Data science study group prerequisite

Outline

You will be qualified to registered for 1-3 credit independent study if you complete the following work before the semester starts.

1. Register for an online class
   (a) Visit `https://www.coursera.org/`
   (b) Search for "machine learning"
   (c) Register for the class
      - Title: Machine Learning
      - Institute: Stanford University
      - Instructor: Andrew Ng
2. Finish week 1 - 8 of the online class as required below. A week's session is considered to be complete if you
   (a) Watch the video lecture.
   (b) Compelte online quiz offered by the online course.
   (c) Complete the coding assignment offered by the online course and get a 100% percent grade.
   (d) Create your own **DIGITAL TYPESET** notes for each topic. Latex is recommended but not required.
   (e) Complete the question sheet (next section) and type your answer along with your notes.
3. Send the proof of your work and your digital note to schen@uwlax.edu or cvidden@uwlax.edu one week before the semester starts.

Question sheet

Remark:

1. All questions are based on the online course.
2. For the "summarize" questions, I recommend you to use bullet list.
3. For all the formulas mentioned in the class, try to write it in matrix form.

Week 1-2

1. What's the difference between supervised and unsupervised learning?
2. Give three examples of supervised learning problem.
3. Give three examples of unsupervised learning problem.
4. Summarize the idea of linear regression.
5. Summarize the idea of gradient descent method.
6. Why do we prefer gradient descent method to normal equations?
7. Why do we need the cost function to be convex in order to apply gradient descent method?
8. What will go wrong if the learning rate α is too big or too small?
9. What is a feature normalization? Why do we need it?

Week 3

1. Give three examples of regression problem.
2. Give three examples of classification problem.
3. What is a decision boundary?
4. Summarize the idea of logistic regression.
5. Why can't we use the cost function of linear regression (the least square) with logistic regression?
6. Can we use logistic regression for regression problem?
7. Summarize the idea of multi-classification.
8. What do we mean by overfitting?
9. For regularization terms, what happens if you increase/decrease value λ? Explain why.

Week 4-5

1. Summarize the idea of neural network.
2. How to choose number of layers and number of nodes in each layer?
3. Why is the training process of neural network called "Backpropagation"?
4. What are the advantage/disadvantages of neural network compared to the logistic regression?
5. Why do we want to do random initialization for neural network? Why don't we need it for the logistic regression?

Week 6

1. What's the strategy for selecting which machine learning model to use?
2. Why do you need a cross validation set?
3. How to determine the percentage distribution of the training, cross validation and test set?
4. Group the following concepts by their relationship

    - High bias
    - High variance
    - Overfitting
    - Underfitting
    - Too "linear"
    - Too "nonlinear"
    - Increase λ
    - Decrease λ
    - Collect more samples
    - Big gap between the learning curve of the training and cross validation set.
    - High training error.

5. How to handle skewed classes?

Week 7

1. Summarize the idea of the SVM (support vector machine).
2. What's the advantage and disadvantage of SVM over logistic regression?
3. Why the SVM can find the largest margin?
4. Explain why manipulating the value of C to will change the sensitivity of the machine to outliers.
5. How to choose land marks?
6. Explain your understanding of the use of kernel function.
7. Why is the method called **SUPPORT VECTOR** machine?

8. (Required by math major. Optional for others.) Read and summarize the mathematical intuition of SVM (knowledge about the Lagrange multiplier is required).
    - `http://cs229.stanford.edu/notes/cs229-notes3.pdf`
    - `http://www.engr.mun.ca/~baxter/Publications/LagrangeForSVMs.pdf`
9. Google search "supervised learning". Find one more supervised learning method not introduced in the class. Try you best to summarize the idea.

Week 8

1. Give five applications of supervised learning.
2. Give five applications of unsupervised learning.
3. Summarize the idea of K-mean.
4. Why do we need random initialization of K-mean?
5. How to choose the number of clusters?
6. Google search "unsupervised learning" or "cluster analysis". Find two more unsupervised learning methods other than K-mean. Try you best to summarize their ideas.
7. Summarize the idea of the PCA (principal component analysis).
8. (Required by math major. Optional for others) Explain the mathematical idea of the algorithm of PCA using eigenvalue theories.

Week 9 - Week 11: optional