NTUST, CSIE
Machine Learning (CS5087701), Fall 2017
Midterm
Total: 111 pts

Name:_____

Student ID #_____

Question	Score
1 (6%)	
2 (45%)	
3 (15%)	
4 (15%)	
5 (15%)	
6 (15%)	
Total	

1. **[True or False 6%]** No explanations are necessary.
   (a) [2%] The training error is generally larger than the test error.

   (b) [2%] The ID3 algorithm for decision tree induction is a greedy algorithm.

   (c) [2%] Multi-layer neural network can only deal with linearly separable data when the sigmoid function is not used (i.e., without the thresholding).

2. **[Simple questions 45%]** Answer the following questions with as brief answer as possible.
   (a) [3%] Name any classifier that can deal with non-linearly separable data.

   (b) [3%] When can we reduce the Maximum A Posteriori to Maximum likelihood principle for prediction?

(c) [3%] Name one classifier where we have a complete hypothesis space.

(d) [6%] Use your own words (i.e., no need to copy from any other materials) to define the term “overfitting”.

(e) [6%] Give two strategies to avoid the overfitting situation.

(f) [6%] Give one example when the ID3 algorithm may not be perfect to find the optimal classifier.

(g) [6%] Explain the pros and cons about using a hypothesis of high VC dimension.

(h) [6%] Explain what the Laplace smoothing is for and why it is quite often a good strategy to apply.

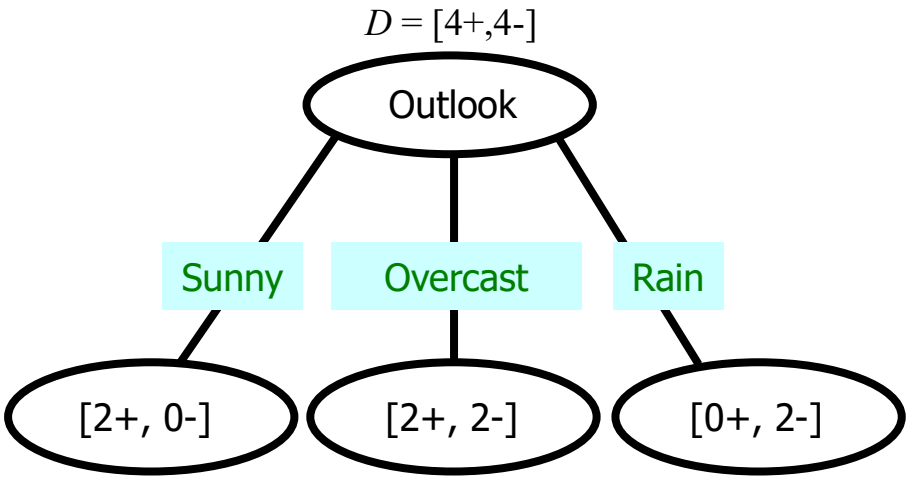
(i) [6%] Explain why using the “date” as one of the features in the decision tree induction may not be a good idea. How to avoid this situation?

3. **[Bayesian learning 15%]** Let us apply Bayes rule for several times for the coin flipping case, with the three flips given by $(x_1, x_2, x_3) = (1,0,1)$. Now suppose we also know that the posterior probability for the first two flips $(x_1, x_2) = (1,0)$ is given by:

$$p(h \mid x_1, x_2) = 6\theta(1-\theta)$$

for $\theta = p(x_i = 1)$. Can you compute $p(h \mid x_1, x_2, x_3) = p(h \mid (1,0,1))$?

4. **[Decision trees 15%]** Compute the information gain for the following case.



5. **[Neural networks 15%]** Step-by-step derive the learning rule with only one perceptron with a sigmoid threshold function σ applied. (hint: computing partial derivatives on the error function to find the minimizer!)
   The error function is given by:

$$\frac{1}{2}\sum_{i=0}^{m}(y_i - h(x_i))^2$$

   where the function *h* is given by:

$$h(x_i) = \sigma\left(\sum_j w_j x_{ij}\right)$$

6. **[Mixed 15%]** Answer the following questions.
   (a) [8%] For the graph on the right, give your support on (1) the squared model, (2) the curved model by some explanations.
   (b) [7%] How many undecided parameters do we have for a Gaussian Naïve Bayes on a dataset of n attributes? Write down those parameters.