Danning Yu, 305087992
CS M146 Discussion 1A
PS5
3/8/2020

1. (a) Denoting the cost function as J:

$$\frac{\partial J}{\partial \beta_t} = \frac{\partial}{\partial \beta_t}\left(\left(e^{\beta_t} - e^{-\beta_t}\right)\varepsilon_t + e^{-\beta_t}\right)$$

Note that ε_t is not dependent on β_t, so we can treat it as a constant. Setting the derivative to 0:

$$\frac{\partial}{\partial \beta_t}\left(\left(e^{\beta_t} - e^{-\beta_t}\right)\varepsilon_t + e^{-\beta_t}\right) = 0$$

$$\left(e^{\beta_t} + e^{-\beta_t}\right)\varepsilon_t - e^{-\beta_t} = 0$$

$$\varepsilon_t e^{\beta_t} = e^{-\beta_t}\left(1 - \varepsilon_t\right)$$

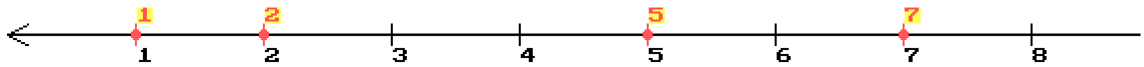
$$\frac{e^{\beta_t}}{e^{-\beta_t}} = \frac{1-\varepsilon_t}{\varepsilon_t}$$

$$e^{2\beta_t} = \frac{1-\varepsilon_t}{\varepsilon_t}$$

$$\beta_t = \frac{1}{2}\log\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$$

(b) If the training set is linearly separable and no slack is allowed, then there will be no misclassification error, and so ε_t will go to 0, and based on the result of part (a), if ε_t goes to 0, then β_1 will go to infinity.

2. (a) Plotting these points on a number line, we get:



For the case of K = 3, the optimal clustering is to have the centers at $\mu_1 = 1.5$, $\mu_2 = 5$, and $\mu_3 = 7$, where x_1 and x_2 are assigned to μ_1, x_3 is assigned to μ_2, and x_4 is assigned to μ_3. The value of the objective is

$$(1-1.5)^2 + (2-1.5)^2 + (5-5)^2 + (7-7)^2 = 0.5$$

(b) A possible suboptimal assignment would be $\mu_1 = 1$, $\mu_2 = 2$, and $\mu_3 = 6$, where x_1 is assigned to μ_1, x_2 is assigned to μ_2, and x_3 and x_4 are assigned to μ_3. The value of the objective is

$$(1-1)^2 + (2-2)^2 + (5-6)^2 + (7-6)^2 = 2$$

which is clearly greater than the value of 0.5 obtained in part (a), but if Lloyd's algorithm is applied to this current assignment, x_1 is closest to μ_1, x_2 is closest to μ_2, and x_3 and x_4 are closest to μ_3, thus resulting in no change of assignments or centroids, so we are at a suboptimal solution that is only a local minimum and not the global minimum.

3. (a) The multivariate normal distribution of a variable with *d* dimensions is defined as:

$$N(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu}) \right)$$

Substituting this into the expression for $l(\boldsymbol{\theta})$ and taking the gradient with respect to $\boldsymbol{\mu}_j$:

$$l(\boldsymbol{\theta}) = \sum_k \sum_n \gamma_{nk} \log \omega_k + \sum_k \sum_n \gamma_{nk} \log\left(\frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}_k|}} \exp\left(-\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k) \right) \right)$$

$$l(\boldsymbol{\theta}) = \sum_k \sum_n \gamma_{nk} \log \omega_k + \sum_k \sum_n \gamma_{nk} \left(\left(\log \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}_k|}} \right) - \frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k) \right)$$

The first summation is not a function of $\boldsymbol{\mu}_\mathrm{j}$ and the first term within the second summation is a constant, so the gradient of both evaluates to 0. Take the gradient to get:

$$\nabla_{\boldsymbol{\mu}_j} l(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\mu}_j} \sum_n \left(-\frac{1}{2} \gamma_{nj} (\mathbf{x}_n - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_j) \right)$$

$$\nabla_{\boldsymbol{\mu}_j} l(\boldsymbol{\theta}) = \sum_n \left(-\frac{1}{2} \gamma_{nj} (2)(-1)(\mathbf{x}_n - \boldsymbol{\mu}_j) \boldsymbol{\Sigma}_j^{-1} \right)$$

$$\nabla_{\boldsymbol{\mu}_j} l(\boldsymbol{\theta}) = \boldsymbol{\Sigma}_j^{-1} \sum_n \left(\gamma_{nj} (\mathbf{x}_n - \boldsymbol{\mu}_j) \right)$$

(b) Setting the answer obtained in part (a) to $\mathbf{0}$ and solving for $\boldsymbol{\mu}_j$ to obtain the desired answer:

$$\nabla_{\boldsymbol{\mu}_j} l(\boldsymbol{\theta}) = \boldsymbol{\Sigma}_j^{-1} \sum_n \left(\gamma_{nj} (\mathbf{x}_n - \boldsymbol{\mu}_j) \right) = \mathbf{0}$$

$$\sum_n \gamma_{nj} \mathbf{x}_n = \boldsymbol{\mu}_j \sum_n \gamma_{nj}$$

$$\boldsymbol{\mu}_j = \frac{\sum_n \gamma_{nj} \mathbf{x}_n}{\sum_n \gamma_{nj}}$$

(c) From the lecture notes, ω_k and $\boldsymbol{\mu}_k$ are given by

$$\omega_k = \frac{\sum_n \gamma_{nk}}{\sum_k \sum_n \gamma_{nk}} \qquad \boldsymbol{\mu}_k = \frac{\sum_n \gamma_{nk} \mathbf{x}_n}{\sum_n \gamma_{nk}}$$

Substituting in the values from the table, we get the following values:

$$\omega_1 = \frac{0.2+0.2+0.8+0.9+0.9}{0.2+0.2+0.8+0.9+0.9+0.8+0.8+0.2+0.1+0.1} = \frac{3}{5} = 0.6$$

$$\omega_2 = \frac{0.8+0.8+0.2+0.1+0.1}{5} = \frac{2}{5} = 0.4$$

$$\mu_1 = \frac{1}{3}\left(0.2(5)+0.2(15)+0.8(25)+0.9(30)+0.9(40)\right) = 29$$

$$\mu_2 = \frac{1}{2}\left(0.8(5)+0.8(15)+0.2(25)+0.1(30)+0.1(40)\right) = 14$$