112學年第二學期第一次考試　　**科目**：機器學習

班級：資管碩一　　**應考人數**：36　**日期**：113/03/28　　**考試時間**：2:10-4:00

注意：可使用計算器，不可使用任何字典。

1. What is generalization error? (5%)
   (a) The error measure on the training set.
   (b) The ability to perform well on previously unobserved inputs.
   (c) The gap between the training error and test error is too large.
   (d) The expected value of the error on a new input.

2. Which activation function is usually used for a binary classification problem at the last layer? (5%)
   (a) Linear (b) Softmax (c) Sigmoid (d) ReLU.

3. The difference between the ground truth and the average of the estimations is referred to as ____. (5%)
   (a) Bias (b) Variance (c) Average error (d) MSE.

4. What regularization technique builds multiple models and evaluating multiple models on each test example? (5%)
   (a) Dropout (b) Bagging (c) Multi-task learning (d) Adversarial training.

5. What are the functions of 1*1 convolution? (5%)
   (a) To reduce the number of parameters. (b)To increase the number of feature maps.
   (c) To increase attention. (d) To decrease the number of feature maps.

6. What statements are true? (5%)
   (a) Comparing to shallow neural networks, deeper neural networks may overfit more.
   (b) Using deeper neural networks can reduce the amount of test error.
   (c) Shallow neural network may need more width.
   (d) Using deeper neural networks can reduce the number of units required to represent the desired function.

7. Please explain the following terms. (Note: do not just translate) (20%)
   (a) *K*-fold cross validation (b) Dropout (c) Zero padding (d) Global maximum pooling.

8. For a **multiclass, single-label classification** problem, what the activation function of the last layer of a neural network shall be used? (5%)

9. (a) Please explain the difference between **bias** and **variance**. (5%)
   (b) Please explain the difference between **spatial attention** and **channel attention**. (5%)

10. (a) Please explain what **momentum** is and what advantage momentum can bring. (5%)
    (b) Please explain what **adaptive learning rate** is and what advantage it can bring. (5%)

11. (a) Assume the output of the last layer before the activation function is [5 2 1 1 2 4] and the activation is **softmax**. What is the **output vector** of the softmax activation function? (8%)
    (b) Assume the ground truth is [0 0 0 0 0 1], what is the loss $L(y,\hat{y})$ of **categorical cross entropy**? (7%)

$$\mathrm{softmax}(\mathbf{x}) = \frac{1}{\sum_{j=1}^{K}\exp(x_j)}\begin{bmatrix}\exp(x_1)\\ \exp(x_2)\\ \dots\\ \exp(x_k)\end{bmatrix} \quad \text{where} \exp(x_i) = e^{x_i} \text{ and } e = 2.718$$

$$L(y,\hat{y}) = -\sum_{i=1}^{k} y_i \log \hat{y}_i \quad \text{where } k\text{: number of categories.}$$

12. Assume batch size is 3, the output of the batch from a neural network is (0.8, 0.2, 0.7), and the ground truths is (0, 1, 0). What is the loss of **binary cross entropy**? (10%)
 Hint: Binary cross entropy: loss = $-\frac{1}{N}\sum_{n=1}^{N} y_n \log\hat{y}_n + (1-y_n)\log(1-\hat{y}_n)$