AIT 636: INTERPRETABLE MACHINE LEARNING

Course Syllabus

Instructor: Dr. Mahdi Hashemi
Email: mhashemi@gmu.edu
Office hours: Monday 2-4 PM

Description

One of the most common tasks performed by data scientists and data analysts is prediction and machine learning. Machine learning combines advanced topics in statistics, probabilities, linear algebra, and calculus to design mathematical models that learn from data or experience to solve new problems. Computers usually do not explain their predictions which is a barrier to the adoption of machine learning. This course focuses on making the decisions from algorithms more understandable for humans. In other words, making machine learning models and their decisions interpretable. This course covers simple, interpretable models such as decision trees, decision rules and linear regression. It also covers general model-agnostic methods for interpreting black box models like feature importance and model settings.

Requirements

A basic understanding of statistics and familiarity with Python is required.

Objectives

On successful completion of this course, students will be able to:

- implement and evaluate major machine learning models,
- interpret and critically discuss different machine learning models,
- understand how prediction models work under the hood,
- understand their strengths and weaknesses and interpret their outputs,
- select and correctly apply the methods that are most suitable for a project,
- overcome the challenges of working with real-world datasets,
- understand basic topics in machine learning, including classification, regression, clustering, feature selection and generation.

Textbook


References

**Grading**

<table>
<thead>
<tr>
<th>Grade</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>97-100</td>
<td>A+</td>
</tr>
<tr>
<td>93-96</td>
<td>A</td>
</tr>
<tr>
<td>90-92</td>
<td>A-</td>
</tr>
<tr>
<td>87-89</td>
<td>B+</td>
</tr>
<tr>
<td>83-86</td>
<td>B</td>
</tr>
<tr>
<td>77-82</td>
<td>B-</td>
</tr>
<tr>
<td>70-76</td>
<td>C</td>
</tr>
<tr>
<td>0-69</td>
<td>F</td>
</tr>
</tbody>
</table>

- Quiz given at the second session (open-book): 5%
- Five homeworks: 5%
- Midterm Exam: 25%
- Final Exam: 25%
- Python implementation: 10%
- Surveying essay on a topic related to interpretable machine learning: 15%
- Project: 15%

**Topics**

- Introduction to interpretability and machine learning
  - Simple interpretable models
    - Bayesian classifier
      - How the likelihood and priors affect the model’s output
      - Interpreting parametric vs. nonparametric models
      - Explaining the effect of hyperparameters
    - Is naïve Bayes classifier a simply interpretable model?
    - Interpretation of Bayesian belief networks
  - How to interpret content-dependent classification
  - Linear models
    - Perceptron
    - SVM
    - Least squares
    - Logistic regression
  - Decision trees
    - How to interpret a decision tree
    - What is a decision
    - Rules and their consequences
    - Interpreting feature importance from decision tree
- How to apply machine learning models correctly
  - Accuracy metrics
  - Hyperparameters
  - Different evaluation methods
    - Setting the value of hyper-parameters
- Explaining and understanding the data before applying machine learning
  - Model-agnostic interpretation methods for features
    - Regression-related methods
      - Significance of coefficients
      - Best-subset selection
      - Forward-stepwise selection
      - Backward-stepwise selection
      - Hybrid-stepwise selection
      - Forward-stagewise regression
      - Incremental forward-stagewise regression
      - Shrinking the coefficients
    - Classification-related methods
      - Class separability measures
      - Scalar feature selection
      - Scalar feature selection incorporating correlation information
      - Wrapper approach
      - Filter approach
      - Sequential backward selection
    - Sequential forward selection
    - Floating search method
    - Fisher’s linear discriminant analysis
  - Interpreting clustering approaches and clusters
    - Proximity measures, their interpretation and consequences
      - Proximity measures between two points
      - Proximity functions between a point and a set(cluster)
      - Proximity functions between two sets/clusters
    - Sequential algorithms
    - Hierarchical algorithms
  - Schemes based on function optimization
    - Proper application of clustering in real-world problems
      - Why evaluating clusters?
      - Monte Carlo technique
      - Internal criteria

<table>
<thead>
<tr>
<th>Topic</th>
<th>Week</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explaining and understanding the data before applying machine learning</td>
<td></td>
</tr>
<tr>
<td>Model-agnostic interpretation methods for features</td>
<td></td>
</tr>
<tr>
<td>Regression-related methods</td>
<td></td>
</tr>
<tr>
<td>Classification-related methods</td>
<td></td>
</tr>
<tr>
<td>Interpreting clustering approaches and clusters</td>
<td></td>
</tr>
<tr>
<td>Sequential algorithms</td>
<td></td>
</tr>
<tr>
<td>Hierarchical algorithms</td>
<td></td>
</tr>
<tr>
<td>Schemes based on function optimization</td>
<td></td>
</tr>
<tr>
<td>Proper application of clustering in real-world problems</td>
<td></td>
</tr>
<tr>
<td>Final Exam</td>
<td></td>
</tr>
</tbody>
</table>

Week 8

Week 9

Week 10

Week 11

Week 12

Week 13

Week 14

Week 15

Week 16
Important Dates

Dates for dropping, adding the course etc. are available via http://registrar.gmu.edu/calendars/

Religious Holidays

A list of religious holidays is available on the University Life Calendar page. Any student whose religious observance conflicts with a scheduled course activity must contact the Instructor at least 2 weeks in advance of the conflict date in order to make alternative arrangements.

Attendance Policy

Students in in-class sections are expected to attend each class, to complete any required preparatory work and to participate actively in lectures, discussions and exercises. As members of the academic community, all students are expected to contribute regardless of their proficiency with the subject matter.

Students are expected to make prior arrangements with Instructor if they know in advance that they will miss any class and to consult with the Instructor as soon as possible if they miss any class without prior notice. Any student who expects to miss more than one class session is strongly advised to drop the course and take it in a later semester when he/she can attend every class.

Departmental policy requires students to take exams at the scheduled time and place, unless there are truly compelling circumstances supported by appropriate documentation. Except in such circumstances, failure to attend a scheduled exam will result in a score of zero (0) for that exam, in accordance with Mason policy on final exams. Students should not make travel plans or other discretionary arrangements that conflict with scheduled classes and/or exams. If the University is closed due to weather or other unforeseen conditions, final exams may be rescheduled – students are strongly advised not to make plans that would prevent them from attending exams that may be rescheduled during the entire exam period.

Classroom conduct

Students are expected to conduct themselves in a manner that is conducive to learning, as directed by the Instructor. Any student who negatively impacts the opportunity for other students to learn will be warned – if disruptive behavior continues, the student will be asked to leave the classroom.

Electronic devices are potential distractions in the classroom environment. Cell phones, pagers and other handheld devices must be turned off or set to "silent" mode and not used while class is in session. Laptop computers and similar devices may be used only if such use is directly related to the classroom activity in progress – for some activities the Instructor may require that such devices not be used in order to maximize student engagement.
Communications

Registered students will be given access to a section of the Blackboard Learning System for this course. Blackboard will be used as the primary mechanism (outside of lectures) to disseminate course information, including announcements, lecture slides, homework and other assignments, and scores for homework and exams.

Communication with the Instructor on issues relating to the individual student should be conducted using Mason email, via telephone, or in person - not in the public forums on Blackboard. Mason Mail is the preferred method – for urgent messages, you should also attempt to contact the Instructor via telephone. Federal privacy law and Mason policy require that any communication with a student related in any way to a student's status be conducted using secure Mason systems – if you use email to communicate with the Instructor you MUST send messages from your Mason email account.

When sending an e-mail to the instructor, please include the following:

- Course number
- Your full name

Privacy

Instructors respect and protect the privacy of information related to individual students.

As described above, issues relating to an individual student will discussed via email, telephone or in person. Instructors will not discuss issues relating to an individual student with other students (or anyone without a need to know) without prior permission of the student.

Assessable work other than final exams will be returned to individual students directly by the Instructor (or by a faculty or staff member or a Teaching Assistant designated by the Instructor, or via another secure method). Under no circumstances will a student's graded work be returned to another student.

Faculty and staff will take care to protect the privacy of each student's scores and grades.

Disability Accommodations

The Office of Disability Services (ODS) works with disabled students to arrange for appropriate accommodations to ensure equal access to university services. Any student with a disability of any kind is strongly encouraged to register with ODS as soon as possible and take advantage of the services offered.

Accommodations for disabled students must be made in advance – ODS cannot assist students retroactively, and at least one week's notice is required for special accommodations related to exams. Any student who needs accommodation should contact the Instructor during the first week of the semester so the sufficient time is allowed to make arrangements.

Honor Code
All members of the Mason community are expected to uphold the principles of scholarly ethics. Similarly, graduating students are bound by the ethical requirements of the professional communities they join. The ethics requirements for some of the communities relevant to IST graduates are available via the following links:

- ACM Code of Ethics and Professional Conduct
- IEEE Code of Ethics
- EC-Council Code of Ethics

On admission to George Mason University, students agree to comply with the requirements of the Mason Honor System and Code. The Honor Code will be strictly enforced in this course. Honor Code cases are heard by a panel consisting of students – students who meet the requirements are encouraged to nominate themselves to serve on the Honor Committee.

Any use of the words or ideas of another person(s), without explicit attribution that clearly identifies the material used and its source in an appropriate manner, is plagiarism and will not be tolerated. The Instructor reserves the right to use manual and/or automated means (including such services as SafeAssign.com) to detect plagiarism in any work submitted by students for this course, and to direct Teaching Assistants and/or other faculty and/or staff members to do likewise in support of this course. For this course, the following requirements are specified:

- All assessable work is to be prepared by the individual student, unless the Instructor explicitly directs otherwise.
- All work must be newly created by the individual student for this course for this semester.

Any usage of work developed for another course, or for this course in a prior semester, is strictly prohibited without prior approval from the instructor.

Students may seek assistance with assigned work (and are encouraged to do so if they feel the need), provided:

- Such assistance is acknowledged in the submitted work, clearly identifying the person(s) giving assistance and the nature of the assistance given.
- Any work to be submitted is prepared entirely an exclusively by the student submitting it. Students are expressly prohibited from sharing any assessable work for this course in any manner with other students (except students assigned as Teaching Assistants to this course and the student's section), unless all students involved have had their work graded and returned by the Instructor, or the Instructor has explicitly approved such sharing.

**Counseling and Psychological Services**

Phone: 703-993-2380
https://caps.gmu.edu/

**Title IX Coordinator**
Gender Identity and Pronoun Use

Gender identity and pronoun use: If you wish, please share your name and gender pronouns with me and how best to address you in class and via email. I use “he” for myself and you may address me as “Dr./Prof. Hashemi” in email and verbally.

Anti-Racism Statement

As a member of the George Mason University community, the Department of Information Sciences and Technology plays an integral role in building an educational environment that is committed to anti-racism and inclusive excellence. An anti-racist approach to higher education acknowledges the ways that individual, interpersonal, institutional, and structural manifestations of racism against Black individuals and other people of color contribute to inequality and injustice in our classrooms, on our campuses, and in our communities, and it strives to provide our community members with resources to interrupt cycles of racism so as to cultivate a more equitable, inclusive, and just environment for all of our students, staff, faculty, alumni, and friends, regardless of racial background.

To be anti-racist means:

- To make constant, conscious decisions to interrupt racism and cultivate equity, inclusion, and justice for people of all racial backgrounds, and in particular those from Black communities and other communities of color, who are most likely to bear the direct and indirect costs of systems of white supremacy;
- To interrogate histories of white supremacy and white-dominant culture, and to examine the ways in which these histories have impacted our individual beliefs, our interpersonal relationships, our institutional and structural policies and processes, and our entire society;
- To make a commitment to being responsible for our own relationships to, and actions within, systems of white supremacy; and
- To cultivate a practice of self-awareness and self-reflection that allows us to critically evaluate our own role in upholding white supremacy and identify the ways we can interrupt cycles of racism at the individual, interpersonal, institutional, and structural levels.

We believe that the work of anti-racism starts with each individual, and that in cultivating an anti-racist approach to research, scholarship, and practice, our students will build a skillset rooted in principles of equity, inclusion, and justice that they will carry with them throughout their lives.
For more information on how to continuously cultivate the practice of anti-racism, see this guide from the National Museum of African American History and Culture on how to be anti-racist:
https://nmaahc.si.edu/learn/talking-about-race/topics/being-antiracist

[This antiracism statement was prepared by Dr. Charles Chavis, Assistant Professor in the Jimmy and Rosalyn Carter School of Peace and Conflict Resolution]