

Το Τμήμα Μαθηματικών και Στατιστικής του Πανεπιστημίου Κύπρου διοργανώνει σεμινάριο την Τετάρτη 22/02/2017, ώρα 11:00-12:00, στην αίθουσα 037/ΘΕΕ01 στην Πανεπιστημιούπολη.

**Ομιλητής:** Giles Hooker (Department of Biological Statistics and Computational Biology and Department of Statistical Science, Cornell University)

**Τίτλος:** Decision Trees and CLT's: Inference and Machine Learning

**Περίληψη:** This talk develops methods of statistical inference based around the popular machine learning methods of bagging and Random Forests. Our goal is to provide a limiting normal distribution for the prediction made by these methods -- if we re-collected the data and re-trained the model many times, the histogram of the predictions at a particular point would look Gaussian. This result can then be used to provide a formalized statistical test of the relevance of particular input features, or for the structure of the underlying relationship more generally. We show that when the bootstrap procedure in ensemble methods is replaced by sub-sampling, predictions from these methods can be analyzed using the theory of U-statistics. Moreover, the limiting normal distribution has a variance that can be estimated within the sub-sampling structure. Using this result, we can compare the predictions made by a model learned with a feature of interest, to those made by a model learned without it and ask whether the differences between these could have arisen by chance. By evaluating the model at a structured set of points we can also ask whether it differs significantly from an additive model. We demonstrate these results in an application to citizen-science data collected by Cornell's Laboratory of Ornithology.