COURSE DESCRIPTION:
Classification and clustering, also known as supervised and unsupervised machine learning, are important statistical techniques commonly applied in many social and behavioral science research problems. Both seek to understand social phenomena through the identification of naturally occurring homogeneous groupings within a population. Classification techniques are used to sort new observations into pre-existing or known groupings, while clustering techniques sort the population under study into natural, homogeneous groupings based on their observed characteristics. Both help to reveal hidden structure that may be used in further analyses. This course will compare and contrast these techniques, including many of their variations, with an emphasis on applications.

COURSE REQUIREMENTS:
Participation: 10%  You are expected to attend class and participate in class discussions
Homework problems: 20%  There will be several assigned problems intended to give you practical experience with the methods discussed.
Data Analysis Projects: 70%  There will be two data analysis projects (worth 35% each).

COURSE READINGS: Handouts will be available on Blackboard by the Thursday preceding class. It is the student’s responsibility to print out and review the notes before coming to class.

Late assignment policy: Assignments are to be handed in on time.

SCHEDULE FOR JANUARY 2017 (example)

<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan. 4</td>
<td>Introduction to classification and clustering; what is a cluster; visualization techniques, including principal components. The classification technique you already know (logistic regression). Intro to Hierarchical clustering</td>
</tr>
<tr>
<td>6</td>
<td>Hierarchical clustering: linkage choices; distance measures; the dendogram. Optimization techniques (k-means); choosing the number of groups; evaluating clusters;</td>
</tr>
<tr>
<td>8</td>
<td>HW 1 DUE (Sunday evening, 12 midnight; grace period allow turn in by 8am Monday)</td>
</tr>
<tr>
<td>9</td>
<td>Model-based clustering (including model selection); Nagin clusters</td>
</tr>
<tr>
<td>10</td>
<td>HW 2 DUE (Tuesday evening, 12 midnight; no grace period)</td>
</tr>
<tr>
<td>11</td>
<td>Nagin Clusters (group selection); Classification – intro via K-Nearest-Neighbor; logistic regression; (Linear) Discriminant function analysis (intro)</td>
</tr>
<tr>
<td>13</td>
<td>(Linear) Discriminant function analysis; Support Vector Machines; kernel trick</td>
</tr>
<tr>
<td>16</td>
<td>PROJECT 1 DUE (this is a Monday evening, 12 midnight; grace period allow turn in by 8am Tuesday)</td>
</tr>
<tr>
<td>17</td>
<td>Cross-Validation; Tree-based methods; Naïve Bayes Classifier</td>
</tr>
<tr>
<td>24</td>
<td>PROJECT 2 DUE (this is the Tuesday after last class)*</td>
</tr>
</tbody>
</table>

* PROJECT EXTENSION: to accommodate the data processing/preparation that might be needed, AFTER DISCUSSING WITH ME, you can hand in Project 2 as late as Friday January 28, 2017, 5pm, electronically, without penalty.
Readings


Classes 1 & 2: Handouts 1 & 2


Class 3: Handout 3


Class 4: Handouts 4 & 5

Possible excerpt from “A First Course in Machine Learning”

Class 5: Handouts 5 & 6

Possible excerpt from “A First Course in Machine Learning”

Class 6: Handout 7