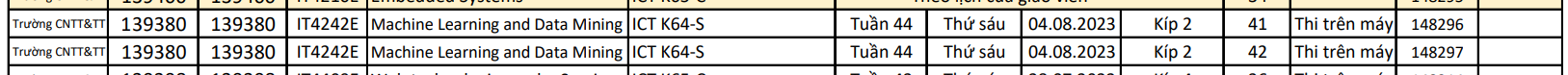
139380 Last lecture discussion

7/2023

# Final exam information

* Computer-based exam
* Multi-choice questions (one or more answer options)
* Do not bring documents, may bring calculator



# Introduction

Concepts

* Machine learning – Data mining
* Model – Leaning function
* Supervised learning – Unsupervised learning
* Design a learning system
* *Underfitting – Overfitting*
* Training – Verification – Testing
* Generalization – Regularization – No-free-lunch theorem

Examples

* Machine learning vs. data mining, techniques of ML/DM
* Underfitting/overfitting vs. prediction accuracy of training/test sets.
* Underfitting/overfitting vs. generalization/regularization
* Data collection/understanding/preparation in ML/DM

# Data preprocessing

Concept

* Data samples – Attribute – Feature – Value
* Data collection – Data crawling – Logging – Scraping
* Data sampling – Variety – Bias
* Data quality – Completeness – Integrity – Homogeneity – Structures
* Cleaning – Integrating – Transforming
* Standardization – Discretization – Normalization
* Data reduction - Feature selection – Dimensional reduction – Abstraction

Examples

* Compare data crawling, logging, scraping
* Methods in data leaning, integrating, transforming
* Purpose of standardization, discretization, normalization

# Linear regression

Concepts

* Linear model – Coefficient vector
* Expected loss – Empirical loss – Generalization error
* OLS – Ridge regression – LASSO – regularization/penalty term (𝜆)

Examples

* OLS vs. Rigde vs. LASSO
* Calculation of coefficient vector for some data samples
* Penalty term 𝜆 vs. underfitting/overfitting, dimension vs size of data

# K-means

Concepts

* Clustering – K-mean process/request/advantages/limitations
* Centroid – Convergence – Loss function
* Problem of outliers, initialization, K-mean++
* Online K-means, algorithm, learning rate, forgetting rate

Examples

* Convergence -> conditions of convergence; centroid -> type of data used
* Calculation of K-means process with small number of samples
* Performance vs size of data, dimensions; lost vs K

# k-NN

Concepts

* Names relating to learning
* Prediction for classification/regression
* Choosing K
* Distances with different data types
* Attribute normalization – Weighting attribute – Weighting neighbors

Examples

* Leaning vs. testing phases
* K vs. M, underfitting/overfitting
* Reason/way of normalization/weighting

# ID3 and RF

Concepts

* Tree representation, ID3 process
* Information gain – Entropy
* Stop learning early – Prune tree
* Attribute selection – Real data – Missing value
* Randomization – Combination – Bagging

Examples

* ID3 vs. DT (regression), DT vs RF (overfitting/variance/intepretable)
* Stop learning early vs. prune tree
* Selecting attribute for branching (calculation)

# SVM

Concepts

* Decision boundary – Hyperplane – Margin – Prediction process – Support vectors
* Hinge loss – Problem formulation – KKT conditions
* Soft margin SVM – Non-linear SVM – Penalty constant

Examples

* Meaning of C, margin, support vectors
* Hard magin vs. Soft margin vs Non-linear SVM
* Calculation of support vectors, decision boundary for dataset of some samples

# Model assessment

Concepts

* Hold-out – Stratified – Repeated hold-out – Cross-validation – Bootstrap (sampling techniques)
* Model selection – Model assessment
* Accuracy (classification/regression)
* Confusion matrix (TP, FP, TN, FN) – Precision – Recall – Micro/macro-averaging

Examples

* Sampling techniques vs. datasets of small/big size
* Calculation of measures from confusion matrix
* Meaning of measures for real life situation

# ANN

Concepts

* Neuron – Activation function
* Architecture – Input – Hiddle – Output layers
* Perceptron – Perceptron algorithm
* Back propagation algorithm – initialization – Learning rate

Examples

* Hard-limited vs. Sigmoid vs. ReLU activation function
* Perceptron vs. back propagation algorithm
* Calculate number of parameters of an ANN

# Probability models

Concepts

* Probability – statistically independent – Bayes’ rule
* Prior/posterior probability – Likelihood – Hypothesis
* Gaussian distribution – GMM
* MLE – MAP – Gaussian/Multinomial Naïve Bayes

Examples

* Calculation on Gaussian distribution, GMM, probability
* MLE vs. MAP
* Features of MLE, MAP, Naïve Bayes

# Association rules

Concepts

* Item – Itemset – k-itemset – Basket – Association rule
* Support – Confidence – Support count – 2-itemset counting
* Monotonicity property – A-priori algorithm – Generate/count/prune process

Examples

* Problem of Association rule mining
* Processes of generate/count/prune of A-priori algorithm.

# All techiniques

* Concepts, ML task (classification/regression/clustering)
* Idea, problem formulation
* Learning vs prediction (process, performance)
* Characteristics of parameters, data (numerical/categorical), leaning (underfitting/overfitting)
* Advantages, limitation