Machine Learning
(IT3190E)

Quang Nhat NGUYEN

quang.nguyennhat@hust.edu.vn

Hanoi University of Science and Technology
School of Information and Communication Technology

Academic year 2020-2021

Proposal of course project

- Select one of the suggested course project examples, *or*
- Modify based on one of the suggested course project examples, *or*
- Propose a new course project
  - *A real application problem that is appropriate to be solved by machine learning*!

Email spam filtering

- **Description of the problem**. To identify (classify) spam emails
- **Input**. Representation of the content of an email (e.g., a vector of keyword weights)
- **Output**. An assigned label of either “spam” or “normal”
- **Approach**. Naïve Bayes classification
- **Dataset**. A set of examples, where each one consists of two parts: email content representation and its label (“spam” or “normal”)

Web page categorization

- **Description of the problem**. For a set of Web pages, the system needs to assign (classify) each Web page into a category (e.g., “Business”, “Sport”, “Technology”, etc.)
- **Input**. The content representation of a Web page (e.g., a vector of keyword weights)
- **Output**. An assigned category of that Web page
- **Approach**. Naïve Bayes classification, or Artificial neural network
- **Dataset**. A set of examples, where each one is represented by the Web page representation and its category

Clustering of student study results

- **Description of the problem**. The system needs to cluster (group) the students based on a number of predefined attributes (e.g., the semester's average grade, gender, the number of registered courses for the semester, the percentage of the lectures participation, etc.)
- **Input**. A vector of attribute values that represent for a student
- **Output**. Clusters of the students' study results
- **Approach**. K-means clustering
- **Dataset**. A set of examples, where each one is a vector of attribute values that represent for a student

Prediction of the level of risk of a loan application

- **Description of the problem**. Given a financial loan application, the system needs to predict (classify) the level of risk of that loan application – in order to decide whether to accept or reject the loan request
- **Input**. The representation of a loan application (e.g., a vector of attribute values)
- **Output**. A predicted level of risk (e.g., “low” – to accept, or “high” – to reject)
- **Approach**. Decision tree classification, or Naive Bayes classification
- **Dataset**. A set of examples, where each one consists of 2 parts: the representation of a loan application and its level of risk (i.e., “low” or “high”)

Web pages recommendation

- **Description of the problem**. Given a set of Web pages that a user has viewed, the system needs to identify (predict) those unseen Web pages of that user's interest. Assumption: If any 2 users who viewed the same Web pages, then they will like to view the same unseen Web pages in future
- **Input**. A list of Web pages viewed by a user (i.e., a Web page is represented by an ID, and not exploiting the Web page's content)
- **Output**. A small and selected set of unseen Web pages recommended to that user
- **Approach**. Nearest neighbour learning, Collaborative filtering
- **Dataset**. A set of examples, where each one consists of the identity (ID) of a user and a list of identities (IDs) of the Web pages that have been viewed (or their ratings) by that user

Experimental comparison of ML algorithms

- **Decsription of the problem**. A real application that can be solved by machine learning (e.g., such one mentioned in the previous slides)
- **Dataset**. A dataset suitable for the selected application problem
- **Tasks:**
  - Select some (2-3) appropriate machine learning algorithms that are suitable for solving the selected application problem
  - For each of the selected algorithms, implement the corresponding system variant to solve the application problem
  - Run the experiments to compare the performance of the system variants on the selected dataset
  - For example, you may compare the performance of the Naive Bayes and the Decision tree classification approaches for the problem of estimation of the risk level of loan application