Security Analytics
Review for Final Exam

Purdue University

Prof. Ninghui Li

Exam Date/Time

- Monday Dec 11 (1:00p - 3:00p)
- HAAS G066

- What are the differences between supervised learning and unsupervised learning?

- Are the following supervised learning or unsupervised learning:
  - Clustering, classification

- Concepts of
  - **Model space**
  - **Scoring function**
  - **Search technique**

Topic 3

- Explain the kNN algorithm for classification.
  - What is the training process?
  - How to predict a sample x?
  - Does a high k value result in a more complex model or a simpler model (smoother decision boundary)?
  - How should one determine k?
  - Is training fast or slow?
  - How large is the model size?

Topic 4: Probability Review

- Able to do conditional probability computation
- Able to judge independent and dependent events
- Understand the base rate fallacy

Topic 5: CLassification

- Accuracy
- Precision and recall
- F1 score
- Naïve Bayes on discrete-valued features
- Smoothing

Topic 6: Percentron and SVM

- Percentron classifier
- Percentron learning rule
  - Training for each instance
- Multilayered percentron doesn’t help without non-linearity
- Intuition behind SVM (margin)
- Linear versus kernel-based SVM

Topic 7: Decision Trees

- Inductive Learning Hypothesis
    - IID assumption
- Understand two sources of inductive bias
    - Language bias
    - Search bias
- Impossibility of bias-free learning
- How to build a decision tree
- Calculating entropy, information gain
- Overfitting

Topic 8: Neural Network (1)

- Types of neurons
  - Linear, binary threshold, rectified Linear, sigmoid (remember)
- The need for hidden layers
  - Without them, limited in the model space
  - Hidden layers learn features
- Backpropagation
  - Compute gradients (partial derivatives) of error function relative to each weight

Neural Network (2)

- Definition of softmax, cross-entropy
- Convolutional neural networks
  - Why we need them? What other things we can do if not using CNN?
  - Replicating feature recognizer
- Comparing different models
  - Is 30 errors significantly better than 40 errors?

Neural Network (3)

- Ways of dealing of overfitting
  - Weight-decay, Weight-sharing, Early stopping
  - Model averaging, Dropout
  - Creating new training data
- Ways to speed up mini-batch learning
  - Momentum, separate adaptive learning rate,

Map-Reduce

- Challenges of cluster computing:
  - Node failures, network bottle-neck, programming
- Meeting the challenges
  - Redundant storage of files, moving jobs to where data is, Map-reduce framework
- Steps involved in Map-reduce framework.
- How to combine Map and reduce to solve problems.
- How the map-reduce framework deal with failures: map worker, reducer, master?

PageRank

- How to compute pagerank for simple examples by power iteration method.
- Random walk interpretation
- Dead ends and spider traps
- How dead ends and spider traps are handled?

Spark

- Dataframes
- Concepts of transformations and actions
- Why it is faster than map-reduce

Adversarial Machine Learning

- What are adversarial examples?
- Not just for Neural Networks
- Relationship to linearity in input
- What do the different maps of Adversarial and Random Cross-Sections mean?
- Concept of transferability