Exercise 12 : Decision Tree, Bias-variance trade-of, Linear Discriminant Analysis, Out-of-Distribution Detection

Exercise 12-1 : Decision Tree

Given the training data about buying computer in the table below

RID	age	income	student	credit rating	Buys computer
1	≤ 30	high	no	fair	no
2	≤ 30	high	no	excellent	no
3	$31 \cdots 40$	high	no	fair	yes
4	> 40	medium	no	fair	yes
5	> 40	low	yes	fair	yes
6	> 40	low	yes	excellent	no
7	$31 \cdots 40$	low	yes	excellent	yes
8	≤ 30	medium	no	fair	no
9	≤ 30	low	yes	fair	yes
10	> 40	medium	yes	fair	yes
11	≤ 30	medium	yes	excellent	yes
12	$31 \cdots 40$	medium	no	excellent	yes
13	$31 \cdots 40$	high	yes	fair	yes
14	> 40	medium	no	excellent	no

(a) Create a decision tree based on Gini index.

(b) Classify the test instances :

i) age$\leq$30, income=medium, student=yes, credit rating=fair

ii) age$>$40, income=excellent, student=no, credit rating=no

iii) age $31 \cdots 40$, income=low, student=no, credit rating=no

Exercise 12-2 : Bias-variance trade-of

You have access to a European database of 1000000 individual trees of various types which include the following entries :

- Tree type (birch, pine, aspen, etc.). In total 98 different classes.
- Age
- Height
- Circumference (at 1-meter height)
- Geographical coordinate of the position of the tree
- Vegetation type (open woodland, mixed wood, highland, wet coniferous etc.)

All parts of Europe are well represented in the database. Consider a regression problem where you want to model the age of a tree based on its height and circumference. We use a linear regression model with two input variables

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \epsilon$$

where the input variables represent the height and the circumference, and the output is the age.

(a) What causes the bias of the model ?

(b) What causes the variance of the model ?

(c) What causes the irreducible error of the model ?

(d) Where do you see the biggest improvement potential of the model (reducing bias, variance or irreducible error) and how would you go about improving it ?

(e) Compare bias and variance of this model with kNN with small k and large k.

Exercise 12-3 : Linear Discriminant Analysis

Scientists of Iowa state have acquired the samples of water from state's reservoirs. Some water samples contain a particular bacterium (class 1) while other do not contain (class 2). The samples have two observed variables x_1(pH) and x_2 (Nitrogen content). The number of instances in each class, the average of the variable vectors and the covariance matrices for the two types of water samples are given as follows :

$$n_1 = 13, \quad n_2 = 10$$
$$\mu_1^T = \begin{pmatrix} 7.8 \\ 45 \end{pmatrix}, \quad \mu_2^T = \begin{pmatrix} 5.9 \\ 20.8 \end{pmatrix}$$
$$\Sigma_1 = \begin{pmatrix} 0.5 & 4.5 \\ 4.5 & 147.2 \end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix} 0.1 & 0.2 \\ 0.2 & 24.2 \end{pmatrix}$$

(a) Determine the discriminant function for the two classes.

(b) Assign the observation $x = \begin{pmatrix} 6 & 52.5 \end{pmatrix}$ to one of the classes.

Exercise 12-4 : Tools : Out-of-Distribution Detection

(a) Import python packages : `train-test-split` from `sklearn.model-selection`, `datasets, metrics` from `sklearn, KNeighborsClassifier` from `sklearn.neighbors, pyplot` from `matplotlib`, and `make-classification` from `sklearn.datasets`.

(b) Load `make-moons` dataset with 2000 samples, split that into random train and test subsets and assign 20 percent of data to test dataset.

(c) Train the K- Nearest Neighbours classifier model with k=3.

(d) Creat ROC curve on this dataset by using `metrics.roc-curve`.

(e) Generate 2 class random dataset with `make-classification`, assign 2000 samples to it, and split that into train and test subsets, assign 20 percent of data to test dataset.

(f) Create ROC curve on random dataset with previous trained model.

(g) Plot both ROC curves along with a random classifier which represents points along the diagonal.

(h) Evaluate the difference between curves and the reason.