CIS 732: Machine Learning and Pattern Recognition
Spring 2008

Hours: 3 hours (additional 3-hour proseminar in data mining, CIS798, available)
Prerequisite: CIS 300 and 501 or equivalent coursework in data structures and algorithms; CIS 301 (set theory/logic), Math 510 (discrete math), Stat 410 (intro probability) recommended
Textbook: *Data Mining: Concepts and Techniques, 2nd edition*, Han & Kamber
Time and Venue: Mon, Wed, Fri 14:30 - 15:20, Room 233 Nichols Hall (N233, CIS Library)
Instructor: William H. Hsu, Department of Computing and Information Sciences
Office: 213 Nichols Hall URL: http://www.cis.ksu.edu/~bhsu (calendar posted here)
Office: +1 785 532 7905 Home: +1 785 539 7180 E-mail: CIS732TA-L@listserv.ksu.edu
Office hours: 12:30 – 14:00 Mon, Wed, Fri
by appointment, Tuesday, Friday AM
Class web page: http://www.kddresearch.org/Courses/Spring-2008/CIS732/

Course Description

This is an introductory course in machine learning for development of intelligent knowledge based systems. The first half of the course will focus on basic taxonomies and theories of learning, algorithms for concept learning, statistical learning, knowledge representation, pattern recognition, and reasoning under uncertainty. The second half of the course will survey fundamental topics in combining multiple models, learning for plan generation, decision support, knowledge discovery and data mining, control and optimization, and learning to reason.

Course Requirements

Exams (35%): in-class midterm exam (15%), take-home final (20%)
Homework (30%): 6 out of 8 programming and written assignments (5% each: 2 written, 2 programming, 2 mixed)
Project (20%): term programming project and report for all students
Paper Reviews (10%): 10 of 12 weekly or semi-weekly paper reviews (1% each)
Class Participation (5%): class and online discussions, asking and answering questions

Computer language(s): C/C++, Java, or student choice (upon instructor approval)

Selected reading (on reserve in K-State CIS Library)

- Koza, J. R., Keane, M. A., Streeter, M. J., Mydlowec, W., Yu, J., & Lanza, G. (2005). *Genetic Programming IV: Routine Human-Competitive Machine Intelligence*. New York, NY: Springer.
- Alpaydin. E. (2004). *Introduction to Machine Learning.* Cambridge, MA: The MIT Press.
- Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation, 2nd edition*. Englewood Cliffs, NJ: Prentice-Hall.
- Mitchell, T. M. (1997). *Machine Learning.* New York, NY: McGraw-Hill.
- Buchanan, B. G. & Wilkins, D. C., eds. (1993). *Readings in Computer Inference and Knowledge Acquisition*. San Francisco, CA: Morgan Kaufmann.
- Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. London, UK: Oxford University Press.
- Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning.* Reading, MA: Addison-Wesley.
- Koza, J. (1992). *Genetic Programming: On The Programming of Computers by Means of Natural Selection*. Cambridge, MA: The MIT Press.

Course Calendar

Lecture	Date	Topic	Source: Han & Kamber 2e
0	Fri 18 Jan 2008	Administrivia; overview of learning	6.1, Handout 1**:** Ch. 1 Mitchell
1	Wed 23 Jan 2008	Concept learning, version spaces	6.2, Ch. 2 Mitchell
2	Fri 25 Jan 2008	Version spaces continued	6.2
3	Mon 28 Jan 2008	Decision trees	6.3
4	Wed 30 Jan 2008	Decision trees continued	6.3, Handout 2
5	Fri 01 Feb 2008	Overfitting and Occam's Razor	6.3
6	Mon 04 Feb 2008	Bayesian classification	6.4
7	Wed 06 Feb 2008	Naïve Bayes	6.4.2
8	Fri 08 Feb 2008	Rule-based classification	6.5
9	Mon 11 Feb 2008	Perceptrons and Winnow	6.6, Handout 3
10	Wed 13 Feb 2008	Classification using MLPs	6.6
11	Fri 15 Feb 2008	Support Vector Machines (SVM)	6.7
12	Mon 18 Feb 2008	Associative Classification	6.8
13	Wed 20 Feb 2008	Lazy Learning: (k-)Nearest-Neighbor	6.9
14	Fri 22 Feb 2008	Genetic Algorithms (GAs)	6.10
15	Mon 25 Feb 2008	Prediction by regression	6.11
16	Wed 27 Feb 2008	Classification/prediction accuracy	6.12
17	Fri 29 Feb 2008	Statistical Evaluation of Hypotheses	6.13
18	Mon 03 Mar 2008	Ensembles: Bagging and boosting	6.14
18	Wed 05 Mar 2008	Model selection: confidence, ROC	6.15
20	**Fri 07 Mar 2008**	**Clustering Intro; midterm review**	**7.1 – 7.3**
21	Mon 10 Mar 2008	Partitioning-based clustering	7.4
	Wed 12 Mar 2008	**No class: CIS 732 midterm**	**Chapters 1, 6**
22	**Fri 14 Mar 2008**	**Agglomerative clustering**	**7.5**
23	Mon 24 Mar 2008	Density- and grid-based clustering	7.6 – 7.7
24	Wed 26 Mar 2008	Model-based clustering: EM	7.8.1 – 7.8.2
25	Fri 28 Mar 2008	Model-based clustering: ANN	7.8.3
26	Mon 31 Mar 2008	Clustering and outliers, concepts	7.9 – 7.11, 2.6
27	Wed 02 Apr 2008	Time series data and data streams	8.1 – 8.2
28	Fri 04 Apr 2008	Sequence patterns	8.3 – 8.4
29	Mon 07 Apr 2007	Genetic programming 1	Koza video 1, Handout 4
30	Wed 09 Apr 2008	Genetic programming 2	Koza video 3, Handout 4
31	Fri 11 Apr 2008	Symbolic regression and GP	Handout 4: Koza
32	Mon 14 Apr 2008	Computational learning theory 1: VC	Handout 5: Kearns & Vazirani
33	Wed 16 Apr 2008	Computational learning theory 2: PAC	Handout 5: Ch. 7 Mitchell
34	Fri 18 Apr 2008	Computational learning theory 3	Handout 5: SVM & COLT
35	Mon 21 Apr 2008	Data mining and KDD overview	1.1 – 1.10
36	Wed 23 Apr 2008	Data preparation and cleaning	2.1 – 2.3
37	Fri 25 Apr 2008	Data reduction and discretization	2.5 – 2.6
38	Mon 28 Apr 2008	Frequent pattern mining: apriori	5.1 – 5.2
39	Wed 30 Apr 2008	Frequent pattern mining continued	5.2
40	Fri 02 May 2008	Graph mining, social network analysis	9.1 – 9.2
41	**Mon 05 May 2008**	**Inductive logic programming (ILP)**	**9.3**
42	Wed 07 May 2008	ILP and multirelational data mining	9.3
	Fri 09 May 2008	Project presentations	
		Final Exam: due Mon 12 May 2008	**Chapters 1, 2, 5-9**

Green-shaded entries denote the due date of a paper review.
Lightly-shaded entries denote the due date of a written problem set.
Heavily-shaded entries denote the due date of a machine problem (programming assignment)

Interim project interviews will be held between the midterm and spring break.
The blue-shaded date is the due date of the draft project report and demo, with interviews and presentations to be held the last two days of class.

Green, **blue** and **red** letters denote **exam review**, **exam**, and **exam solution review** dates.