CS 475 Machine Learning: Binary Classification Datasets
Supervised Classifiers 1 Version 1.0

1 Data

The first part of the semester will focus on supervised classification. We consider four real world binary classification datasets in the biology, finance, natural language processing, and computer vision domains. Note, we have already extracted features and vectorized these datasets for you. You followed this same process in Lab 1 when you created a supervised learning dataset. If you happened to create a binary classification dataset, we encourage you to try it on your Perceptron and Logistic Regression implementions!

For each dataset, we have provided you with three files: train, dev, and test. Both the train and dev files have labels, and you should use these to train and evaluate your algorithms. The test file contains unlabeled examples that we will use to test your algorithm. It is **a very good idea** to run on the test data just to make sure your code doesn't crash. You'd be surprised how often this happens.

1.1 Biology

Biological research produces large amounts of data to analyze. Applications of machine learning to biology include finding regions of DNA that encode for proteins, classification of gene expression data and inferring regulatory networks from mRNA and proteomic data.

Our biology task of characterizing gene splice junction sequences comes from molecular biology, a field interested in the relationships of DNA, RNA, and proteins. Splice junctions are points on a sequence at which "superfluous" RNA is removed before the process of protein creation in higher organisms. Exons are nucleotide sequences that are retained after splicing while introns are spliced out. The goal of this prediction task is to recognize DNA sequences that contain boundaries between exons and introns. Sequences contain exon/intron (EI) boundaries, intron/exon (IE) boundaries, or do not contain splice examples.

For a binary task, you will classify sequences as either EI boundaries (label 1) or non-splice sequences (label 0). Each learning instance contains a 60 base pair sequence (ex. ACGT), with some ambiguous slots. Features encode which base pair occurs at each position of the sequence.

1.2 Finance

Finance is a data rich field that employs numerous statistical methods for modeling and prediction, including the modeling of financial systems and portfolios.[1]

[1]For an overview of such applications, see the proceedings of the 2005 NIPS workshop on machine learning in finance. `http://www.icsi.berkeley.edu/~moody/MLFinance2005.htm`

Our financial task is to predict which Australian credit card applications should be accepted (label 1) or rejected (label 0). Each example represents a credit card application, where all values and attributes have been anonymized for confidentiality. Features are a mix of continuous and discrete attributes and discrete attributes have been binarized.

1.3 NLP

Natural language processing studies the processing and understanding of human languages. Machine learning is widely used in NLP tasks, including document understanding, information extraction, machine translation and document classification.

Our NLP task is sentiment classification. Each example is a product review taken from Amazon kitchen appliance reviews. The review is either positive (label 1) or negative (label 0) towards the product. Reviews are represented as uni-gram and bi-grams; each one and two word phrase is extracted as a feature.

1.4 Vision

Computer vision processes and analyzes images and videos and it is one of the fundamental areas of robotics. Machine learning applications include identifying objects in images, segmenting video and understanding scenes in film.

Our vision task is image segmentation. In image segmentation, an image is divided into segments are labeled according to content. The images in our data have been divided into 3x3 regions. Each example is a region and features include the centroids of parts of the image, pixels in a region, contrast, intensity, color, saturation and hue. The goal is to identify the primary element in the image as either a brickface, sky, foliage, cement, window, path or grass. In the binary task, you will distinguish segments of foliage (label 0) from grass (label 1).

2 Questions?

Remember to submit questions about the assignment to the appropriate group on Piazza: `https://piazza.com/class/kdfbbwz3hwb3an`.