

# Machine Learning, Fall 2020: Project 2

## 1 Linear Regression (15pt)

In this exercise, you will implement a linear regression model to predict the house price. For this exercise use the dataset from the link below. Only use a single feature for your regression model and explain your reasons for selecting that feature. Please explain the data setting and experimental setup similar to Project 1.

The key components of your linear regression model are the cost function and gradient descent method to update the weights.

1. <https://www.kaggle.com/mayanksrivastava/predict-housing-prices-simple-linear-regression/data>

## 2 Decision Trees (15 pt)

### 2.1 ID3

Consider the following set of training examples for the unknown target function  $\langle X_1, X_2 \rangle \rightarrow Y$ .

$Y$	$X_1$	$X_2$	Count
+	T	T	3
+	T	F	4
+	F	T	4
+	F	F	1
-	T	T	0
-	T	F	1
-	F	T	3
-	F	F	5

1. What is the sample entropy  $H(Y)$  for this training data (with logarithms base 2)?
2. What are the information gains  $IG(X_1) \equiv H(Y) - H(Y|X_1)$  and  $IG(X_2) \equiv H(Y) - H(Y|X_2)$  for this sample of training data?
3. Draw the decision tree that would be learned by ID3 (without postpruning) from this sample of training data.

## 3 Perceptron (35 pt)

Please use the notebook provided for you to complete the perceptron exercise.

## 4 Support Vector Machine (35 pt)

In this problem, you will repeat the format of Project 1 but using an SVM. On the Breast Cancer Wisconsin (Diagnostic) Data Set. See associated link.

<http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29>