CS 3190: Foundations of Data Analysis **Due: Friday 12.9 at 11:50pm**

Homework 5: Clustering and Classification

Instructions: Your answers are due at 11:50pm. You must turn in a pdf through canvas I recommend using latex (`http://www.cs.utah.edu/~jeffp/teaching/latex/`) for producing the assignment answers. If the answers are too hard to read you will lose points, entire questions may be given a 0 (e.g. **sloppy pictures with your phone's camera are not ok, but very careful ones are**)

Please make sure your name appears at the top of the page.

You may discuss the concepts with your classmates, but write up the answers entirely on your own. **Be sure to show all the work involved in deriving your answers! If you just give a final answer without explanation, you may not receive credit for that question.**

---

1. **[40 points]** Consider this set of 3 sites: $S = \{s_1 = (-3,-1), s_2 = (1,1), s_3 = (-2,2)\} \subset \mathbb{R}^2$. We will consider the following 5 data points $X = \{x_1 = (-2,0), x_2 = (-2,1), x_3 = (-1,1), x_4 = (0,0), x_5 = (-3,-2)\}$.

   For each of the following points compute the closest site (under Euclidean distance):

   (a) $\phi_S(x_1) =$

   (b) $\phi_S(x_2) =$

   (c) $\phi_S(x_3) =$

   (d) $\phi_S(x_4) =$

   (e) $\phi_S(x_5) =$

   Now consider that we have 3 Gaussian distributions defined with each site s_j as a center μ_j. The corresponding standard deviations are $\sigma_1^2 = 0.3$, $\sigma_2^2 = 1.0$ and $\sigma_3^2 = 1.0$, and we assume they are univariate so the covariance matrices are $\Sigma_j = \begin{bmatrix} \sigma_j^2 & 0 \\ 0 & \sigma_j^2 \end{bmatrix}$.

   (f) Write out the probability density function (its likelihood $f_j(x)$ for each of the Gaussians).

   Now we want to assign each x_i to each site in a soft assignment. For each site s_j define the weight of a point as $w_j(x) = f_j(x)/(\sum_{k=1}^{3} f_k(x))$. For each of the following points calculate the weight for each site

   (g) $w_1(x_1), w_2(x_1), w_3(x_1) =$

   (h) $w_1(x_2), w_2(x_2), w_3(x_2) =$

   (i) $w_1(x_3), w_2(x_3), w_3(x_3) =$

   (j) $w_1(x_4), w_2(x_4), w_3(x_4) =$

   (k) $w_1(x_5), w_2(x_5), w_3(x_5) =$

2. **[20 points]** Construct a data set X with 4 points in $\mathbb{R}^2$ and a set S of $k = 2$ sites so that Lloyds algorithm will have converged, but there is another set S', of size $k = 2$, so that $\textsf{cost}(X, S') < \textsf{cost}(X, S)$. Explain why S' is better than S, but that Lloyds algorithm will not move from S.

3. **[20 points]** Suppose we have a dataset (X, y) where $y \in \{-1, +1\}$ and define a linear function $g(x) := \langle (1, x), \alpha \rangle$ where the overall cost is defined as $\mathcal{L}(g, (X, y)) = \sum_{i=1}^{n} \ell_i(y_i \cdot g(x_i))$ for some loss function ℓ_i

   (a) Explain why setting $\ell_i(x) = x^2$ would be inappropriate

   (b) If we suppose that ℓ_i is the Δ loss function, then explain why we defined $\mathcal{L}(g, (X, y)) = \sum_{i=1}^{n} \ell_i(y_i \cdot g(x_i))$ and **not** $\mathcal{L}(g, (X, y)) = \sum_{i=1}^{n} \ell_i(y_i - g(x_i))$

   For parts (c) and (d) of this question, we'll suppose

$$\ell_i(z) = \begin{cases} 0 & \text{if } z > 1 \\ 1 - z & \text{if } 0 \le z \le 1 \\ 1 & \text{if } z \le 0. \end{cases}$$

   (c) What problems might a gradient descent algorithm have when attempting to minimize $\mathcal{L}$ by choosing the best α?

   (d) Explain if the problem would be better or worse using stochastic gradient descent?

4. **[20 points]**

   (a) Construct and report a set of labeled points (X, y) in $\mathbb{R}^2$ that is not linearly separable (provide a plot).

   (b) Explain what will happen if you run the perceptron algorithm for a linear classifier on this data set? (don't allow a fixed upper bound on T the number of steps)

   (c) Describe another algorithm discussed in the class (Chapters 9.1 - 9.3) which would provides an acceptable linear classifier for the set of points from part (a).