MA 751
M. Kon

Problem Set 12, Part I
(Spring '22)

These problems do not need to be turned in, but please do go through them, as they are part of the syllabus of the course.

Lectures 23, 24

Random forests are one of the most successful machine learning methods, combining the good qualities of decision tree models together with bootstrap aggregation (bagging), which combines estimates of bootstrap samples. The idea of taking variable subsets of the available features for use at different nodes of trees makes the algorithm more powerful and attractive.

This material also covers SVM methods in more detail, deriving the quadratic optimization algorithm in both the perfectly separable (hard margin) case, and the not linearly separable case (soft margin).

This material also begins to covers prototype methods for classification $-$ these are like k nearest neighbors with the change that only selected 'prototype' points for regions of the feature space are kept, rather than the entire training set.

We will study unsupervised methods such as clustering $-$ k-means clustering versus hierarchical clustering methods, which give very visualizable 'dendrograms' - hierarchical trees which subdivide families of feature vectors in a recursive way. Prior to this an application of clustering is also discussed - there are ways of using clustering not just to divide up the feature space into categories based on similarity, but to do (supervised) classification on this basis.

Reading: 15.1-15.3, 12.1, 12.2, 12.3.1-12.3.4, 13.1, 13.2.1, 13.3 Introduction, 13.3.1-13.3.2, 13.5

Problems: 15.4, 12.1, 12.2, 13.3 (please look at/do these before the final exam)