ML Questionnaires

1. Talk about your understanding of supervised learning, semi-Supervised learning, and unsupervised learning, as well as their application scenarios.

The main difference between these learning methods lies in the training data.
Supervised Learning uses labeled data for training, meaning the training data includes corresponding correct answers. This approach is commonly used for classification, regression, and similar problems.
Semi-Supervised Learning involves using a small amount of labeled data along with a large amount of unlabeled data. The goal is to improve the model's performance by leveraging the structure of the unlabeled data. This method is often used when labeling data is expensive, such as in text classification or image recognition.
Unsupervised Learning does not use labeled data for training. The aim is for the model to discover specific features, structures, or patterns from the data itself. It is typically used for data exploration and feature extraction, such as in market research and big data analysis. Generative models, which are common today, also fall under the category of unsupervised learning.

2. Discuss your understanding of deep learning and how it differs from traditional machine learning methods.

Deep learning, on the other hand, involves a structure composed of multiple layers of neural networks with a large number of parameters. This requires substantial data and longer training times. However, deep learning can address a broader range of problems compared to traditional machine learning.
Traditional machine learning requires manual feature engineering and algorithm design to extract features, resulting in simpler models such as linear regression, decision trees, and support vector machines. These models are suitable for smaller datasets and are generally more interpretable.
Deep learning is capable of learning complex data patterns and handling problems that traditional machine learning methods cannot solve.

3. Please explain the problem of "overfitting" in machine learning, and how you avoid or address it. Please explain from the perspectives of both traditional machine learning and deep learning.

During the training process, there is a tendency to increase the accuracy on the training data. However, when the model's performance on the validation dataset starts to worsen, it indicates overfitting. You can think of it as the machine memorizing the answers, which leads to poorer performance on new data. This issue might arise because the training data contains too much noise, preventing the model from learning the necessary features.

Solutions for Overfitting in Traditional Machine Learning:

- Simplify the Model: Use simpler models with fewer parameters.
- Reduce Features: Decrease the number of features used for training.
- Add Regularization: Implement regularization techniques to constrain the model.
- Early Stopping: Stop training before overfitting occurs.

Solutions for Overfitting in Deep Learning:

- Regularization: Use techniques like L1 and L2 regularization to reduce model complexity.
- Dropout: Apply dropout to randomly exclude units during training, preventing over-reliance on any single neuron.
- Early Stopping: Stop training before overfitting occurs.
- Data Augmentation: Enhance the training data by introducing variations, increasing its diversity, and improving the model's generalization ability.
- Ensemble Learning: Combine the predictions of multiple simpler models to enhance overall generalization.

4. How do you evaluate the performance of a machine learning model? Please explain some commonly used performance metrics.

For classification tasks

We can use confusion matrix to visualize the results.

	Predict label		
True label		Positive	Negative
	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Different classification problems emphasize different metrics. Here are some common ones:

Accuracy	$\frac{\mathbf{TP + TN}}{\mathbf{TP + FP + TN + FN}}$
Precision	$\frac{\mathbf{TP}}{\mathbf{TP + FP}}$
Recall	$\frac{\mathbf{TP}}{\mathbf{TP + FN}}$
F1-score	$\frac{\mathbf{2 \times Precison \times Recall}}{\mathbf{Precision + Recall}}$
Specificity	$\frac{\mathbf{TN}}{\mathbf{FP + TN}}$

For Regression Metrics

1. Mean Absolute Error:

average of the absolute differences between predicted and actual values.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |Predicted_i - Actual_i|$$

2. Mean Squared Error(MSE):

The average of the squared differences between predicted and actual values.

$$MSE = \frac{1}{n}\sum_{i=1}^{n} |Predicted_i - Actual_i|^2$$

3. Root Mean Squared Error(RMSE):

The square root of the mean squared error.

$$RMSE = \sqrt{MSE}$$

5. How do you handle imbalanced datasets? Would the approach differ when the imbalance ratio is 1:10 versus 1:10000?

Imbalanced datasets can be addressed from two aspects: data handling and model adjustments.

For data handling, we can use methods such as oversampling or undersampling. Common techniques include:

- SMOTE (Synthetic Minority Over-sampling Technique): Generates synthetic examples by interpolating between existing minority class instances.
- ADASYN (Adaptive Synthetic Sampling): Similar to SMOTE but focuses on generating more synthetic samples near difficult-to-classify instances.
-

For model adjustments, we can increase the weight of the minority class data.

- Ensemble methods: Ensemble techniques can be highly effective in dealing with severe imbalances by combining the strengths of multiple balanced models.
-

For moderate imbalance (1:10): Apply a combination of resampling techniques, adjust the weight of dataset, these methods might could handle this.

For extreme imbalance (1:10,000): emphasize advanced resampling methods, ensemble models might help for extreme imbalance, and carefully selected evaluation metrics.