CSCE633 Exam 03

Lu Sun

November 2020

(1 point) (1) Assume that you would like to run an ensemble learning method through multi-expert combination with voting on a set of 10 features. You will be using linear perceptron as the base classifier. What would be three possible ways to ensure the diversity of the linear perceptron models? Please explain your answer.

Answer:

Way 1: Different hyper parameters in the same model.

Way 2: Different input features for models.

Way 3: Different kinds of model classification.

In way 1, for example the number of perceptron, the max iteration and the learning rate in each model can make the final model different.

In way 2, for example, some models are given image form of input, some are give string or number form of input.

In way 3, for example, some models are parametric model and some are non-parametric model which can finally result in the diversity of the linear perceptron models.

(1.5 points) (2) Please select the correct answer(s) and justify.

Answer: A

For first dimension space, the mean value is $\mu_1 = \frac{1}{3}\Big(x_1(1)+x_2(1)+x_3(1)\Big) = (1+0-1)/3 = 0$; the variance value is $var_{11} = \frac{1}{3}\sum_{i=1}^{3}(x_i(1)-\mu_1)^2 = \frac{1}{3}(1^2+0^2+(-1)^2) = 2/3$.

For second dimension space, the mean value is $\mu_2 = \frac{1}{3}\Big(x_1(2) + x_2(2) + x_3(2)\Big) = (-1+1+0)/3 = 0$; the variance value is $var_{22} = \frac{1}{3}\sum_{i=1}^{3}(x_i(2)-\mu_2)^2 = \frac{1}{3}((-1)^2+1^2+(0)^2) = 2/3$.

$var_{12} = var_{21} = \frac{1}{3}\sum_{i=1}^{3}(x_i(1)-\mu_1)(x_i(2)-\mu_2) = \frac{1}{3}(-1+0+0) = -1/3$.

Finally, $\Sigma = \begin{bmatrix} var_{11} & var_{12} \\ var_{21} & var_{22} \end{bmatrix} = \frac{1}{3}\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$. Namely, choose A.

(1.5 points) (3) Please select the correct answer(s) and justify.

Answer: B

According to problem (2), the corresponds covariance matrix is $\Sigma = \frac{1}{3}\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$. Setting $0 = det|\lambda I - \Sigma| = (\lambda - \frac{2}{3})^2 - \frac{1}{3}^2$. The corresponding eigenvalue are $\frac{1}{3}, 1$. The largest one is 1. Then $\lambda_{max} I - \Sigma = \begin{bmatrix} 1/3 & 1/3 \\ 1/3 & 1/3 \end{bmatrix}$. Since u_1 is eigen-vector of 1, $(\lambda_{max} I - \Sigma)u_1 = 0$.

A: 1/3-1/3=0. However, $||u_1||_2^2 = 1+1 = 2$, not one, which can't be eigen-vector.

B: $1/3 \times \frac{\sqrt{2}}{2} - 1/3 \times \frac{\sqrt{2}}{2} = 0$. $||u_1||_2^2 = 1/2 + 1/2 = 1$. The true answer.

C: $1/3 \times \frac{2\sqrt{5}}{5} + 1/3 \times \frac{\sqrt{5}}{5} \neq 0$.

D:$1/3 \times \frac{2}{5} + 1/3 \times \frac{3}{5} \neq 0$.

(1 point) (4) For which of the following datasets is K-Means clustering (K=2) more likely to separate the samples correctly into the blue and yellow categories? Please select the correct answer(s) and justify.

Answer: Dataset 2.

To separate the samples correctly, K-Means clustering needs two separately spherical-like data sets.

Dataset 1: the data are not distributed around 2 centroids and have some overlap in the center. The data in the overlap can be mis-classified.

Dataset 2: 2 centroids can easily be piked up that will perfectly classify all the data.

Dataset 3: the data are not distributed around 2 centroids and have difficult to category into 2 classes.

(2 points) (5) Please select the correct answer(s) and justify.

Answer: (i).

2 closest points need to be merged every time.

First step, merge A and B, for the smallest distance is 1 between A and B.

Second step, merge AB with C, for the smallest distance is 2 between B and C.

Third step, merge ABC with D, for the smallest distance is 3 between C and D.

Finally, the dendogram (i) is generated.

(1 point) (6) Please select the correct answer(s) and justify.

Answer: (i)

The information entropy of (i) and (ii) is as follow:

(i): Class one have 5 'no' and 0 'yes', class two have 4 'yes' and 0 'no', and class three have 3 'yes' and 2 'no'. $Entropy = 0 + 0 + \frac{3}{3+2}(1 - \frac{3}{3+2}) + \frac{2}{3+2}(1 - \frac{2}{3+2}) = 12/25 = 0.48$

(ii): Class one have 4 'no' and 3 'yes', class two have 3 'yes' and 4 'no'. $Entropy = 2 * (\frac{3}{3+4}(1 - \frac{3}{3+4}) + \frac{4}{3+4}(1 - \frac{4}{3+4})) = 48/49 = 0.9796$

The information gain of (i) and (ii) is as follow: (i) -0.48, (ii) -0.9796. $-0.48 > -0.9796$, the more the better. Choose (i).

(1 point) (7) Which of the following statement(s) is/are true when performing a binary classification task with AdaBoost using shallow decision trees as base classifiers? (A) Adaboost will depict a lower bias compared to a single decision tree; (B) The decision tree base classifiers can be trained in parallel when using Adaboost; (C) Adaboost will depict a lower variance compared to a single decision tree. Please select the correct answer(s) and justify.

Answer: (A)(C)

(A) Correct. The weak learners are combined in Adaboost and depict a lower bias compared to a single decision tree.

(B) Incorrect. The decision tree is trained recursively and can't be trained in parallel.

(C) Correct. The weak learners are combined in Adaboost and depict a lower variance compared to a single decision tree.

(1 point) (8) Which of the following statement(s) is/are true regarding the entropy of a probabilistic distribution? (A) It is maximized when we have a uniform distribution; (B) It can be a good criterion for feature selection; (C) It can be a good criterion for picking the attributes to be used as nodes of a decision tree; (D) It can be used for both continuous and discrete probabilistic distributions. Please select the correct answer(s) and justify.

Answer: (A)(B)(C)(D)

(A) Correct. The uniform distribution has the biggest uncertainty and distributed equally, which will result in a maximum information entropy.

(B) Correct. When in unsupervised problem, the feature selection models, which rely on the availability of labeled data, are much less than entropy feature selection models, which more suitable in unsupervised problem.

(C) Correct. The feature with lowest entropy value show the lowest uncertainty, will be picked as the node first.

(D) Correct. Both continuous distribution(through integral operator) and discrete distributions(through sum operator) can use the entropy.