Aaron Dardik

CS567 Homework #3

1.1) For T_1 its left child has 150 examples in class A and 50 in class B. This implies its right child has 50 examples in class A and 150 in class B. For T_2 the left child has 0 examples in class A and 100 in class B. This implies its right child has 200 examples in class A and 100 in class B. Entropy is given by the following formula: $-\sum_{k=1}^{C} p(Y=k)\log\big(p(Y=k)\big)$. For T_1, $P(A|\text{left})$ = $\frac{3}{4}$, $P(B|\text{left})$ = $\frac{1}{4}$. Still for the first tree we have that $P(A|\text{right})$ = $\frac{1}{4}$, $P(B|\text{right})$ = $\frac{3}{4}$. Thus, for T_1 we have that entropy of the left branch is $-(\frac{3}{4}\log\frac{3}{4}+\frac{1}{4}\log\frac{1}{4})$ and the entropy of the right branch is $-(\frac{1}{4}\log\frac{1}{4}+\frac{3}{4}\log\frac{3}{4})$. Now the entropy of $T_1$is ½*(entropy of left branch) + 1/2 *(entropy of right branch). As the two branches are equal, the entropy of T_1 = entropy of left branch = entropy of right branch = 0.56. For T_2 we have that $P(A|\text{left})$ = 0, $P(B|\text{left})$ = 1. On the right branch we have $P(A|\text{right})$ = $\frac{2}{3}$, $P(B|\text{right})$ = $\frac{1}{3}$. For T_2 we have the entropy of the left branch is $-(0\log 0+1\log 1)=0$ and the entropy of the right branch is $-(\frac{2}{3}\log\frac{2}{3}+\frac{1}{3}\log\frac{1}{3})$. For T_2 we have that its entropy is equal to ¼ * (entropy of the left branch) + ¾ * (entropy of the right branch) = $\frac{1}{4}*0+\frac{-3}{4}*\left(\frac{2}{3}\log\frac{2}{3}+\frac{1}{3}\log\frac{1}{3}\right)=0.48$. Gini impurity is given by $\sum p(Y=k)*\big(1-p(Y=k)\big)$. For T_1 on the left branch we have $P(A|\text{left})$=$\frac{3}{4}$, $P(B|\text{left})$=$\frac{1}{4}$. Therefore, for this branch the Gini impurity is $\frac{3}{4}*\frac{1}{4}+\frac{1}{4}*\frac{3}{4}=\frac{3}{8}$. As this tree displays symmetry, the Gini impurity of the right branch is also $\frac{3}{8}$ and as the Gini impurity of T_1 is ½ *(impurity of the left branch) + ½ *(impurity of the right branch) = $\frac{1}{2}*\frac{3}{8}+\frac{1}{2}*\frac{3}{8}=\frac{3}{8}$. For T_2 the Gini impurity of the left branch is $0*(1-0)+1*(1-1)=0$. For the right branch the Gini is $\frac{2}{3}*\frac{1}{3}+\frac{1}{3}*\frac{2}{3}=\frac{4}{9}$. The total Gini for $T_2=\frac{1}{4}*0+\frac{3}{4}*\frac{4}{9}=\frac{1}{3}$. We will now calculate each tree's classification error. T_1 assigns the left branch to A and the right to B, so on the left side $\frac{50}{200}=\frac{1}{4}$ are misclassified. On the right branch $\frac{50}{200}=\frac{1}{4}$ are misclassified so the classification error for $T_1=\frac{1}{2}*\frac{1}{4}+\frac{1}{2}*\frac{1}{4}=\frac{1}{4}=0.25$. For T_2 the classification error on the left branch is 0, and on the right branch is $\frac{1}{3}$ so the total classification error of $T_2=\frac{1}{4}*0+\frac{3}{4}*\frac{1}{3}=\frac{1}{4}=0.25$.

1.2) For T_1 we have entropy = 0.56, Gini impurity is 0.38 and classification error is 0.25. For T_2 entropy is 0.48, Gini impurity is 0.33 and classification error is 0.25. Based on classification error, the two trees are of the same quality, but using entropy and Gini impurity as metrics, causes T_2 to come out ahead. Therefore T_2 is of higher quality.

2.1) Let $L(\beta_t)=\varepsilon_t\big(e^{\beta_t}-e^{-\beta_t}\big)+e^{-\beta_t}$. As L is a convex function, its local minimum is also its global minimum. We will find the local, and therefore global minimum. Taking the derivative and setting it equal to 0, we get that $\frac{\partial L}{\partial\beta_t}=\varepsilon_t\big(e^{\beta_t}+e^{-\beta_t}\big)-e^{-\beta_t}=0$. We will perform a change of variables and let

$\beta_t = \ln z$. Thus, $0 = \varepsilon_t\left(z + \frac{1}{z}\right) - \frac{1}{z} \Rightarrow 0 = \varepsilon_t(z^2 + 1) - 1$. Rearranging, we get $z^2 = \frac{1}{\varepsilon_t} - 1 \Rightarrow z = \pm\sqrt{\frac{1}{\varepsilon_t} - 1}$. Now, as $z = \beta_t, \Rightarrow \beta_t^* = \ln\sqrt{\frac{1}{\varepsilon_t} - 1} = \frac{1}{2}\ln\left(\frac{1}{\varepsilon_t} - 1\right) = \frac{1}{2}\ln\frac{1-\varepsilon_t}{\varepsilon_t}$. The second derivative test confirms this is indeed a minimum and as the local minimum of a convex function is the global minimum, we have shown that $\beta_t^* = \frac{1}{2}\ln\frac{1-\varepsilon_t}{\varepsilon_t}$.

2.2) We seek to prove that $\sum_{n:h_t(x_n)\neq y_n} D_{t+1}(n) = \frac{1}{2}$. $\varepsilon_1 = \sum_{n:y_n\neq h_1(x_n)} \frac{1}{N} = \frac{\#(\text{examples misclassified by } h_1)}{N}$ where the notation #(…) means "number of…" Similarly, $1 - \varepsilon_1 = \frac{\#(\text{examples correctly classified by } h_1)}{N}$ and that $\beta_1 = \frac{1}{2}\log\frac{\#(\text{examples correctly classified by } h_1)}{\#(\text{examples incorrectly classified by } h_1)}$.

Now, $\sum_{n:y_n\neq h_1(x_n)} D_2(n) = \frac{\frac{1}{N} * e^{\beta_1} * \#(\text{examples misclassified by } h_1)}{\frac{\#(\text{examples correctly classified})}{N} * e^{-\beta_1} + \frac{\#(\text{examples incorrectly classified})}{N} * e^{\beta_1}} = \frac{\varepsilon_1 * e^{\beta_1}}{(1-\varepsilon_1) * e^{-\beta_1} + \varepsilon_1 * e^{\beta_1}}$. Multiplying top and bottom by β_1 we see that the preceding equation simplifies to the following expression: $\frac{\varepsilon_1 * e^{2\beta_1}}{(1-\varepsilon_1) + \varepsilon_1 * e^{2\beta_1}}$. Remembering that $\beta_1 = \frac{1}{2}\ln\frac{1-\varepsilon_1}{\varepsilon_1}$, so that $e^{2\beta_1} = \frac{1-\varepsilon_1}{\varepsilon_1}$ and our preceding quotient will be reduced to $\frac{\varepsilon_1 * \frac{1-\varepsilon_1}{\varepsilon_1}}{(1-\varepsilon_1) + \varepsilon_1 * \frac{1-\varepsilon_1}{\varepsilon_1}} = \frac{1-\varepsilon_1}{(1-\varepsilon_1)+(1-\varepsilon_1)} = \frac{1-\varepsilon_1}{2(1-\varepsilon_1)} = \frac{1}{2}$. This demonstrates the base case. We will now assume the inductive hypothesis, that $\sum_{n:h_{t-1}(x_n)\neq y_n} D_t(n) = \frac{1}{2}$. Then, $\sum_{n:h_t(x_n)\neq y_n} D_{t+1}(n) = \sum_{n:h_t(x_n)\neq y_n} {(D_t(n) * e^{-\beta_t y_n h_t(x_n)})}\Big/{(\sum_{n'=1}^{N} D_t(n') * e^{-\beta_t y_n' h_t(x_n')})}$. As we are summing over the cases where $h_t(x_n) \neq y_n$, we can see that the coefficient of β_t in the numerator will never have a minus sign in front. This allows us to simplify the expression to $e^{\beta_t} * (\sum_{n:h_t(x_n)\neq y_n} {D_t(n)}\Big/{\left(\sum_{n'=1}^{N} D_t(n') * e^{-\beta_t y_n' h_t(x_n')}\right)}) = \frac{e^{\beta_t} * \sum_{n:h_t(x_n)\neq y_n} D_t(n)}{\sum_{n:h_t(x_n)\neq y_n} D_t(n) * e^{\beta_t} + \sum_{n:h_t(x_n)=y_n} D_t(n) * e^{-\beta_t}}$. Now, using the inductive hypothesis and substituting back in for β_t, ε_t as well as recognizing that $e^{2\beta_t} = \frac{1-\varepsilon_t}{\varepsilon_t}$ and that $\sum_n D_t(n) = 1$, we can now simplify the quotient to $\frac{e^{2\beta_t} * \sum_{n:h_t(x_n)\neq y_n} D_t(n)}{e^{2\beta_t} * \sum_{n:h_t(x_n)\neq y_n} D_t(n) + \sum_{n:h_t(x_n)=y_n} D_t(n)}$, and now by plugging in the values and simplifying further we reduce the quotient to the following form: $\frac{\frac{1-\varepsilon_t}{\varepsilon_t} * \varepsilon_t}{(1-\varepsilon_t)+(1-\varepsilon_t)}$ as $\sum_n D_t(n) = \sum_{n:h_t(x_n)\neq y_n} D_t(n) + \sum_{n:h_t(x_n)=y_n} D_t(n)$ and the first sum on the right hand side of the preceding expression is ε_t. Thus, our quotient reduces to $\frac{1-\varepsilon_t}{(1-\varepsilon_t)+(1-\varepsilon_t)} = \frac{1}{2}$, and therefore the result holds for all t.

3.1) We would like to solve $\underset{q}{\text{argmax}} \sum_k a_k \ln q_k$, such that $q_k \geq 0, \sum q_k = 1$. This is equivalent to finding $\underset{q}{\text{argmin}} -\sum_k a_k \ln q_k$ such that $-q_k \leq 0, \sum q_k - 1 = 0$. The Lagrangian of this expression $L(q, \alpha, \beta)$is given by $L = -\sum_k a_k \ln q_k - \sum_k \alpha_k q_k + \beta(\sum_k q_k - 1)$ and the solution is found by the values that satisfy, $\frac{\partial L}{\partial q_k} = 0, \frac{\partial L}{\partial \beta} = 0, \alpha_k \geq 0, \alpha_k q_k = 0, \nabla_\alpha L \leq 0$. Before differentiating, note that 0 is not a possible solution as it is not in the domain of the function due to $\ln 0$ being undefined. Now, taking

$\frac{\partial L}{\partial q_k}$ and setting it equal to zero, we have $\frac{\partial L}{\partial q_k} = \frac{-a_k}{q_k} - \alpha_k + \beta = 0$. Multiplying through by q_k we see that $-a_k - \alpha_k q_k + \beta q_k = 0$, and note that $\alpha_k = 0$ for all k as $q_k \neq 0$. This implies that $\beta q_k = a_k \Rightarrow q_k = \frac{a_k}{\beta}$. Now, as $\sum_k q_k = 1 \Rightarrow \sum_k a_k = \beta$ and using this to substitute back in to the expression for $q_k$we have $q_k = \frac{a_k}{\sum_k a_k}$.

3.2) Here we would like to solve the following problem: $\underset{q}{\text{argmax}}\ \sum_k (q_k b_k - q_k \ln q_k)$, such that $q_k \geq 0, \sum q_k = 1$. This is the same as finding $\underset{q}{\text{argmin}}\ \sum_k (q_k \ln q_k - b_k q_k)$ such that $-q_k \leq 0, \sum q_k = 1$. We will now construct the Lagrangian. However, by the same reasoning as from problem 3.1 we see that the α_k are all zero. So, we have $L(q, \beta) = \sum_k (q_k \ln q_k - b_k q_k) + \beta(\sum_k q_k - 1)$. Taking the derivative with respect to q_i, and setting it equal to zero, we have $\frac{\partial L}{\partial q_i} = \frac{q_i}{q_i} + \ln q_i - b_i + \beta = 0$. Simplifying, we have $\ln q_i = b_i - \beta - 1 \Rightarrow q_i = e^{b_i - \beta - 1}$. Note that this is equal to $e^{-\beta-1} * e^{b_i}$, which, as $e^{-\beta-1}$ is a constant, which we can call C, is equal to Ce^{b_i}. Again noting that $\sum_k q_k = 1 \Rightarrow C \sum e^{b_i} = 1$ so we can conclude that $C = \frac{1}{\sum e^{b_i}}$ and therefore, $q_k = \frac{e^{b_k}}{\sum_i e^{b_i}}$.

4.1) To solve this problem we will begin by solving to separate maximization problems. The first problem we will solve is $\underset{w_k}{\text{argmax}}\ \sum_n \sum_k \gamma_{nk} \ln w_k$ such that $w_k \geq 0, \sum_k w_k = 1$. Maximizing our function over w_k is the same as minimizing -1 * our function. We also require "shift" the greater than conditions to be $-w_k \leq 0$. We create the Lagrangian, $L(w_k, \alpha_k, \beta) = -\sum_n \sum_k \gamma_{nk} \ln w_k - \sum_k \alpha_k w_k + \beta(\sum_k w_k - 1)$ with the requirement that $\frac{\partial L}{\partial w_k} = 0$ implying $-\sum_n \frac{\gamma_{nk}}{w_k} - \alpha_k + \beta = 0$. Rearranging, we get that $w_k = \frac{\sum_n \gamma_{nk}}{\beta - \alpha_k}$. However, we see that together $\beta(\sum w_k - 1) = 0, \sum \alpha_k w_k = 0$,and using the previous result to see that $w_k > 0$ is a strict inequality, $\Rightarrow\ \alpha_k = 0$. Therefore, $w_k = \frac{\sum_n \gamma_{nk}}{\beta}$. And since $\sum_k w_k = 1 \Rightarrow \sum_k \frac{\sum_n \gamma_{nk}}{\beta} = 1 \Rightarrow \beta = \sum_{n,k} \gamma_{nk}$ and therefore, $w_k = \frac{\sum_n \gamma_{nk}}{\sum_{n,k} \gamma_{nk}} = \frac{\sum_n \gamma_{nk}}{N}$. The second problem to solve is $\underset{\mu_k, \Sigma_k}{\text{argmax}} \sum_n \sum_k \gamma_{nk} \ln \mathcal{N}(x_n | \mu_k, \Sigma_k) = \underset{\mu_k, \Sigma_k}{\text{argmax}}\ \sum_n \gamma_{nk} \left(\ln \frac{1}{|\Sigma_k|^{\frac{1}{2}}} - \frac{[x_n - \mu_k]^T \Sigma_k^{-1} [x_n - \mu_k]}{2} \right)$. We put the equation into the required form for the KKT conditions as we did in the last step, and begin by taking the derivative of the Lagrangian with respect to μ_k and setting it equal to zero. Thus, $\sum_n \gamma_{nk} [x_n - \mu_k] = 0 \Rightarrow \mu_k = \frac{\sum_n \gamma_{nk} x_n}{\sum_n \gamma_{nk}}$. Taking the derivative with respect to Σ_k and setting it equal to zero, we get $\sum_n \gamma_{nk} (\frac{-1}{2} \Sigma_{\mathrm{k}}^{-1} + \frac{1}{2} * \Sigma_{\mathrm{k}}^{-1} [x_n - \mu_k][x_n - \mu_k]^T \Sigma_k^{-1}) = 0$. Our previous equation simplifies to $\sum_n \gamma_{nk} \Sigma_k^{-1} = \Sigma_k^{-1} [x_n - \mu_k][x_n - \mu_k]^T \Sigma_k^{-1}$. By multiplying both sides of the equation by Σ_k on the right, then dividing both sides by $\sum_n \gamma_{nk}$ and then multiplying both sides on the left by Σ_k we get $\Sigma_k = \frac{[x_n - \mu_k][x_n - \mu_k]^T}{\sum_n \gamma_{nk}}$.

4.2) From problem 3.2 we see that $q_k^* = \underset{q}{\text{argmax}} \sum_k (q_k b_k - q_k \ln q_k) = \frac{e^{b_k}}{\sum_i e^{b_i}}$. Now solving $\underset{q_k \in \Delta}{\text{argmax}}\ \mathbb{E}_{z_n \sim q_n}[\ln p(x_n, z_n; \theta^t)] - \mathbb{E}_{z_n \sim q_n}[\ln q_n]$ and comparing to q_k^* we have that $b_k = \ln p(x_n, z_n = k; \theta^t) \Rightarrow q_k^* = \frac{e^{\ln p(x_n, z_n = k; \theta^t)}}{\sum_i e^{\ln p(x_i, z_i; \theta^t)}} = \frac{p(x_n, z_n = k; \theta^t)}{\sum_i p(x_n, z_n = i; \theta^t)} = \frac{p(x_n, z_n = k; \theta^t)}{p(x_n; \theta^t)} = p(z_n = k | x_n; \theta^t)$ and the result holds.

4.3) The loss function for K-means clustering can be written as $\underset{S}{\text{argmin}}\ \sum_{i=1}^n \sum_{j=1}^k 1\left(x_i \in S_j\right)\left(x_i - \mu_j\right)^2$. If we consider a w_{ij} such that $\sum_j w_{ij} = 1$ with the w_{ij} thus representing "how much," or "the proportion of" of x_i is in class S_j. We now write $\underset{S}{\text{argmin}}\ \sum_{i=1}^n \sum_{j=1}^k w_{ij}\left(x_i - \mu_j\right)^2$. From this expression, if we set $w_{ij} = \begin{cases} 1, \text{ if } x_i \in S_j \\ 0 \text{ if } x_i \notin S_j \end{cases}$ then we can recover the loss-function from K-means clustering as follows: the MLE for the Gaussian mixture model is found by maximizing $\sum_{k=1}^K (\ln \frac{w_k}{\sqrt{|\Sigma_k|}} - \frac{1}{2}[x_n - \mu_k]^T \Sigma_k^{-1} [x_n - \mu_k])$ and if $\Sigma_k = \sigma^2 I$ for all *k* (i.e. all clusters have the same radius and are spherical) and w_{ij} is as described, then K-means clustering can be seen as a special case of the mixture model. The model parameters are therefore $w_i = \frac{1}{K}, \forall i$ and $\Sigma_k = \sigma^2 I, \forall k$. This shows that K-means clustering is a specific instance of the Gaussian mixture model where we consider the mixture weights to be equal and the variance to be spherical with each cluster having the same radius. The responsibility $p(z_n = k | x_n) \triangleq \gamma_{nk}$ is given by $\frac{w_k e^{\frac{-1}{2\sigma} * \|x_n - \mu_k\|^2}}{\sum_i w_i e^{\frac{-1}{2\sigma} * \|x_n - \mu_i\|^2}}$, which as variance goes to zero is equal to $\frac{1}{K}$.