Assignment 1: Classification

CS485/685 – Winter 2012

Out: January 17, 2012
Due: February 2, 2012, at the beginning of the lecture. Late assignments may be submitted in the pink drop off box on the third floor of MC within 24 hrs for 50% credit.

Be sure to include your name and student number with your assignment.

1. **[18 pts]** Prove the following properties of the logistic sigmoid function σ:
   - $\sigma(-a) = 1 - \sigma(a)$
   - $\sigma^{-1}(a) = ln(a/(1-a))$
   - $\frac{\partial \sigma}{\partial a} = \sigma(a)(1-\sigma(a))$
2. **[60 pts]** Classication. Implement the following three classification algorithms. Download the dataset posted on the course web page. It is a modified version of the Optical Recognition of Handwritten Digits dataset from the UCI repository. It contains preprocessed black and white images of the digits 5 and 6. Each attribute indicates how many pixels are black in a patch of 4 x 4 pixels. There is one row per image and one column per attribute. The class labels are 5 and 6. Test the algorithms by 10-fold cross validation.
   (a) **[20 pts]** K-Nearest Neighbors: Classify an input x according to the most frequent class amongst its k nearest neighbors. Break ties at random.
   (b) **[20 pts]** Mixture of Gaussians: let $\pi = \Pr(y = C_1)$ and $1 - \pi = \Pr(y = C_2)$. Let $\Pr(x|C_1) = N(x|\mu_1, \Sigma)$ and $\Pr(x|C_2) = N(x|\mu_2, \Sigma)$. Learn the parameters π, μ_1, μ_2 and Σ by likelihood mazimization. Use Bayes theorem to compute the probability of each class given an input x: $\Pr(C_j|x) = k \Pr(C_j) \Pr(x|C_j)$.
   (c) **[20 pts]** Logistic regression: let $\Pr(C_1|x) = \sigma(w^T x + b)$ and $\Pr(C_2|x) = 1 - \sigma(w^T x + b)$. Learn the parameters w and b by conditional likelihood maximization. More specifically use Newton's algorithm derived in class to optimize the parameters. 10 iterations of Newton's algorithm should be sufficient for convergence.

   What to hand in:

   - In k-NN, find the best k by 10-fold cross validation. Draw a graph that shows the accuracy as k increases from 1 to 30.
   - Report the accuracy of mixtures of Gaussians, logistic regression and k-NN (for the best k) obtained by 10-fold cross validation. Measure the accuracy by counting the average number of correctly labeled images. For mixture of Gaussians and logistic regression, an image is correctly labeled when the probability of the correct label is greater than 0.5. For k-NN, an image is correctly labeled when more than half of the k nearest neighbors return the correct label. Break ties randomly.
   - Briefly discuss the results:
     – Mixture of Gaussians and logistic regression both find a linear separator, but they use different parameterizations and different objectives. What can you conclude about the parameterizations and the objectives?

- Mixture of Gaussians and logistic regression find a linear seperator where as k-NN finds a non-linear separator. What can you conclude about the different separators?

- Printout the parameters found for each model.
- Printout your code.

3. **[22 pts]** Linear separability

   (a) **[16 pts]** Consider a threshold perceptron that predicts $y = 1$ when $w^T x + b \geq 0$ and $y = 0$ when $w^T x + b < 0$. It is interesting to study the class of Boolean functions that can be represented by a threshold perceptron. Assume that the input space is $X = \{0, 1\}^2$ and the output space is $Y = \{0, 1\}$. For each of the following Boolean functions, indicate whether it is possible to encode the function as a threshold perceptron. If it is possible, indicate some values for w and b. If it is not possible, indicate a feature mapping $\phi : X \rightarrow \hat{X}$ with values for w and b such that $w^T \phi(x) + b$ is a linear separator that encodes the function.

   - and
   - or
   - exclusive-or
   - iff

   (b) **[6 pts]** Is the dataset used in Question 2 linearly separable? Give a rigorous explanation based on the results obtained in Question 2.