate treatment describes the more "obvious" instances of discrimination; it describes instances of discrimination that use protected variables, for example. Disparate impact describes the less "obvious" forms

of discrimination that are facially neutral but nevertheless are unintentionally discriminating and avoidable; these forms of discrimination may use unbiased data but are nevertheless discriminating.

Relevant outcomes for both workers and companies

For workers, the relevant outcomes are hired versus not-hired on the basis of their application, ideally not on the basis of a protected variable like race. For the company, the relevant outcome is maximizing profit. Both of these are measurable.

Properties of the algorithm

The algorithm should match, for example, the best sales representatives to the company (or those applicants with the potential to be the best sales representatives), and it should do this in a fashion which is independent of race by optimizing for fairness and being well calibrated. The fairness criteria means that the probability of an applicant being hired given a particular race and a set of other non-protected predictors is equal to the probability of an applicant being hired given any other race and the same set of non-protected predictors. Essentially, and ideally, the applicant should be hired solely on their ability to be a great sales representative. This will optimize both the social goods of ethical applicant evaluation and profit maximization.

In the Forever 28 case, we are told that "African-American sales representatives have significantly fewer average sales than white sales representatives, and the most profit comes from interactions between white female sales representatives and white male customers." On this measure, it may seem that race is an important predictor for maximizing the company's profit such that we cannot both maximize profit and achieve our fairness criterion. However, this simple statement may have several underlying complexities and thus should be reconsidered in that light through data exploration. The latter observation may itself be a reinforcing bias for several reasons. Data exploration should be a must to give proper context to analyses, as we show briefly below.

It may very well be the case that the reason behind why we arrive at the above observation - that African American sales representatives see less profit for the company than their white counterparts - is because of unequal training, for example. The analysis which yields the discrepancy between sails of African Americans and whites may be the result of the underlying discrepancy in training; the African American and white sales representatives may both possess the same underlying attributes of sales performance but this latent ability may not be brought out in African American sales representatives. This difference in training may be the result of human bias that the data analysis and subsequent algorithmic training cannot pick up on. Further, the company may be a chain such that results of the sales are aggregated across various stores. Thus, we should consider and factor in non-protected demographics; perhaps those stores in predominately African American communities, where income is on average lower than predominantly white communities, see less profit as a result of this demographic. Thus, the analysis given to us may not be the result of an apparent inadequacy of African American sales representatives compared to whites - both the overall observation and the race-to-race interaction - but instead be a result of the this underlying demographic and the history that comes with it. Finally, in light of current population demographic trends, where the population is becoming more diverse and heterogeneous and the population's attitudes are changing as a result, it may very well be ideal to hire a diverse sales force which mirrors the population, a result we can achieve by evaluating applicant on underlying potential and not race.

When the potential algorithm is trained on data including the protected race variable (i.e. without the data exploration mentioned above), it may not only be that the results are biased but it may also be the case that the results are inaccurate in the sense that the "best" applicants - those with a higher "latent" ability - are actually given a lower evaluation than they should have received because of the race indicator - as may be what is going on in the Forever 28 example. As a result, the company does not achieve as high a profit. Thus, it could be the case that when we account for the potential biases in the data - like we noted above - that the algorithm performs better, both in terms of identifying the "best" applicants that will maximize profit and in terms of being non-discriminating. To this extent, we would not have to make comprises for our algorithm as we maximize both social goods simultaneously: the company maximizes profit and the fairness criterion is met. The key is to explaining and making sense of our data well in addition to having good, non-protected variables.

Development Strategy

To produce a training dataset which aligns with the above, a number of measures need to be carried out. In the case of Forever 28, we need to consider performance of sales representatives not on race, but on the context of other factors such as level of training and store demographics. Obtaining demographics of the regions surrounding stores - median income of the surrounding area for example - should not be difficult but quantifying the level of training obtained by the sales representatives may be difficult. Training may not be standardized across stores and obtaining non-biased measures may be difficult. Thus, to ameliorate this process in the future, it should be a priority for the company to create or review its training procedures to ensure training is uniform across stores and across representatives. Further feature engineering may be appropriate as well to develop this dataset; perhaps we could look into quantifying variables which we might believe make a great sales representative for the company e.g. perhaps the more outgoing, personable, friendly applicants make better representatives or those applicants knowledgeable about “fashion trends” or what “goes together better” in the case of FOrever 28 - if such variables can be quantified, perhaps by some test. Essentially, we would like to produce a dataset which account for quality of employee outside of protected variables while also possessing concrete, dense data on non-protected potential performance indicators and measures. It may take some time to standardize and develop this data, but carrying out this task will allow companies to fully maximize profit down the road by identifying the best applicants irrespective of race which should in turn produce a diverse sales force.

- Name: Meriton Ibrahimi
- Email: meritonibrahimi@college.harvard.edu
- Collaborators: N/A
- Approximately how long did this homework take you to complete (in hours): a couple hours each day over the week