UNIVERSITY OF CALIFORNIA, LOS ANGELES
DEPARTMENT OF COMPUTER SCIENCE

CS 276A assignment 2

Yang Pei
304434922

December 14, 2014

1 PROBLEM 1

After the re-weight, all examples classified correctly have weight $w_{n+1}(x) = w_n(x) \times e^{-\alpha_+}$ and those classified wrong have weight $w_{n+1}(x) = w_n(x) \times e^{\alpha_+}$. We could divide the entire sample into two set $\Omega = A, \bar{A}$ correspoding to the samples classified correctly and wrong. Since the total sample weight sum to 1, we could have

$$\sum_{x \in A} w_{n+1} + \sum_{x \in \bar{A}} w_{n+1} = \sum_{x \in A} w_n \times e^{-\alpha_+} + \sum_{x \in \bar{A}} w_n \times e^{\alpha_+} = 1 \tag{1.1}$$

We notice that $\sum_{x \in A} w_n = 1 - err_n(h_+)$ and $\sum_{x \in \bar{A}} = err_n(h_+)$. Then, we could obtain

$$(1 - err_n(h_+))e^{-\alpha_+} + err_n(h_+)e^{\alpha_+} = 1 \tag{1.2}$$

Then we plugin $\alpha_+ = \frac{1}{2} log \frac{1-err_n(h_+)}{err_n(h_+)}$ into the equation 1.2 and obtain

$$2\sqrt{(1 - err_n(h_+))err_n(h_+)} = 1 \tag{1.3}$$

From 1.3 we know that $\sqrt{(1 - err_n(h_+))err_n(h_+)} = \frac{1}{2}$. We know that

$$err_{n+1}(h_+) = err_n(h_+)e^{\alpha_+} = \sqrt{(1 - err_n(h_+))err_n(h_+)} \tag{1.4}$$

Combine 1.4 and the result from 1.3 we have $err_{n+1}(h_+) = \frac{1}{2}$, therefor $h_+()$ won't be selected in the next step.

2 PROBLEM 2

1. We could have the following deduction:

$$\begin{aligned} -\int p(x)\log p(x)dx &= -\int p(x)\log(\frac{p(x)}{unif(x)}unif(x))dx \\ &= -\int p(x)\log(\frac{p(x)}{unif(x)})dx - \int p(x)\log(unif(x))dx \\ &= -KL(p(x)||unif(x)) - \int p(x)\log(unif(x))dx \end{aligned} \tag{2.1}$$

   Since the $unif(x)$ is uniform distribution, $unif(x) = \frac{1}{N}$ and it is a constant. And since we use model $p(x)$ to estimate the underlaying distribution $f(x)$, $p(x)$ must be convergent that is $\int p(x)dx = c$ where c is some finiate constant. So, the second term in 2.1 is constant. Thus we could re-write

$$\begin{aligned} p^* = arg\max -\int p(x)\log p(x)dx &= arg\max C - KL(p(x)||unif(x)) \\ &= arg\max -KL(p(x)||unif(x)) \end{aligned} \tag{2.2}$$

   which means the maximum entropy principle is the same with minimizing the Kullback-Leibler divergence to the uniform distribution.

2. If we constraint that $p(x)$ normalizes to 1, then we are solving the problem with a constraint and thus we could utilize Lagrange Multiplier. We plugin the constraints and obtain

$$L = -\int p(x)\log p(x)dx + \lambda_0(\int p(x)dx - 1) = 0 \tag{2.3}$$

   then we have

$$\frac{\partial L}{\partial p(x)} = \int -\log p(x) + \lambda_0 - 1 dx = 0 \tag{2.4}$$

   That is to say $-\log P(x) + \lambda_0 - 1 = 0$, and we have $p(x) = e^{\lambda_0 - 1}$ Since we know $\int p(x)dx = 1$, then we know $e^{\lambda_0 - 1}$ must be bounded and hence it should be 1. Thus $p(x) = \frac{1}{Z}$ where Z is a term used to normalize the distribution and the optimal probability $p^*(x)$ is uniform distribution.

   From the conclusion above, we know that the uniform distribution has the maximum distribution and is the solution for maximum entropy. Kullback-Leibler divergence measure the "distance" or similarity between to distribution. If the two distributions are all same, then they would have 0 KL divergence. So minimizing KL divergence with unifrom distribtuion, we are try to find the distribution that is the most possbile same with uniform distribution, which is the same goal as in maximum entropy.

3 PROBLEM 3

1. Given the data set D, we could have $P(D|\Theta) = \prod_{i=1}^{N} p(y_i;\Theta)$. So the log-likelihood function $\ell(\Theta)$ for the observed data D is

$$\ell(\Theta) = \sum_{i=1}^{N}(logP(y_i;\Theta)) = \sum_{j=1}^{K} n_i \log\theta_i \tag{3.1}$$

2. We consider the class i and take partial derivation. We could have the following deduction:

$$\frac{\partial\ell(\Theta)}{\partial\theta_i} = \sum_{i=1}^{N}\frac{\partial logP(y_i;\Theta)}{\theta_i} = \sum_{i=1}^{N}\frac{1}{P(y_i;\Theta)}\frac{\partial P(y_i;\Theta)}{\partial\theta_i} = 0 \tag{3.2}$$

   Since we could separate the whole data into two parts: labeled with i and not i. For the sample labeled with i we have $P(y_i;\Theta) = \theta_i$ and for the sample not labeled with i we have $P(y_i;\Theta) = 1-\theta_i$. So we could get

$$\sum_{y=i}\frac{1}{P(y_i;\Theta)} - \sum_{y\neq i}\frac{1}{P(y_i;\Theta)} = 0 \tag{3.3}$$

   Then we plugin $p(y=i;\Theta) = \theta_i$ and could obtain $\frac{n_i}{\theta_i} - \frac{n-n_i}{1-\theta_i} = 0$ and obtian $\theta_i = \frac{n_i}{n}$.

4 PROBLEM 4

1. The log-likelihood function $\ell(\Theta)$ would be

$$\begin{aligned}\ell(\Theta) &= \log\prod_{i=1}^{N} p(x_i;\Theta) = \sum_{i=1}^{N}\log p(x_i;\Theta)\\ &= -N\log Z + \sum_{i=1}^{N}\sum_{j=1}^{K}\lambda_j\phi_j(x)\end{aligned} \tag{4.1}$$

2. Since Z is used to normalize the probablity, we know $Z = \sum_{i=1} Ne^{-\sum_{j=1}K\lambda_j\phi_j(x_i)}$. Then we could have

$$\begin{aligned}\frac{\partial\ell(\Theta)}{\theta_i} &= -\sum_{i=1}^{N}\phi_j(x_i) - N\frac{\partial Z}{\partial\theta_i}\\ &= -\sum_{i=1}^{N}\phi_j(x_i) + \frac{N}{Z}\sum_{i=1}^{N}e^{-\sum_{j=1}^{K}\lambda_j\phi_j(x_i)}\phi_j(x_i)\\ &= -\sum_{i=1}^{N}\phi_j(x_i) + \frac{N}{Z}\int e^{-\sum_{j=1}^{K}\lambda_j\phi_j(x)}\phi_j(x)dx\\ &= -\sum_{i=1}^{N}\phi_j(x_i) + N\int p(x;\Theta)\phi_j(x)dx = 0\end{aligned} \tag{4.2}$$

Solving the 4.2 we could obtain

$$\int p(x;\Theta)\phi_j(x)dx = \frac{1}{N}\sum_{i=1}^{N}\phi_j(x_i) \tag{4.3}$$

for all $j = 1, 2, ..., K$.

5 PROBLEM 5

1. For $d = 1$ the $l = r$ and for $d = 2$ the $l = r^{\frac{1}{2}}$. With $r = 0.1$, then $l = 0.1$ and $l = 0.3162$ each.

2. Since $l^d = r$, then we have $l = r^{\frac{1}{d}}$. With $d = 100$ and $r = 0.1$, $l = 0.9772$.

3. For $d = 10100$ and $r = 0.01$, then $l = 0.9995$. For $r = 0.1$, then $l = 0.9997$. From the calculation, we could see that with the increase of the dimension, to obtain the same amount informaiton, we need larger length of the hypercube. This means that the points are distributed scatteredly among all dimensions in the space. And we could also see that with large d, we increase a little bit of the length, we increase a quite large amount of the points, this means that the points are laying near the sufaces of the unit hypercube.