CS188, Winter 2017
Problem Set 5: Boosting, Unsupervised Learning
Due March 16,2017

1 Problem 1

(a) Problem 1a

Solution: We have

$$(h_t(x), \beta_t x) = (e^{\beta_t} - e^{-\beta_t}) \sum_n w_t(n) I(y_n \neq h_t(x)) + e^{-\beta_t} \sum_n w_t(n)$$

Since $\epsilon_t = \sum_n w_t(n) I(y_n \neq h_t(x))$ and $\sum_n w_t(n) = 1$, we have:

$$(h_t(x), \beta_t) = (e^{\beta_t} - e^{-\beta_t})\epsilon_t + e^{-\beta_t} = e^{\beta_t}\epsilon_t - e^{-\beta_t}\epsilon_t + e^{-\beta_t}$$

$$\frac{\delta(h_t(x), \beta_t}{\delta\beta} = \epsilon_t e^{\beta_t} + \epsilon_t e^{-\beta_t} - e^{-\beta_t} = 0$$

$$\epsilon_t e^{\beta_t} = e^{-\beta_t} - \epsilon_t e^{-\beta_t}$$

$$\epsilon_t e^{\beta_t} = e^{-\beta_t}(1 - \epsilon_t)$$

$$log(\epsilon_t) + \beta_t = log(1 - \epsilon_t) - \beta_t$$

$$2\beta_t = log(\frac{1 - \epsilon_t}{\epsilon_t})$$

$$\beta_t = \frac{1}{2}(\frac{1 - \epsilon_t}{\epsilon_t})$$

giving us the update for β_t.

(b) Problem 1b

Solution: We have $\epsilon_1 = \sum_n w_t(1) I(y_n \neq h_1(x_n))$ and $h_1(x_n) = sign(w_1^T x_n + b)$, The optimal weak learner at this iteration is given by $h_1(x) = argmin_{h_1(x)} \epsilon_t$. Since we are usinga hard-margin linear SVM to train on lienarly separable data, the SVM will be able to perfectly classify the data. Therefore, the minimum error $\epsilon_1 = 0$. Plugging this into the formula for β_t we see that $\beta_t \to \infty$. In practice, we may want to make sure that we are indeed using weak classifiers that don't obtain 0 error, or add a small delta to the denominator of β_t to make sure that we don't run into divide by zero issues.

2 Problem 2

Problem 2a

Solution: From the lecture, we have

$$\mu_k = \frac{\sum_n z_{nk} x_n}{\sum_n z_{nk}}$$

This is equivalent to

$$\frac{1}{\sum_n z_{nk}}(z_{1k}x_1 + z_{2k}x_2 + ... + z_{nk}x_n)$$

$$= \frac{z_{1k}x_1}{\sum_n z_{1k}} + ... + \frac{z_{nk}x_n}{\sum_n z_{nk}} = \sum_n \frac{z_{nk}}{\sum_n z_{nk}} x_n$$

Since we have

$$\mu_k = \sum_{n=1}^{N} \alpha_{nk} x_n$$

, we obtain

$$\alpha_{nk} = \frac{z_{nk}}{\sum_n z_{nk}}$$

Problem 2b

Solution:

$$||x_1 - x_2||^2 = ||x_1||^2 + ||x_2||^2 - 2x_1 \cdot x_2$$

$$= x_1^T x_1 + x_2^T x_2 - 2x_1^T x_2 =< x_1, x_1 > + < x_2, x_2 > -2 < x_1, x_2 >$$

Now the inner products may be replaced with a kernel.

Problem 2c

Solution:

$$||x_n - \mu_k||^2 = ||x_n - \sum_{i=1}^{N} \alpha_{ik} x_i||^2$$

=

$$x_n^T x_n + (\sum_{i=1}^{N} \alpha_{ik} x_i)^T (\sum_{i=1}^{N} \alpha_{ik} x_i) - 2x_n \cdot \sum_{i=1}^{N} \alpha_{ik} x_i$$

=

$$< x_n, x_n > + \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_{ik} \alpha_{jk} < x_i, x_j > -2 \sum_{i=1}^{N} \alpha_{ik} < x_n, x_i >$$

3 Problem 3

Problem 3a

Solution: We have the clusters $k \in 1...3$ and $x_1 = 1, x_2 = 2, x_3 = 5, x_4 = 7$. The optimal clustering for this data is then $A(1) = A(2) = 1$, $A(5) = 2$, $A(7) = 3$. 1 and 2 are assigned to cluster 1, 5 is assigned to cluster 2, and 7 is assigned to cluster 3. Then, $\mu_1 = 1.5$, $\mu_2 = 5$, $\mu_3 = 7$. The value of the objective function is then $(1-1.5)^2 + (2-1.5)^2 + (5-5)^2 + (7-7)^2 = \frac{1}{2^2} + \frac{1}{2^2} = \frac{1}{2}$.

Problem 3b

Solution: A possible initialization is $A(x_3) = A(x_4) = 1$, $A(x_2) = 2$, $A(x_1) = 3$. Namely, 5 and 7 are assigned to cluster 1, 2 is assigned to cluster 2, and 1 is assigned to cluster 3. Then, $\mu_1 = 6$, $\mu_2 = 2$, $\mu_3 = 1$.
The value of the objective function is then $(5-6)^2 + (7-6)^2 + (1-1)^2 + (2-2)^2 = 2$ which is suboptimal since $2 > 0.5$ which was the value for our objective function for the assignments in part a.
After the values for $\mu_k, k \in 1...3$ are computed, the points will be reassigned to clusters based on which μ_k the point is closest to. However, each point is closest to the μ_k that corresponds to the cluster it is assigned to: 5 and 7 are closest to $\mu_1 = 6$, 1 is closest to $\mu_3 = 1$, and 2 is closest to $\mu_2 = 2$. So, the assignment of points to clusters will not clusters will not change, even though there is a better setting of cluster means and assignemnts that produce a lower objective fucntion value.

4 Problem 4

Problem 4a

Solution: The transition probabilities $q_{21} = P(q_{t+1} = 2|q_t = 1)$ and $q_{22} = P(q_{t+1} = 2|q_t = 2)$ are missing. Since $\sum_k q_{k1} = 1$ and $\sum_k q_{k2} = 1$, both q_{22} and q_{21} are 0.
The output probabilities $e_2(A) = P(O_t = A|q_t = 2)$ and $e_1(B) = P(O_t = B|q_t = 1)$ are also missing. We know that we must always output $O_t = A$ or $O_t = B$ with some probabilities regardless of the state of q_t. So if $q_t = 1$ then we require $P(O_t = B|q_t = 1) + P(O_t = A|q_t = 1) = 1$, so $e_1(B) + e_1(A) = 1$ and $e_1(B) = 1 - .99 = 0.01$.
Similarly, if $q_t = 2$ we must still emit either A or B. So, $P(O_t = B|q_t = 2)+P(O_t = A|q_t = 2) = 1$, so $e_2(A)+e_2(B) = 1$ and $e_2(A) = 1-0.51 = 0.49$.

Overall, we have $q_{22} = q_{21} = 0$, $e_2(A) = 0.49$, and $e_1(B) = 0.01$.

Problem 4b

Solution: We have

$$P(O_1 = A) = \sum_k P(O_1 = A, q_1 = k) = P(O_1 = A, q_1 = 1)+P(O_1 = A, q_1 = 2)$$

Applying the chain rule, this is

$$P(O_1 = A) = P(O_1 = A|q_1 = 1)P(q_1 = 1) + P(O_1 = A|q_1 = 2)P(q_1 = 2)$$

$$= e_1(A)\pi_1 + e_2(A)\pi_2 = (0.99)(0.49) + (0.49)(0.51) = 0.735$$

Similarly, we have

$$P(O_1 = B) = \sum_k P(O_1 = B, q_1 = k)$$

Which can be written using the same methodology as above as

$$P(O_1 = B) = P(O_1 = B|q_1 = 1)P(q_1 = 1) + P(O_1 = B|q_1 = 2)P(q_1 = 2)$$

$$= e_1(B)\pi_1 + e_2(B)\pi_2 = (0.01)(0.49) + (0.51)(0.51) = 0.265$$

Since $P(O_1 = A) > P(O_1 = B)$, A is the most frequent output symbol to appear in the first position of sequences that this model generates.

Problem 4c

Solution: We are interested in maximizing the joint probability $P(O_1 : 3, q_1 : 3)$ Since the probabilities indicate that we either start in state 1

and then stay in state 1, or start in state 2 and immediately go to state 2, we can compute $P(O_1 : 3, 111) and P(O_1 : 3, 211)$ for each of the 8 possible output states, and take the maximum. In general, we also have $P(y_1 : T, x_1 : T) = \prod_{t=1}^{T} p(y_t|x_t)\pi_{x_1} \prod_{t=1}^{T} q_{x_{t+1},x_t}$.

The probability of observing $AAA = P(AAA, 111) + P(AAA, 211)$

$$P(AAA, 111) = P(A|1)P(A|1)P(A|1)\pi_1 q_{11} q_{11} = (0.99)^3(0.49)$$

.

$$P(AAA, 211) = (0.49)(0.99)^2(0.51)$$

similarly. Therefore, the probability of observing the sequence of AAA is $(0.99)^3(0.49) + (0.99)^2(0.49)(0.51)$.

Considering the values for our emission, initial, and transition probabilities, we can see $P(A|1)$ is the largest, there will always be a contribution from both π_1, π_2 in all cases, and the transition probabilities will not change the probability since they are always 1. So picking $P(A|1)$ maximizes our joint probability since it is the highest. As an example as another sequence we can compute the probability of, and show it is less, we can consider:

$$P(AAB) = P(AAB|111)+P(AAB|211) = (0.99)^2(0.01)(0.49)+(0.99)(0.49)(0.01)(0.51)$$

which is less than P(AAA).

We can compute all eight sequences symbols, but since $P(A|1)$ is largest at each step, the sequence AAA clearly has the highest probability of being observed.