CSI 5325 Assignment 3

Greg Hamerly

Assigned: 2/7/2017; Due: 2/21/2017

Instructions

The instructions for this assignment are the same as for all assignments in this course; for details refer to Assignment 1. As a refresher: use \LaTeX{}, make your document beautiful (proofread it!), use well-labeled figures to illustrate things, turn in a hardcopy and an email copy, and keep attachments small.

Problems

Do the following exercises. Do not just give your answers; show your work (in \LaTeX{}) and explain/analyze your results.

1. Compare two algorithms on a classification task: the Pocket algorithm (which is designed for classification), and linear regression (which is not designed for classification). For linear regression, after learning the weights $w$, we use $h(x) = \text{sign}(w^T x)$ to classify $x$. Here is a starting point for the dataset (given as Octave / MATLAB code):

   ```
   c = 3 * randn(2, 2);  % c(:,1) is the center of the +1 class, c(:,2) is the center of the -1 class
   y = [-1 * ones(50, 1); ones(50, 1)];  % create the +1/-1 labels
   x = randn(100,2) + c((y + 3) / 2, :);  % create the data, and center it using the center labels
   x = [ones(100,1), x];  % add a constant dimension to create the data matrix
   plot(x(y==1,2), x(y==1,3), 'go', x(y==-1,2), x(y==-1,3), 'kx');  % plot it to see what it looks like
   axis equal;
   ```

Create another dataset using the same methods as above, which we will use to estimate $E_{out}$.

Try the following three approaches using multiple randomized experiments and explain which works best in terms both $E_{out}$ and the amount of computation required.

(a) The Pocket algorithm, starting from $w = 0$.

(b) Linear regression (applied as a classification method by taking the sign of the regression output).
(c) The Pocket algorithm, starting from the solution given by linear regression.

Also, try adding some significant outliers to the $y = +1$ class (arbitrarily chosen) of the training dataset and explain how that affects your results.

2. Consider the logistic regression model and its likelihood function:

$$\sigma(\alpha) = \frac{1}{1 + \exp(-\alpha)}$$

$$P(y = 1|x) = \sigma(w^T x)$$

$$P(y = -1|x) = 1 - \sigma(w^T x)$$

$$P(y|x) = P(y = 1|x)^{(y+1)/2}P(y = -1|x)^{(1-y)/2}$$

$$\ell(w) = \log \prod_{n=1}^{N} P(y_n|x_n) = \sum_{n=1}^{N} \log P(y_n|x_n)$$

(a) Show that

$$\frac{d\sigma}{d\alpha} = \sigma(\alpha)(1 - \sigma(\alpha)).$$

(b) Derive the gradient of the log-likelihood, $\nabla_w \ell(w)$.

(c) Write down the update step for gradient ascent of $\ell(w)$ using the gradient you just derived. (Note that in your book, it uses gradient descent on $-\ell(w)$, i.e. minimizing the negative log-likelihood. But we are performing gradient ascent on $\ell(w)$, since we are maximizing the log-likelihood. These two approaches are the same.)

(d) Implement gradient ascent to learn a logistic regression model using the derivations you’ve just performed. Apply it to the datasets you produced in the previous question. Refer to your book for a good description of learning rate, initialization, and stopping conditions.

Experiment and explain your results on the following, in terms of how many iterations it takes to learn, how well it can learn (measure $E_i$ by thresholding the probabilities at 0.5), the weights that are learned, etc. Compare your results with those of the pocket and linear regression algorithms.

- Vary the (fixed) learning rate $\eta$. In other words, try learning with different fixed values of $\eta$.
- Make the classes completely separable, e.g. by changing $c = 3 \cdot \text{randn}(2, 2)$; to be $c = 6 \cdot \text{randn}(2, 2)$. What happens to gradient ascent?