1 Theory

1. **Decision Tree (3 points)** Consider the Gini index, classification error and entropy in a simple classification setting with two classes. Create a single plot that displays each of these quantiites as a function of $\hat{p_{m1}}$. The $x-$axis should display $\hat{p_{m1}}$, ranging from 0 to 1, and the $y-$axis should display the Gini index, classification error and entropy. Note that $\hat{p_{mk}}$ represents the proportion of training observations in the mth region that are from kth class. **(3 points)**

2. **Decision Boundaries (3 points)** With $p = 2$ dimensions (that is, number of features), a linear decision boundary takes the form:

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 \ = \ 0.$$

We now investigate a non-linear decision boundary.

(a) Sketch the curve

$$(1 + x_1)^2 + (2 - x_2)^2 \ = \ 4.$$

(b) On your sketch, indicate the set of points for which

$$(1 + x_1)^2 + (2 - x_2)^2 > 4,$$

as well as the set of points for which

$$(1 + x_1)^2 + (2 - x_2)^2 \leq 4.$$

(c) Suppose that a classifier assigns an observation to the blue class if

$$(1 + x_1)^2 + (2 - x_2)^2 > 4,$$

and to the red class otherwise. To what class is the observation $(0, 0)$ classified? $(-1, 1)$? $(2, 2)$? $(3, 8)$?

(d) Argue that while the decision boundary in (c) is not linear in terms of x_1 and x_2, it is linear in terms of x_1, x_1^2, x_2 and x_2^2.

3. **Bootstrap Confidence Intervals** (4 **points)** We have a random sample of 30 observed tip percentages. This data has a positive skewness.

    22.7 16.3 13.6 16.8 29.9 15.9 14.0 15.0 14.1 18.1 22.8 27.6 16.4 16.1 19.0

    13.5 18.9 20.2 19.7 18.2 15.4 15.7 19.0 11.5 18.4 16.0 16.9 12.0 40.1 19.2

    (a) Plot a histogram of the given data sample.

    (b) Compute the mean of the data and a 95% confidence interval, assuming normal distribution.

    (c) From the given observations, draw 1000 **new samples**, each of size 30. The samples are drawn uniformly, and **with replacement**. Compute the mean from each of the new samples. Call the mean x^*. Therefore, you have 1000 such means $x_1^*, x_2^* \dots x_{1000}^*$.

    (d) Plot a histogram of the 1000 means computed in (c). What do you observe?

    (e) Compute a 95% confidence interval from the bootstrap samples.

2 Project

2.1 Tasks

1. Decision Tree (5 **points)**

    - Implement the decision tree algorithm for classification on the **Titanic data set**. Split the data into training and test sets in a $60 - 40$ ratio. That is, use 60% of the data for training and the remaining 40% for testing. You may utilize Python or Matlab libraries for the task.
    - Calculate the accuracy of prediction on the training **and** the test set. Plot the accuracies versus the depth of the tree. Comment if and when the model is overfitting.
    - Explain how a k-fold cross-validation is implemented.
    - Perform a 5-fold cross-validation. Comment on the performance.

2. Support Vector Machine (5 **points)**

    - Use Support Vector Machine (SVM) algorithm to perform classification task on the **Wisconsin Breast Cancer data set**. You may utilize Python or Matlab libraries for the SVM task.
    - Perform data visualization with scatter plots (or seaborn.pairplot) in feature space.
    - Plot decision regions of the classifier in two-dimensional feature space.

2.2 Data Set

- For the decision tree project, you will be working with the Titanic data set. There are a total of 891 observations. Each row of the data has 12 columns. The second column **Survived** is the class label or the target variable which we want to predict. It takes two values, namely, 1 (survived) and 0 (not survived). Some of the features are **Age**, **Gender**, **Plcass** (ticket class) with values 1, 2 and 3, **SibSp**, the number of siblings/spouses aboard the ship, **Parch**, the number of parents/children aboard the ship and **Embarked**, the port of embarkation. The data set and its description are posted on the blackboard. Note that some features some as **PassengerId**, **Name** and **Ticket** may not be relevant for classification.

- For the support vector machine (SVM) project, you will be working with the Wisconsin Breast Cancer data set. There are a total of 699 observations. Each row of the data has 11 columns. The last column **Class** is the class label which we want to predict. It takes two values, namely, 2 (benign) and 4 (malignant). Some of the features are **Clump Thickness** and **Cell Size Uniformity**. These features take value in the $1-10$ range. The first column **Sample Id** is not relevant for classification. The data set and its description are posted on the Blackboard.

3 Project Report

The project **must** include a report with several sections, as suggested below. Please limit the report to **ten pages**.

- Introduction: Describe the project background and goals.

- Description: Detailed description of the project including any necessary equations.

- Algorithm and Implementation: Describe the algorithms, and the names of the Python/Matlab libraries utilized. You may use code snippets to describe the algorithms.

- Results and Conclusions: Analysis, plots, feature reduction/selection, interpretations and numerical errors. How do the results compare with theory? Summarize your findings.

- References and Citations: Please provide all citations, for example, the libraries utilized,