

2 Programming (8 points)

1. Decision Tree (required) (4 points)

Task description Fit (*i.e.*, regression) the real variable *Sales* in the **Carseats** dataset, using decision tree, bagging, and random forests. All these algorithms can be implemented by calling **sklearn** in Python. The loss is set as sum of squared error (SSE).

Dataset **Carseats** contains 400 data points (saved in 400 rows). For each data, the first column is the value of target variable *Sales* that we want to fit; the remaining 9 columns indicate 9 features (or attributes), as shown in Fig. 1. There is no fixed train/test splitting. In this project, you have two options: simply set the first 300 rows as the training set, and the remaining 100 rows as the testing set; randomly split the whole dataset to 300 train + 100 test, and try multiple times. The dataset can be obtained from <https://github.com/selva86/datasets/blob/master/Carseats.csv>

What you should do

- **Data statistics:** analyze the statistics of the target variable and each feature, and try to visualize the statistics (*e.g.*, histogram) (0.5 point)
- **Decision tree:** solve the above problem using decision tree method; report the train/test errors with respect to different maximum depths, different least node sizes; plot the learned tree (2 points)
- **Bagging of trees:** solve the above problem using the bagging method, with decision tree as the base learner; report the train/test errors with respect to different depths, different number of trees (2 points)
- **Random forests:** solve the above problem using the random forest method, with decision tree as the base learner; report the train/test errors with respect to different number of trees, different values of m (the number of candidate attributes to split in every step, see Slides ‘W7-Decision Tree’, Page 68) (2 points)
- Plot the curve of bias^2 with respect to different number of trees in random forests, *e.g.*, $\#tree = 10, 20, \dots, 100$. Then, describe the relationship between bias^2 and different number of trees; repeat the procedure for variance. (1.5 points)

Carseats

Sales	CompPrice	Income	Advertising	Population	Price	ShelveLoc	Age	Education	Urban	US
9.5	138	73	11	276	120	Bad	42	17	Yes	Yes
11.22	111	48	16	260	83	Good	65	10	Yes	Yes
10.06	113	35	10	269	80	Medium	59	12	Yes	Yes
7.4	117	100	4	466	97	Medium	55	14	Yes	Yes
4.15	141	64	3	340	128	Bad	38	13	Yes	No
10.81	124	113	13	501	72	Bad	78	16	No	Yes
6.63	115	105	0	45	108	Medium	71	15	Yes	No
11.85	136	81	15	425	120	Good	67	10	Yes	Yes
6.54	132	110	0	108	124	Medium	76	10	No	No
4.69	132	113	0	131	124	Medium	76	17	No	Yes
9.01	121	78	9	150	100	Bad	26	10	No	Yes
11.96	117	94	4	503	94	Good	50	13	Yes	Yes
3.98	122	35	2	393	136	Medium	62	18	Yes	No
10.96	115	28	11	29	86	Good	53	18	Yes	Yes
11.17	107	117	11	148	118	Good	52	18	Yes	Yes
8.71	149	95	5	400	144	Medium	76	18	No	No
7.58	118	32	0	284	110	Good	63	13	Yes	No
12.29	147	74	13	251	131	Good	52	10	Yes	Yes
13.91	110	110	0	408	68	Good	46	17	No	Yes
8.73	129	76	16	58	121	Medium	69	12	Yes	Yes
6.41	125	90	2	367	131	Medium	35	18	Yes	Yes
12.13	134	29	12	239	109	Good	62	18	No	Yes
5.08	128	46	6	497	138	Medium	42	13	Yes	No
5.87	121	31	0	292	109	Medium	79	10	Yes	No
10.14	145	119	16	294	113	Bad	42	12	Yes	Yes

Figure 1: Some examples of Carseats.

2. Handwritten Digit Recognition using sk-learn (4 points)

Task description Use the function "MLPClassifier" of sk-learn to construct a fully connected network to classify the MNIST data. The dataset can be downloaded from <http://yann.lecun.com/exdb/mnist/>. The training-testing spilling has been given. Show the performance of your neural network with different structures:

- number of hidden layers chosen from $\{1, 2, 3\}$
- number of hidden nodes chosen from $\{50, 200, 784\}$ (let all hidden layers have the same number of nodes)

Note that you should submit [A3_StudentID.pdf](#) (report, together with the written answers), and [A3_StudentID.ipynb](#) (code). Please zip them into "A3_StudentID.zip". The reference report is in Assignment 1. You can check it on BlackBoard. (You can submit several files in one submission. Don't submit them in different submissions.) **Your report for the programming questions should include necessary formulas, charts, and explanations. The number of pages should be 4-5.**