Homework 4

Due Date: May 17, 2016

You may groan about the amount of homework this time. But this is the last homework! So, take a deep breath and go ahead :)

1 Required

1.1 Learning Theory:PAC Learning

PAC stands for "Probably Approximately Correct" and concerns a nice formalism for deciding how much data you need to collect in order for a given classifier to achieve a given probability of correct predictions on a given fraction of future test data.

- True or false: (if true, give a 1 sentence justification; if false, give a counter example.) Within the setting of PAC learning, it is **impossible** to assure with probability 1 that the concept will be learned perfectly (i.e., with true error = 0), regardless of how many training examples are provided.

1.2 Regularization: L2 norm

1. For linear regression, the regularized cost function is:

$$J(\theta) = (y - X\theta)^T(y - X\theta) + \lambda\|\theta\|_2^2$$

, where y is a m-dimensional vector, X is a $m \times n$ matrix, θ is a n-dimensional vector, $\|\theta\|_2$ is the L2 norm of θ. m is sample size, n is feature size. Please find the θ that minimize $J(\theta)$.

2. For logistic regression, the regularized log likelihood is:

$$l(\theta) = \log \prod_{i=1}^{m} h_\theta(x^{(i)})^{y^{(i)}}(1 - h_\theta(x^{(i)}))^{1-y^{(i)}} - \lambda\|\theta\|_2^2$$

, where $h_\theta(x)$ is sigmoid function, m is sample size. Please find the updating rule for θ in gradient descent.

1.3 Generalize EM algorithm

When attempting to run the EM algorithm, it may sometimes be difficult to perform the M step exactly recall that we often need to implement numerical optimization to perform the maximization, which can be costly. Therefore, instead of finding the global maximum of our lower bound on the log-likelihood, and alternative is to just increase this lower bound a little bit, by taking one step of gradient ascent, for example. This is commonly known as the Generalized EM (GEM) algorithm.

Put slightly more formally, recall that the M-step of the standard EM algorithm performs the maximization

$$\theta := \arg\max_{\theta} \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}$$

The GEM algorithm, in contrast, performs the following update in the M-step:

$$\theta := \theta + \alpha \nabla_\theta \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}$$

where α is a learning rate which we assume is chosen small enough such that we do not decrease the objective function when taking this gradient step.

1. Prove that the GEM algorithm described above converges. To do this, you should show that the the likelihood is monotonically improving, as it does for the EM algorithm — i.e., show that $\ell(\theta^{(t+1)}) \geq \ell(\theta^{(t)})$.

2. Instead of using the EM algorithm at all, suppose we just want to apply gradient ascent to maximize the log-likelihood directly. In other words, we are trying to maximize the (non-convex) function

$$\ell(\theta) = \sum_i \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)}; \theta)$$

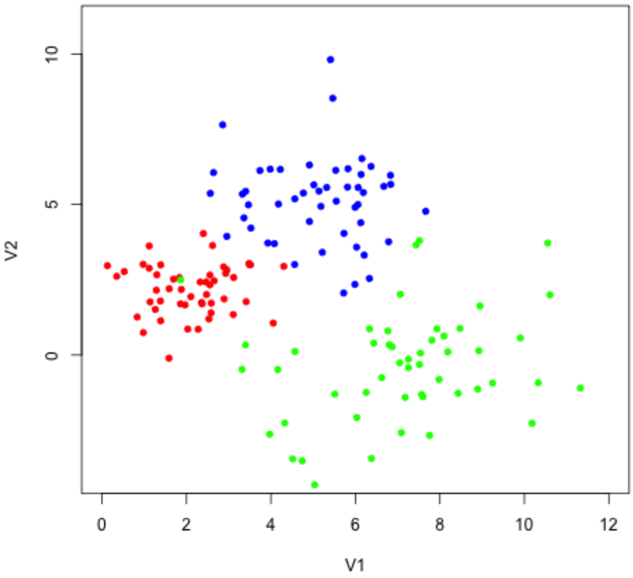
so we could simply use the update

$$\theta := \theta + \alpha \nabla_\theta \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}$$

Show that this procedure in fact gives the same update as the GEM algorithm described above.

2 Optional

2.1 GMM and K-means



Your TAs have drawn some data from three different distributions (see red, green, blue dots in the graph). But they forgot to write down the corresponding labels for each data point. Please help them by using GMM and K-means to cluster the data. Things to be included in your report:

1. a graph indicating different clusters with different colors
2. list the corresponding parameters for the model. (For GMM, they are the weights, means, covariances; for k-means, they are the positions of the centroids).
3. the comparison between GMM and K-means