Data Application Lab

Tuesday Machine Learning Basis Quiz

1. **What is supervised machine learning? What is unsupervised? What is the difference? Use examples to explain this.**

Supervised Learning: if you are training your machine learning task for every input with corresponding target, it is called supervised learning, which will be able to provide target for any new input after sufficient training.

An example: You have a dataset including three-cluster data, you want to train a model and predict which class it belongs when there is new input.

Unsupervised Learning: if you are training your machine learning task only with a set of inputs, it is called unsupervised learning, which will be able to find the structure or relationships between different inputs.

An example: You have a dataset without knowing clusters, you want to train a model and divide the dataset into several classifications basing on their own properties.

2. **What is the differences between classification and regression?**

In classification, the target variable is categorical. In regression, the target variable is continuous.

3. **What are some evaluation methods of regression? Using MSE, RMSE, MAE, to evaluate following values. (Given y_true = [100, 50, 30, 20]; y_pred = [90, 50, 50, 30])**

Mean Absolute Error (MAE) is the mean of the absolute value of the errors:

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Mean Squared Error (MSE) is the mean of the squared errors:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Root Mean Squared Error (RMSE) is the square root of the mean of the squared errors:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

MAE = 10.0
MSE = 150.0
RMSE= 12.2474487139

4. What's the disadvantages of linear regression?

- Linear regressions are sensitive to outliers.
- Linear regressions are meant to describe linear relationships between variables. (However, this can be compensated by transforming some of the parameters with a log, square root, etc. transformation.)
- Linear regression assumes that the data are independent.
- …

5. Explain what precision and recall are. How do they relate to the ROC curve?

TN / True Negative: case was negative and predicted negative

TP / True Positive: case was positive and predicted positive

FN / False Negative: case was positive but predicted negative

FP / False Positive: case was negative but predicted positive

The precision: TP/(TP+FP)

the probability that a the true positive among the predicted positive, a measure of how many of the samples predicted by the classifier as positive is indeed positive.

The recall: TP/(TP+FN) the probability that a the true positive among the conditioned positive, a measure of how many of the positive samples have been identified as being positive.

ROC curve represents a relation between sensitivity (RECALL) and specificity(NOT PRECISION) and is commonly used to measure the performance of binary classifiers.

Reference:

https://en.wikipedia.org/wiki/Precision_and_recall