Free-response questions (5 points)
Understanding SVMs (1 point)

1. (0.5 points) Explain why a support vector machine using a kernel, once trained, does not directly use the decision boundary to classify points.

**A :**

**Reason to use a kernel: data might be linearly inseparable. Therefore, we need to transform the data into something that can be separated by a hyperplane, and this requires transform $\phi(\dot{})$. After that, we can optimize on $L_d$ and calculate the class label with $\phi(x)\phi(z)$ directly, which does not require explicit knowledge of $\phi(\dot{})$.**

**Reason not to directly use a decision boundary:**
**If we don’t calculate $\phi(x)\phi(z)$ directly, we need to calculate $\phi(dot)$ separately. Some Kernels, like Gaussian kernel which includes an exponential term, will implicitly expand to an infinite dimensional feature space through taylor expansion. Therefore, using kernel instead of feature vectors for classifying points is more feasible.**

2. (0.5 points) If the support vector machine does not directly use the decision boundary to classify points, how does it, in fact, classify points. Hint, what are the support vectors?

**A: We are trying to optimize on $L_d$ with respect to $$\phi(\dot)$$, and $$X_s$$ with positive $$\alpha_s$$ will be support vectors. Then, instead of calculating $$\phi(X_s)$$ and the new point’s transform $$\phi(z)$$ separately, which might be expanded into infinite dimensional space, we can calculate their product $$\phi(x)\phi(z)$$ directly through a pre-identified operation $$K(x,z)$$. Then, classify the new point as**

$$g(\mathbf{z}) = \sum_{s}^{S} \alpha_s y_s K(\mathbf{x}_s, \mathbf{z}) + b$$

the MNIST data (1 point)

3. (0.5 points) How many images are there in the MNIST data? How many images are there of each digit? How many different people's handwriting? Are the digit images all the same size and orientation? What is the color palette of MNIST (grayscale, black & white, RGB)?

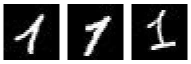
**A:**

```
digit:  0 number of images:  5923
digit:  1 number of images:  6742
digit:  2 number of images:  5958
digit:  3 number of images:  6131
digit:  4 number of images:  5842
digit:  5 number of images:  5421
digit:  6 number of images:  5918
digit:  7 number of images:  6265
digit:  8 number of images:  5851
digit:  9 number of images:  5949
Total Size of Database:  70000
```

**Number of Writers: Approx. 500 writers**
**Size: fixed size.**
**Orientation: Not the same**
**Color Palette: black & white**

4. (0.5 points) Select one of the digits from the MNIST data. Look through the variants of this digit that different people produced. Show us 3 examples of that digit you think might be challenging for a classifier to correctly classify. Explain why you think they might be challenging.

**A:**
**There are 1s that look like 7, also there are ones that look very different from any major pattern. So irregular patterns of the same number is what makes classification challenging - those patterns might be misclassified.**



Selecting training and testing data (.5 points)

5. (0.5 points) Now you have to decide how to make a draw from the data for training and testing a model. Think about the goals of training and testing sets - we pick good training sets so our classifier generalizes to unseen data and we pick good testing sets to see whether our classifier generalizes. Explain how you should select training and testing sets. (Entirely randomly? Train on digits 0-4, test on 5-9? Train on one group of hand-writers, test on another?). Justify your method for selecting the training and testing sets in terms of these goals.

**A: Yann LeCun stated on his website that “Drawing sensible conclusions from learning experiments requires that the result be independent of the choice of training set and test among the complete set of samples.”  Therefore, I randomly select a group of 250**

**hand-writers' handwriting for all 10 digits as my training set. The other group of 250 hand-writers' handwriting will be my testing set. Then, I will keep repeating this training testing step and find the training set that gives the lowest error rate. This way, we can "simulate" the result we have when we have a new handwriter, while making sure there is no loss of independence.**

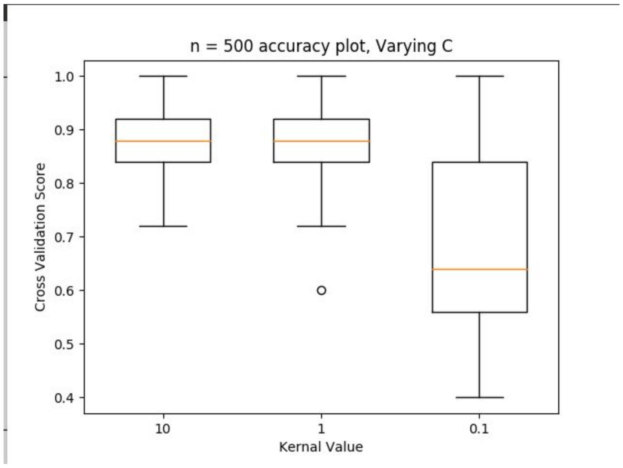
Finding the best hyperparameters (2.5 points)

To answer the following questions you should use your **GridSearch** hyperparameter tuner. We want to find the **best kernel and slack** cost, C, for handwritten digit recognition on MNIST using a support vector machine. To do this, we're going to try **different kernels** from the set {Linear, Polynomial, Radial Basis Function}. Use the **default value of 3 for the degree** of the polynomial. We will combine each kernel **with a variety of C** values drawn from the set { 0.1, 1, 10 }. This results in 9 variants of the SVM. For each variant we will be running 20 fold cross validation which was specified in the worker.py. You can simply call the run function within experiment.py with the right parameters to get all the results that you nedd.

9. (0.5 point) Create a table with 3 rows (1 kernel per row) and 3 columns (the 3 slack settings). Rows and columns should be clearly labeled. For each condition (combination of slack and kernel), show the following 3 values: the mean accuracy a of the trials, the standard deviation of the accuracy std and the number of trials/experiments n, written in the format: mean(a),std(a),n.

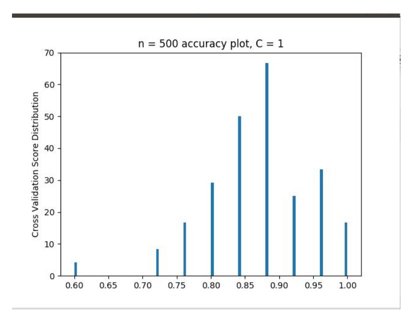
| | C = 1 | C= 10 | C = 0.1 |
|---|---|---|---|
| 'rbf' | 0.89, 0.057, 20 | 0.91, 0.053, 20 | 0.56, 0.071, 20 |
| poly | 0.83, 0.09, 20 | 0.86, 0.076, 20 | 0.6, 0.09, 20 |
| linear | 0.88, 0.064, 20 | 0.88, 0.064, 20 | 0.88, 0.063, 20 |

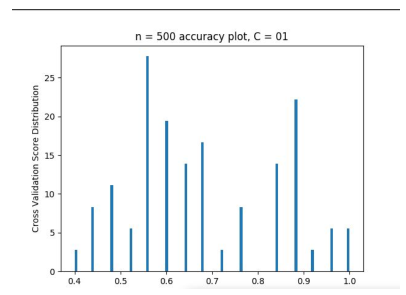
10. (0.5 points) Make a **boxplot graph** that plots accuracy (vertical) as a function of the **slack C** . There should be 3 boxplots in the graph, **one per value of C**. Use results across all kernels. Indicate n on your plot, where **n** is the number of trials per boxplot. Don't **forget to label your dimensions.**

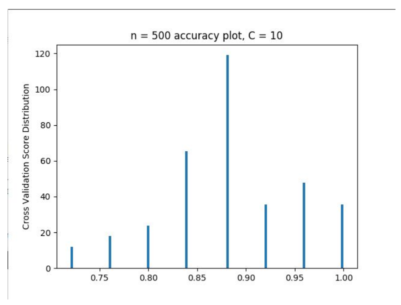


**CORRECTION: n = 60**

11. (0.25 points) What statistical test should you use to do comparisons between the values of C plotted in the previous question? Explain the reason for your choice. Consider how you selected testing and training sets and **the skew of the data** in the boxplots in your answer. Note: Your boxplots will show you whether a distribution is skewed (and thus, not normal), but will not show you what the shape of each distribution. There are distributions that are not skewed, but are still not bell curves (normal distributions). It would be a good idea to look at the histograms of your distributions to decide which statistical test you should use.







**CORRECTION: n = 60**

**A:**

**Between C = 1 and C = 10, I want to use Mann-Whitney signed rank test, as there is one outlier in the C=1 distribution.**

**For the rest, I want to use Mann-Whitney test as well because they are non-gaussian distributions, also because they are independent.**

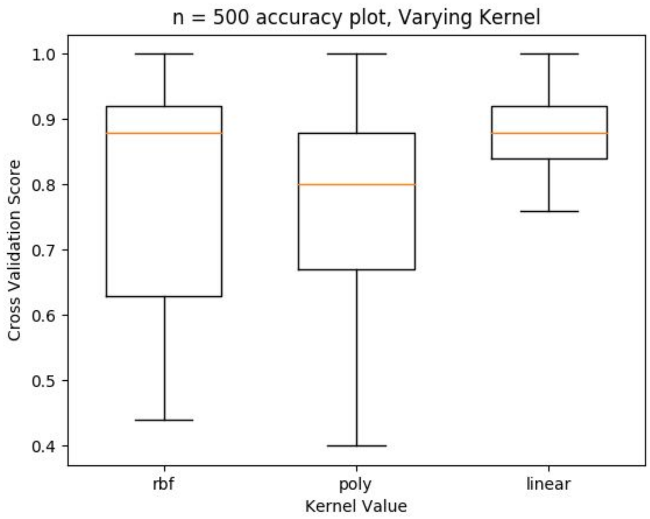
12. (0.25 points) Give the p value reported by your test. Say what that p value means with respect to the 5% rule.

**A:**

**C = 10 vs C= 1: 0.1488610119454875**
**C = 10 vs C = 0.1: 5.5381293962832595e-11**
**C = 1 vs C = 0.1: 1.1445293125993956e-09**

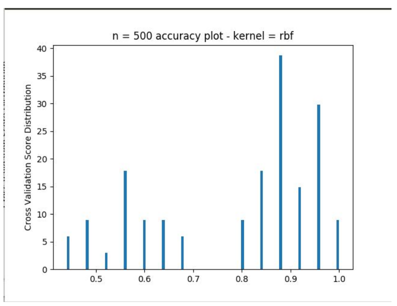
**Since the second and third p values are less than 5%, the distributions are different, and there is a statistical significance in their differences. However, the first comparison does not yield a statistically significant comparison, with a p value above 5%. Therefore, varying on C does not necessarily have an effect on the performance of SVM's true error rate.**

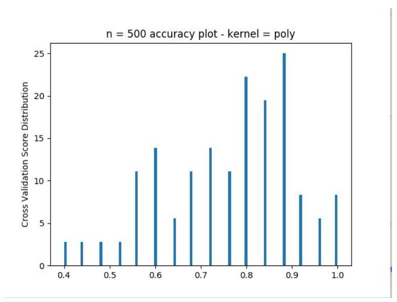
13. (0.5 points) Make a boxplot graph that plots accuracy (vertical) as a function of kernel choice. There should be 3 boxplots in the graph, one per kernel. Use results across all values for C. Don't forget to indicate n on your plot, where n is the number trials per boxplot. Don't forget to label your dimensions.

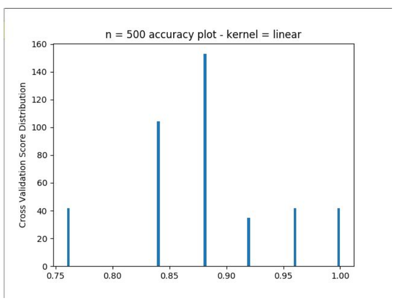


**CORRECTION: n = 60**

14. (0.25 points) What statistical test should you use to determine whether the difference between the best and second best kernel is statistically significant? Explain the reason for your choice. Consider how you selected testing and training sets and the skew of the data in the boxplots in your answer.







**A:**

**From the boxplot, we can find out that linear and rbf kernels yield the highest median cross-validation score.**

**However, the skew values in the boxplot are not sufficient to evaluate if the distributions are gaussian (or close to Gaussian). Therefore, I am using histograms as a reference. The two distributions can be regarded as independent,. Since neither distribution looks close to Gaussian, I am going to use Mann-Whitney test.**

15. (0.25 points) What is the result of your statistical test? Is the difference between the best and second best value of kernel statistically significant?

**A:**

**The result is 0.022337189265663422, less than 5%. Therefore, the difference between the two kernels has statistical significance.**