Security Analytics
Review for Final Exam

Purdue University

Prof. Ninghui Li

Exam Date/Time

- Monday Dec 10 (8am – 10am)
- LWSN B134

Organization of the Course

- Basic machine learning algorithms
- Neural networks
- Big data analytics
- Advasarial machine learning

Topic 2

- Tasks: Exploratory, Descriptive, Predictive, Pattern Discovery
- What are the differences between supervised learning and unsupervised learning?

Topic 2

- Concepts of
  - **Model space**
  - **Scoring function**
  - **Search technique**
- Distance metrics
  - Minkowski: Manhattan, Euclidean, L_0, L_\infty
  - Jaccard

Topic 2

- 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
- Under Conditional Independence
- Able to compute Bernoulli and Binomial

Topic 5: CLassification

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

Topic 6: Logistic Regression and SVM

- Linear regression
- Sum-square Error (SSE)
- Logistic-regression
  - Intuition, Odds-Ratio,
- Maximum likelihood estimation
- 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, Gini impurity
- Overfitting, prepruning, postpruning (reduced error pruning)

Topic 8: Bagging and Random Forest

- **Bagging: B**ootstrap **agg**regat**ing**
- Bootstrap sampling
- Limitations of bagging with decision trees (i.d. not i.i.d.)
- Random forests
  - Need for feature selection
  - Increasing number of trees causes no overfitting

Topic 8: Neural Network (1)

- Types of neurons
  - Linear, binary threshold, rectified Linear, sigmoid (remember)

Neural Network (2)

- Architecture of NN
  - Feed-forward, recurrent
- Percentron classifier
- Percentron learning rule
  - Training for each instance
- Multilayered percentron doesn’t help without non-linearity
- The need for hidden layers
  - Without them, limited in the model space
  - Hidden layers learn features

Neural Network (3)

- Backpropagation
  - Compute gradients (partial derivatives) of error function relative to each weight
- Online, full batch, and mini-batch

Neural Network (4)

- Definition of softmax,
- Definition of cross-entropy

Neural Network (5)

- Convolutional neural networks
  - Why we need them? What other things we can do if not using CNN?
  - Replicating feature recognizer

Neural Network (6)

- Ways to speed up mini-batch learning
  - Momentum, separate adaptive learning rate, rprop, rmsprop

Neural Network (9)

- Ways of dealing of overfitting
  - Weight-decay, Weight-sharing, Early stopping
  - Model averaging, Dropout
  - Creating new training data

Recurrent Neural Networks

- Types of input-output
- Understand issue of Vanishing gradients
- Gated recurrent units
- LSTM

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?

Spark

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

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?

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