Data mining (2019/11/4)

1. (20%)With the following dataset, please build a two-level decision tree. Then report its prediction accuracy, recall rate, precision rate and F1 measures with the same dataset as the testing dataset.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| X | Y | Z | No. of Class C1 Examples | No. of Class C2 Examples |
| 0 | 0 | 0 | 5 | 40 |
| 0 | 0 | 1 | 0 | 15 |
| 0 | 1 | 0 | 10 | 5 |
| 0 | 1 | 1 | 45 | 0 |
| 1 | 0 | 0 | 10 | 5 |
| 1 | 0 | 1 | 25 | 0 |
| 1 | 1 | 0 | 5 | 10 |
| 1 | 1 | 1 | 5 | 20 |

1. (20%) Answer the following questions:
   1. Give the definition of data mining. (5%)
   2. What are support vectors in an SVM? (5%) What is the difference between a hard margin SVM and a soft margin SVM? (5%) How kernel functions are useful in training an SVM? (5%)
2. (20%) (Saturday morning problem—To play or not to play?) The very first example for explaining Decision Trees is the Saturday morning problem. I have modified the class distribution of the dataset. Please use the Naïve Bayes classifier to determine the class label of the following testing example:

**X**=（Outlook=Rain, Temperature=Mild, Humidity = Normal, Wind = Strong）

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Day** | Outlook | Temperature | Humidity | Wind | PlayTennis |
| **D1** | Sunny | Hot | High | Weak | No |
| **D2** | Sunny | Hot | High | Strong | No |
| **D3** | Overcast | Hot | High | Weak | Yes |
| **D4** | Rain | Mild | High | Weak | Yes |
| **D5** | Rain | Cool | Normal | Weak | Yes |
| **D6** | Rain | Cool | Normal | Strong | No |
| **D7** | Overcast | Cool | Normal | Strong | Yes |
| **D8** | Sunny | Mild | High | Weak | No |
| **D9** | Sunny | Cool | Normal | Weak | Yes |
| **D10** | Rain | Mild | Normal | Weak | Yes |
| **D11** | Sunny | Mild | Normal | Strong | Yes |
| **D12** | Overcast | Mild | High | Strong | Yes |
| **D13** | Overcast | Hot | Normal | Weak | Yes |
| **D14** | Rain | Mild | High | Strong | No |

1. (20%) Given a simple MLP with two inputs a and b, one hidden node c and one output node d, please find the updated weight of **Wcd** after training the network once with the following two samples. Assume that the initial values on the weights of wac, wbc, wcd are 0.1, 0.1, 0.1, respectively, and the learning rate is 0.3.

(notes:1. Let not consider the biases (no biases) and the changes on Wac, Wbc , for simplicity. **Wcd** is the weight on the link between node **c** and node **d**

2. Use batch update where updates are accumulated until finishing the calculation on all the samples in the dataset.

3. The accumulated updates need to be averaged before updating the weights.)



|  |  |  |
| --- | --- | --- |
| a | b | d |
| 1 | 0 | 1 |
| 0 | 1 | 0 |

Appendix：

**Output nodes**

**Input nodes**

**Hidden nodes**

**Output vector**

**Input vector: *xi***

*wij*



1. (20%) (Gini index) Gini index= 
   1. Find the maximum value of the Gini index for a dataset with three different class labels. (Hint 1: we now have three probabilities P1, P2, and P3 for three different classes of the dataset with the constraint of P1 +P2 + P3 = 1； Hint 2: You can use the Lagrangian method to solve this simple constrained optimization.) (10%)
   2. Find the maximum value of the Gini index for a dataset with *n* different class labels. (10%)