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: $S = \{s_1 = (1, 1), s_2 = (0, 3), s_3 = (-1, 0)\} \subset \mathbb{R}^2$.

   We will consider the following 5 data points $X = \{x_1 = (0, 0), x_2 = (-1, 4), x_3 = (2, 1), x_4 = (0, 1), x_5 = (0, 2)\}$.

   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 $\mu_j$. The corresponding standard deviations are $\sigma_1 = 1/\sqrt{2}$, $\sigma_2 = 1.0$ and $\sigma_3 = \sqrt{2}$, and we assume they are univariate so the covariance matrices are $\Sigma_j = \begin{bmatrix} \sigma^2_j & 0 \\ 0 & \sigma^2_j \end{bmatrix}$.

   (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 $w_j(x) = f_j(x) / (\sum_{j=1}^{3} f_j(x))$. 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 5 points in $\mathbb{R}^2$ and a set $S$ of $k = 3$ sites so that Lloyds algorithm will have converged, but there is another set $S'$, of size $k = 3$, so that $\text{cost}(X, S') < \text{cost}(X, S)$. Explain why $S'$ is better than $S$, but that Lloyds algorithm will not move from $S$.

3. [25 points] Consider a family of linear classifiers defined by the sign of function $g_{w,b}(x) = \langle w, x \rangle + b$, where $x \in \mathbb{R}^2$ and so $w \in \mathbb{R}^2$ and $b \in \mathbb{R}$. Given a data point $x_i$ and label $y_i \in \{-1, +1\}$. We require that $\|w\| = 1$.

Now consider a uncertainty zone misclassification goal $\Lambda$ (in place of $\Delta$). In this setting, we want to penalize a classifier with a cost of $1/4$ for any point within a distance of 2 of the classification boundary – even if it has the correct sign. So the cost is

$$
\Lambda(g_{w,b}, (x_i, y_i)) = \begin{cases} 
1 & \text{if } (x_i, y_i) \text{ is misclassified and } |g_{w,b}(x_i)| > 2 \\
1/4 & \text{if } 0 \leq |g_{w,b}(x_i)| \leq 2 \\
0 & \text{if } (x_i, y_i) \text{ is classified correctly and } |g_{w,b}(x_i)| > 2
\end{cases}
$$

(a) Explain $\Lambda(g_{w,b}, (x_i, y_i))$ as a function of $z_i = y_i g_{w,b}(x_i)$.

(b) Design a loss function $\ell_\Lambda(z)$ as proxy for $\Lambda(z)$ that is (i) convex, (ii) has a derivative defined for all $z$, and (iii) for all values of $z$ satisfies $\ell_\Lambda(z) \geq \Lambda(z)$.

4. [15 points] Consider a data set $(X, y)$, where each data point $(x_{1,i}, x_{2,i}, y_i)$ is in $\mathbb{R}^2 \times \{-1, +1\}$. Provide the psuedo-code for the Perceptron Algorithm using a polynomial kernel of degree 2. You can have a generic stopping condition, where the algorithm simply runs for $T$ steps for some parameter $T$. (There are several correct ways to do this, but be sure to explain how to use a polynomial kernel clearly.)