CSE 546 **Introduction to Machine Learning** Spring 2023

# Mid-Term Exam 1

# Feb 24, 2023

# (*Posted Friday 2/24 at 11:59am and due Saturday 2/25 at 3:00pm*)

# Submit ONE pdf file that has the answers to all questions, and

# ONE notebook file that has the code for problems 4 and 5. The output of the Notebook must match the answers in your pdf file.

**Problem #1 (10 points)**

Create a 2-dimensional data set with 30 samples that has the following properties:

* Samples should belong to 2 classes (15 samples per class)
* All samples can be classified correctly using Linear SVM with no regularization.
* Two samples from class 1 can be misclassified using Linear SVM with regularization.

Generate a scatter plot of your data. Use a different color/symbol for each class. **Indicate the 2 samples that cannot be classified correctly in (b). Explain why these points are classified correctly in (a) and incorrectly in (b).**

1. Provide a justification why the solution in (b) can be better than the one in (a)
2. Provide a justification why the solution in (a) can be better than the one in (b)

***Note: This data should be generated manually and you do not need to run any code on it***

**Problem #2 (10 points)**

Create a 2-dimensional data set with 30 samples that has the following properties:

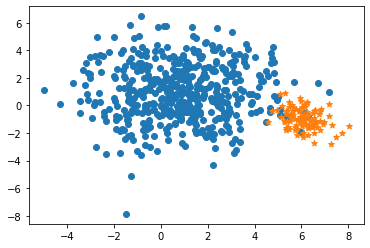
* Samples should belong to 2 classes (15 samples per class)
* All samples can be classified correctly using kNN with k=3
* 29 samples are classified correctly and 1 sample is misclassified using kNN with k=5
* Not all samples can be classified correctly using a linear classifier.

Generate a scatter plot of your data. Use a different color/symbol for each class. **Indicate the samples that cannot be classified correctly with different colors/symbols and justify your answers.**

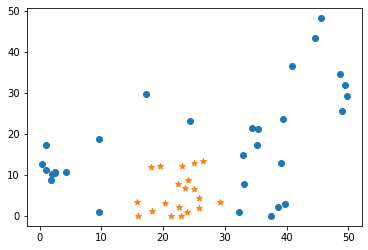
***Note: This data should be generated manually and you do not need to run any code on it***

**Problem #3 (15 points)**

1. Given the following 2-Dim data set with 2 classes. Out of all classifiers that we have covered in class, which one is the best candidate for this data? **Justify your choice**.



1. Given the following 2-Dim data set with 2 classes. Out of all classifiers that we have covered in class, which one is the best candidate for this data? **Justify your choice**.



**Problem #4 (30 points)**

For this problem, you need to use the built-in sklearn ***California Housing*** dataset. You can load this data using

from sklearn.datasets import fetch\_california\_housing

cal\_housing = fetch\_california\_housing()

Divide the data into training and test sets using train\_test\_split and random\_state=38

The goal is to experiment with few **regression** algorithms and compare their performance on this data.

1. Build and train a Ridge Regression model. Vary the constraint **parameter **and analyze the results by identifying cases of **overfitting** and **underfitting**. Select the optimal value of  and justify your choice.
2. Build and train a Decision Tree regression model. Vary the **pruning parameter**and analyze the results by identifying cases of **overfitting** and **underfitting**. Select the optimal pruning and justify your choice.
3. Build and train a Random Forest regression model. Vary the **n\_estimators, max\_features, and max\_depth,** and analyze the results by identifying cases of **overfitting** and **underfitting**. Select the optimal pruning and justify your choice.
4. Compare the **accuracy** of the 3 methods and the **relevant features** identified by each method and comment on the results.

**Problem #5 (35 points)**

For this problem, you need to use the built-in sklearn ***digits*** dataset. You can load this data using

from sklearn.datasets import load\_digits

digits = load\_digits()

(data will be stored in ***digits.data*** and desired output will be stored in **digits.target**)

Divide the data into training and test sets using train\_test\_split and random\_state=0

The goal is to train a ***Gradient Boosted Regression Trees*** classifier and optimize its performance on this data.

1. Design your experiment to learn the best values for the following parameters using *4-fold cross validation*:***n\_estimators, learning\_rates****,* and***max\_depth***
2. Analyze the results of the classifier using its optimal parameters and comment on its generalization capability.
3. **Visualize** and **explain** the relevant features identified by the Random Forest classifier.

* Create a white 8x8 image that represents the original 64 features. Map each identified relevant feature to this 2D image and display it using a grey scale that reflects its importance (e.g. 0 🡺most relevant feature and 255 🡺 least relevant feature).

1. Identify one misclassified sample from each class (if they exist). Visualize each misclassified sample as an 8x8 image, and use its nearest neighbors and the learned important features to explain why it was misclassified.

Hint: for examples on how to read this data and visualize it, check

<https://scikit-learn.org/stable/auto_examples/classification/plot_digits_classification.html#sphx-glr-auto-examples-classification-plot-digits-classification-py>