JOHNS HOPKINS
CAREY BUSINESS SCHOOL

Final Review

BU.330.775 Machine Learning

Minghong Xu, PhD.
Associate Professor

Course Evaluation Time!

Final Exam Information

- Length: 1.5 hours
- Format: in-person, closed-book
- Cheat sheet:
  - One A4 paper, single-side only
  - Sign and submit after exam
- Question Types:
  - 10 multiple-choices (10 pt)
  - 8 fill-in-blanks (8 pt)
  - 3 essay questions (17 pt)

Software

- You will use Respondus Lockdown Browser for the final

https://www.respondus.com/lockdown/download.php?id=123533816

- You will not be able to access the exam if you try to access it via any other browser (Chrome, Firefox, Safari, etc.)

- Sample test on Canvas

Final Week Office Hours

- Wednesday Dec 11th, 11:30am-1:30pm
- **Saturday Dec 14th, 5pm-6pm**
- **Sunday Dec 15th, 5pm-6pm**
- On Zoom link:

https://jhucarey.zoom.us/j/4658557490?pwd=Y2NvL0M0RjdFb3RpUjlVOFBSSkFLZz09

- **Monday Dec 16th, 1pm-2pm**
- **At HBC 458C**

Scope

- All class materials we covered including
  - Lecture slides (except those having “Optional” in title)
  - Labs & assignments
- Focus on understanding and application
- Python code
  - Only core code
  - Understanding only, in multiple choices
  - No plot, no matplotlib

Introduction

- AI paradigm
  - AI vs machine learning vs deep learning vs generative AI
- Machine learning definition and use cases
- Three categories of machine learning
  - Supervised learning
  - Unsupervised learning
  - Reinforcement learning
- Machine learning problem formulation

Data Preparation and Preprocessing

- Observation/instance
- Representation and feature: continuous vs categorial
- Target/output/label
- Training, testing, validation and cross-validation
- Sampling strategies: stratifying and shuffling
- Encode categorial data: ordinal vs one-hot
- Missing value: remove vs imputation; zero, mean, median
- Outliers: issues and how to identify, drop or keep
- Feature scaling: normalization vs standardization

Supervised I

- Classification vs regression
- Parameters and hyperparameters
- Model optimization, loss function
- Gradient descent
  - Stochastic, batch, vs mini-batch
- Learning rate, general idea of adaptive learning rates
- Training epoch
- Logistic regression is classification

Regularization

- Overfitting vs appropriate-fitting vs under-fitting
- Bias and variance, generalization error
- L2 regularization, ridge regression
- L1 regularization, Lasso regression
- L1 vs L2
- Early stopping

Model Evaluation

- H0 and H1
- Type I and Type II errors, why
- Confusion matrix
- Performance measures: accuracy, precision, recall, F1-score, specificity
- Precision-recall trade-off
- ROC curve and AUC

Supervised II

- K-nearest neighbors and use cases
- Pros and cons
- Decision tree (just the idea), strengths and weaknesses
- Ensemble
  - Bagging not required
- Boosting
  - Adaptive boosting vs gradient boosting

Unsupervised I

- Training in unsupervised
- Testing: alterative evaluations
  - Internal vs external vs generalization
- Curse of dimensionality
- Reduce dimensions, information loss
- Project methods and issues
- PCA, process and steps, explained variance ratio
- Manifold methods and t-SNE
- Feature engineering: extraction vs selection vs creation

Clustering

- Definition and similarity
- Clustering vs classification
- Use cases
- K-means and steps, centroid, inertia
- Hard clustering vs soft clustering
- Mini-batch k-means
- Issues of clustering

Reinforcement Learning

- Bellman equation
- Definition and objective
- environment, actions, rewards
- Policy, policy parameters, policy gradient
- Why reinforcement learning

JOHNS HOPKINS
CAREY BUSINESS SCHOOL

Q & A