Multi-label classification

March 16, 2016; due April 6, 2016

1 (everyone)

For the classification problems that we'e covered so far, the true label y has been assumed to have a single value—either $\{1, 0\}$ in the binary case or $\{1, \ldots, K\}$ in the multiclass case. Multi-*label* classification is the problem of predicting a *vector* of values for a given input data point. Consider, for example, the problem of movie genre prediction, as illustrated in table 1. Rather than assuming that each movie has a single genre, we can hold the perhaps more realistic belief that movies can belong to multiple genres at the same time; this can be cast as a binary multi-label problem by associating each genre with a label and predicting whether or not each movie holds that label ($= 1$) or not ($= 0$).

x = movie	y = [sci-fi, space opera, drama, action, comedy]
Star Wars	[1, 1, 1, 1, 0]
Bridesmaids	[0, 0, 0, 0, 1]
Independence Day	[1, 0, 0, 1, 0]

Table 1: Data for multi-label classification

At this point, we have considered many algorithms that could be adapted for multi-label classification, including the perceptron, logistic regression, graphical models, neural networks, and k-nearest neighbors (among others), along with other methods that could be used in the service of it (e.g., clustering, PCA). Your task is to design **two** different approaches to multi-label classification using the methodology that we've learned so far; at its simplest, this could involve simply training a separate binary classifier for each label in isolation, but be ambitious in thinking through the possibilities that these methods present. For each of the two methods, describe the high-level algorithm you'd use to a.) train that model and b.) make predictions from it. Contrast them on the following dimensions:

- How does each method capture, or fail to capture, the correlations that exist between the labels in y?
- How does each method capture, or fail to capture, the correlations that exist between the features of x?
- How does each method scale as the cardinality of y increases? In the genre example above, the cardinality of $y = 5$; how would it fare if there were 1,000 or 10,000 labels?

Deliverable: 1.5-page paper (single-spaced).