Alex Lemon
EE263 (Introduction to Linear Dynamical Systems)
Problem of the Day 5
July 8, 2014

1 Least-squares classification

Consider a corpus of N documents. The ith document in the corpus has an associated feature vector $x^{(i)} \in \mathbb{R}^n$, and binary label $y_i \in \{-1, +1\}$. For example, each component of the feature could be the number of occurrences of a certain term in the document, and the label could be $+1$ for a document that is a legitimate email, and -1 for a document that is spam email. We want to find a classifier that assigns a label to a feature vector $x \in \mathbb{R}^n$ using the formula

$$\hat{y} = \operatorname{sgn}(w^\mathsf{T} x + v),$$

where $w \in \mathbb{R}^n$ and $v \in \mathbb{R}$ are the parameters of the classifiers.

(a) Explain how to choose the parameters w and v in order to minimize

$$J = \sum_{i=1}^{N} (w^\mathsf{T} x^{(i)} + v - y_i)^2.$$

State any assumptions that are needed for your method to work.

(b) The file `ls_classification_data.m` defines the following variables.

- `X`, a matrix whose columns are the observed feature vectors $x^{(1)}, \ldots, x^{(N)}$
- `y`, the vector of observed labels
- `Xtest`, a matrix whose columns are another set of observed feature vectors
- `ytest`, the vector of observed labels for `Xtest`

Apply your method to the set of observations in `X` and `y`. Use the w and v that you find to predict the labels for the observed feature vectors in `Xtest`. Report the number of correct predictions, the number of false negatives ($\hat{y}_i = +1$ and $y_i = -1$) and the number of false positives ($\hat{y}_i = -1$ and $y_i = +1$).