EOSC 510 Winter session 2013-2014 (term 1) 3 credits

Course title: Data Analysis in Atmospheric, Earth & Ocean Sciences

This is an online graduate course on applying machine learning and statistical methods to environmental sciences.

Course web site: http://www.ocgy.ubc.ca/~william/EOSC510/

Instructor: William Hsieh e-mail: whsieh 'at' eos.ubc.ca tel: 604-822-2821
Office: EOS-South Bldg. room 162

Course description:
Regression and spectral analysis, multivariate statistical methods (principal component and canonical correlation analysis), classification and clustering methods, neural networks, tree-based models and kernel methods.

Course outline: [Note: each "hour" is about 40 minutes of video lecture material, not counting the time needed by the student to solve questions posed by the instructor during the lecture.]

0. Introduction and course setup (0.5 hr.)

1. Correlation and regression (2.5 hrs.).

2. Principal component analysis (PCA) and rotated PCA (7.5 hrs.)

3. Canonical correlation analysis (CCA) (2 hrs.)

4. Time series: Fourier spectral analysis, windows, filters, and singular spectrum analysis (4 hrs.).

5. Classification and clustering (3.5 hrs.)

6. Feed-forward Neural Network (NN) models: multi-layer perceptrons (MLP), classification, radial basis functions (RBF), conditional density networks (CDN), mixture models (5.5 hrs.).

7. Nonlinear optimization methods: Deterministic and stochastic approaches (2 hrs.)

8. Learning and generalization: variance & bias errors, regularization, cross-validation, (Bayesian NN), ensemble methods (bootstrapping, boosting), effects of time averaging, regularization of linear models (ridge regression and lasso) (4.5 hrs.)

9. Tree-based methods: classification and regression trees (CART), random forests (1.5 hrs).

10. Forecast verification (1.5 hrs)

11. Nonlinear principal component analysis (1 hr)

12. Kernel methods (1 hr) [optional material]

Marking scheme:

Homework (see www.ocgy.ubc.ca/~william/EOSC510)	40%
1 Mid-term test	20%
1 Final exam	40%

Rationale:

Environmental scientists have to work with large datasets in the space-time domain. A graduate-level course covering classical multivariate statistical methods and modern machine learning methods (e.g. neural networks) is deemed to be a course which would enhance our graduate training. In the assignments, students will analyze real datasets from meteorology and oceanography. This course should also be of interest to graduate students of other disciplines.

Pre-requisites: This course assumes the student knows linear algebra (e.g. eigenvectors), vector calculus and some basic probability.

Textbook:

Hsieh, W.W., 2009. Machine Learning Methods in the Environmental Sciences. Cambridge Univ. Pr., 349 pp.

Reference books:

Hastie, T., R. Tibshirani, and J. Friedman, 2009. Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer.
Free pdf: http://www-stat.stanford.edu/~tibs/ElemStatLearn/

Bishop, C.M. 2006. Pattern Recognition and Machine Learning. Springer.
Free pdf:
http://www.inf.unideb.hu/valseg/dolgozok/ispany/DataMine/Alkalmazasok/Bishop__Pattern_Recognition_and_Machine_Learning.pdf

von Storch, H. & F. Zwiers, 1999. Statistical Analysis in Climate Research. Cambridge Univ. Pr.

Wilks, D.S. 2011 (3rd ed.). Statistical Methods in the Atmospheric Sciences. Academic Pr.

Bishop, C.M. 1995. Neural networks for Pattern Recognition. Clarendon Pr.

Cherkassky, V., and F. Mulier, 2007. Learning from Data. Wiley.

Draper, N. and H. Smith, 1981. Applied Regression Analysis, 2nd ed. Wiley-Interscience.

Emery, W.J. & R.E. Thomson. 1997. Data Analysis Methods in Physical Oceanography. Pergamon.

Haupt, S. E., A. Pasini, and C. Marzban, Eds., 2009: Artificial Intelligence Methods in the Environmental Sciences. Springer.

Krasnopolsky, V. M., 2013: The Application of Neural Networks in the Earth System Sciences: Neural Network Emulations for Complex Multidimensional Mappings. Springer.

Mardia, K.V. et al. 1979. Multivariate Analysis. Academic Pr.

Nabney, I.T. 2002. NETLAB: Algorithms for Pattern Recognition. Springer.
Free MATLAB software: http://www1.aston.ac.uk/eas/research/groups/ncrg/resources/netlab/

Scholkopf, B. and A.J. Smola, 2002. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Pr.

Shawe-Taylor, J. and N. Cristianini, 2004. Kernel Methods for Pattern Analysis. Cambridge Univ. Pr.

Witten, I.H., E. Frank and M.A. Hall, 2011. Data Mining: Practical Machine Learning Tools and Techniques. Free software: http://www.cs.waikato.ac.nz/ml/weka/