Discussion 1
Probabilistic Machine Learning,Fall 2016

1 Linear Classifiers: Concepts

1. Consider the AND function defined over three binary variables: $f(x_1, x_2, x_3) = (x_1, x_2, x_3)$. We aim to find a θ such that, for any $x = [x_1, x_2, x_3]$, where $x_i \in \{0, 1\}$:

$$\theta x + \theta_0 > 0 \text{ when } f(x_1, x_2, x_3) = 1, \quad \text{and } \theta x + \theta_0 < 0 \text{ when } f(x_1, x_2, x_3) = 0.$$

(a) If $\theta_0 = 0$ (no offset), would it be possible to learn such a θ?

(b) Would it be possible to learn the pair θ and θ_0?

2. You are given the following labeled data points: Positive examples: $[-1, 1]$ and $[1, -1]$,
Negative examples: $[1, 1]$ and $[2, 2]$.

For each of the following parameterized families of classifiers, find the parameters of a classification model in each family that can correctly classify the above data, or explain (with a clear diagram and/or words) why no such choice of parameters exists.

(a) Inside or outside of an origin-centered circle with radius r,

(b) Inside or outside of an $[x, y]$-centered circle with radius r,

(c) Above or below a line through the origin defined by θ,

(d) Above or below a line defined by θ and offset θ_0.

3. Which of the above are families of linear classifiers?

4. Suppose we are given a linear classification model that predicts whether or not it is going to rain based upon the temperature (in degrees Celsius) and humidity (expressed as a percentage from 0-100). The model has weights defined such that if the sum of the temperature and the humidity exceeds 110, then it predicts rainfall instead of clear weather.

(a) Assume that an output of +1 corresponds to predicted rainfall. This model has a weight vector θ of length 2 and a nonzero offset θ_0. What are the values of θ and θ_0?

(b) Consider what happens when we feed this model a data point from the planet Mercury (where it never rains) on which the temperature is observed to be 400° C with a humidity of zero. What does this model predict will happen on Mercury? What does this say about the generalization ability of this model?

2 ROC and AUC

For given data points $\mathbf{X}$, and their corresponding labels $\mathbf{Y}$:

$$\mathbf{X} = [-6,5; -7,5; 0,5; -7,2; 0,1; 1,1; 1,-1; -1,-3]; \quad \mathbf{Y} = [1,1,1,1,0,0,0,0];$$

We have two classifiers:

$$\text{Classifier A} : f(\mathbf{x}) = -2x_1 + 4x_2 - 8$$
$$\text{Classifier B} : f(\mathbf{x}) = -2x_1 + 0x_2 - 8$$

Please compute the following:

(a) Plot the training set as a labeled scatter plot in 2 dimensions.

(b) Plot the decision boundaries for the two classifiers

(c) Calculate the confusion matrix for each classifier

(d) Calculate the ROC curve for each classifier, keeping in mind that the curve goes diagonally to the upper right if a positive and negative example have the same function value.

(e) Calculate the AUC for each classifier (remember to normalize it so the maximum possible value is 1)

(f) Compare the two classifiers with respect to their decision boundaries, ROC and AUC, What can you see from the partition of the data points and the corresponding AUC values? Please make a conclusion about the quality of the classifiers.

(g) If the set of labels were Y = [1, 0, 0, 1, 1, 0, 1, 0]; what would the ROC curve for the classifier $f(\mathbf{x}) = -2x_1 + 4x_2 - 5$ look like? What is the AUC?

3 Linear Classifiers for Multiple Classes

We want to extend the binary classifier shown above to solve a multi-class classification problem. Multi-class classification is similar to binary classification, where one class is assigned to each observation. (This is unlike multi-label classification where multiple labels can be assigned to each observation). One potential method for doing multi-class classification is to divide up the p dimensional input space into K classes. Let $p = 3$ and $K = 3$. This classification problem can be visualized as two or more planes which partition a 3D into several regions. As for the binary case, the region of space which an input point falls into determines its classification.

(a) Assuming our linear classifiers are not collinear, determine the number of regions into which two planes split the 3D input space. Is it equal to K?

(b) A better way to implement multi-class classification is to designate one linear classifier for each class k such that $h_k(x; \theta_k) = \theta_k^T \cdot x$. Then, for any input x, we select the class k such that $h_k(x; \theta_k)$ is maximal. Write an expression for the number of parameters of this model as a function of the number of classes K and the dimensionality of the input N.