(1)

a) $-1 \leq \cos(y-t) \leq +1 \Rightarrow \begin{cases} \hat{L}_{max} = 2 \Rightarrow \cos(y-t) = -1 \Rightarrow y = \frac{7\pi}{4} \\ \hat{L}_{min} = 0 \Rightarrow \cos(y-t) = 1 \Rightarrow y = \frac{3\pi}{4} \end{cases}$

b) $$\frac{\partial L}{\partial w} = \frac{\partial (1-\cos(w^T x + b - t))}{\partial w} = x \sin(w^T x + b - t)$$

$$\frac{\partial L}{\partial b} = \frac{\partial (1-\cos(w^T x + b - t))}{\partial w} = \sin(w^T x + b - t)$$

c) The batch cost function for N input data is written by:

$$\ell(y,t) = \sum_{i=1}^{N} L(y_i, t_i) = \sum_{i=1}^{N} 1 - \cos(y_i - t_i)$$

$$\frac{\partial \ell}{\partial w} = \sum_{i=1}^{N} x_i \sin(w^T x_i + b - t_i)$$

$$\frac{\partial \ell}{\partial b} = \sum_{i=1}^{N} \sin(w^T x_i + b - t_i)$$

Gradient Descent:

1) initialize $\hat{w}$, $\hat{b}$, γ

2) until convergence:

update $\hat{w} \leftarrow \hat{w} - \gamma \frac{\partial \ell}{\partial w}$

update $\hat{b} \leftarrow \hat{b} - \gamma \frac{\partial \ell}{\partial b}$

(2)-1

* For simplicity, we assume two categories of age: $\begin{cases} \text{young if age} \leq 30 \\ \text{senior if age} > 30 \end{cases}$

a) For Bayes classifier, see Lecture 4b.

* we assume an uniform prior distribution over class labels $\Rightarrow$

$$p(t_{new}=k \mid X, t, x_{new}) = \frac{P(x_{new} \mid t_{new}=k, X, t)\ P(t_{new}=k)}{\sum_j P(x_{new} \mid t_{new}=j, X, t)\ P(t_{new}=j)}$$

$$= \frac{p(x_{new} \mid t_{new}=k, X, t)}{\sum_j p(x_{new} \mid t_{new}=j, X, t)}$$

$$p(t_{new}=L \mid X, t, x_{new}=(young, sports)) = \frac{p(x_{new} \mid t_{new}=L, X, t)}{\sum_{j \in \{L,H\}} p(x_{new} \mid t_{new}=j, X, t)}$$

$$= \frac{2/2}{2/2 + 1/4} = 4/5$$

$$p(t_{new}=H \mid X, t, x_{new}=(young, sports)) = \frac{p(x_{new} \mid t_{new}=H, X, t)}{\sum_{j \in \{H,L\}} p(x_{new} \mid t_{new}=j, X, t)}$$

$$= \frac{1/4}{2/2 + 1/4} = 1/5$$

b) Naive Bayes assumes the independence of components of x_{new} given the class label, i.e.,

$$p(x_{new} \mid X, t, t_{new}=k) = \prod_{d=1}^{D} p(x_d^{new} \mid X, t, t_{new}=k)$$

thus:

$$\begin{cases} p(x_{new} \mid t_{new}=L, X, t) = p(X_{(1)}^{new} = young \mid t_{new}=L, X, t)\, p(X_{(2)}^{new} = sports \mid t_{new}=L, X, t) \\ \qquad = 2/2 \times 2/2 = 1 \\ p(x_{new} \mid t_{new}=H, X, t) = p(X_{(1)}^{new} = young \mid t_{new}=H, X, t)\, p(X_{(2)}^{new} = sports \mid t_{new}=H, X, t) \\ \qquad = 3/4 \times 1/4 = 3/16 \end{cases}$$

$$\Rightarrow \begin{cases} p(t_{new}=L \mid X, t, x_{new}) = \dfrac{1}{1+3/16} = 16/19 \\ p(t_{new}=H \mid X, t, x_{new}) = \dfrac{3/16}{1+3/16} = 3/19 \end{cases}$$

c) The training data is not sufficient to include data of type track.

In the naive-Bayes, we can write/assume:

$$p(X_{(2)}^{new} = track \mid X, t, t_{new}=L) = p(X_{(2)}^{new} = track \mid X, t, t_{new}=H)$$

$$\Rightarrow p(t_{new}=L \mid X, t, x_{new}) = \frac{2/2}{2/2 + 3/4} = \frac{4}{7}$$

$$p(t_{new}=H \mid X, t, x_{new}) = \frac{3/4}{2/2 + 3/4} = \frac{3}{7}$$

(3)

a) $O(nkd) \Longrightarrow$ the method requires a full pass of data for each center to be sampled.

b, c, d): there are several solutions/possibilities, for instance:

in each iteration $j \in [2, 3, \dots k]$ we construct a Markov chain of length m using Metropolis-Hasting method.

The proposal dist. is defined to be an independent and uniform distribution $q(x) = 1/N$.

So with start with a random initial state x_1 and in each iteration $i \in [2, \dots, m]$ we sample a candidate y_i using $q(x)$.

We accept this candidate (i.e., $x_i = y_i$) with probability

$$\min\left(\frac{p(y_i)}{p(x_{i-1})}\,\frac{q(x_{i-1})}{q(y_i)}, 1\right) = \min\left(\frac{d^2(y_i, C)}{d^2(x_{i-1}, C)}, 1\right)$$

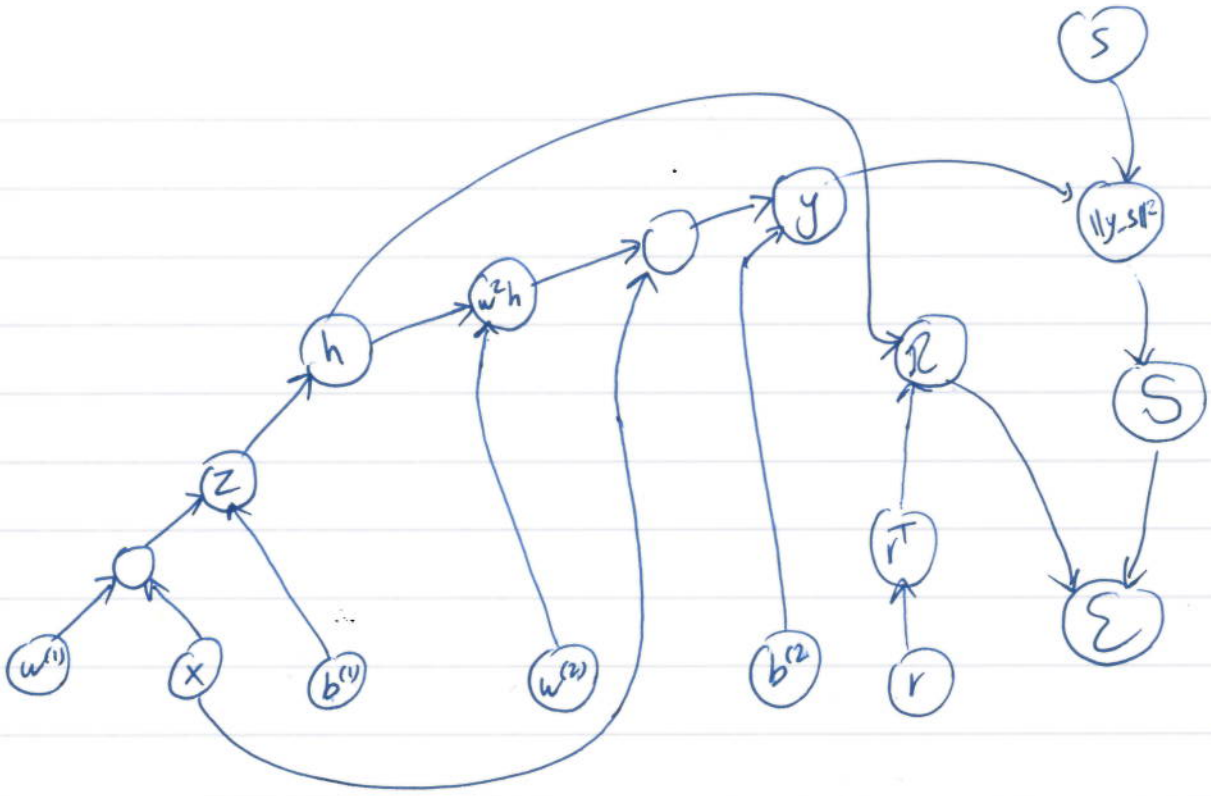
Then, in each of the $k-1$ iterations, we only need to compute the distances between m objects and $k-1$ centers.

Thus the complexity would be $O(mk^2d)$.

* for a very detailed & accurate solution see:
[out of the scope of the cours and the exam]
las.inf.ethz.ch/files/bachem16fast.pdf
or search: Fast and Provably Good seeding for k-means.

(4)

a)



b) $N*k + k + N*k + N$

c) we use chain rule in backpropagation, for example:

$$\frac{\partial \Sigma}{\partial w^{(1)}} = \frac{\partial R}{\partial w^{(1)}} + \frac{\partial S}{\partial w^{(1)}}$$

$$\frac{\partial R}{\partial w^{(1)}} = \frac{\partial R}{\partial h}\frac{\partial h}{\partial w^{(1)}} = \frac{\partial R}{\partial h}\frac{\partial h}{\partial z}\frac{\partial z}{\partial w^{(1)}}$$

$$\frac{\partial S}{\partial w^{(1)}} = \frac{\partial S}{\partial y}\frac{\partial y}{\partial w^{(1)}} = \frac{\partial S}{\partial y}\frac{\partial y}{\partial h}\frac{\partial h}{\partial w^{(1)}} = \frac{\partial S}{\partial y}\frac{\partial y}{\partial h}\frac{\partial h}{\partial z}\frac{\partial z}{\partial w^{(1)}}$$

Similarly:

$$\frac{\partial \Sigma}{\partial w^{(2)}} = \frac{\partial R}{\partial w^{(2)}} + \frac{\partial S}{\partial w^{(2)}}$$

$$\frac{\partial R}{\partial w^{(2)}} = \frac{\partial R}{\partial h}\frac{\partial h}{\partial w^{(2)}} \to 0$$

$$\frac{\partial S}{\partial w^{(2)}} = \frac{\partial S}{\partial y}\frac{\partial y}{\partial w^{(2)}}$$

similarly we can obtain $\frac{\partial \Sigma}{\partial b^{(1)}}$ and $\frac{\partial \Sigma}{\partial b^{(2)}}$.