CSE 250B: Homework 3

1. Handwritten digit recognition using a Gaussian generative model

(a) description of choosing c:
step 1: grid search. Try c from 10, 100, 1000, 10000 and select best grid.
step 2: setting steps to be smaller and smaller to find best c.
(b) the figure showing the validation error for all the values of c is the following. Detailed data are shown in Table 1.

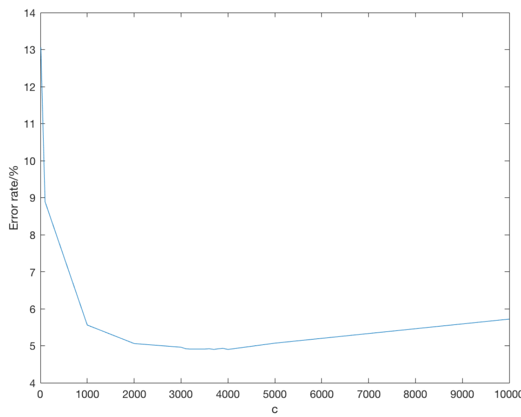

Figure 1: validation error in terms of different c

c	10	100	1000	2000	3000	3100	3200	3300	3400
validation err	13.6	8.89	5.56	5.06	4.96	4.92	4.91	4.91	4.91
c	3500	3600	3700	3800	3900	4000	5000	10000	
validation err	4.91	4.92	4.90	4.92	4.93	4.9	5.07	5.72	

Table 1: Validation Error in terms of different c

(c) By checking Figure 1, we choose $c = 3700$.
(d) The overall error rate on the MNIST test set is 4.44%.

2. Entropy calculation

(a) coin of bias 2/3: we know that $p_0 = 1/3, p_1 = 2/3$.

$$H(P) = \frac{1}{3}log3 + \frac{2}{3}log1.5 = 0.6365 \tag{1}$$

(b) rolling a fair die:

$$H(P) = 6 \times \frac{1}{6} log6 = log6 \tag{2}$$

(c) $X = (X_1, ..., X_{10})$, where the X_i are independent and are each coins of bias 1/2: In this case, there are 2^{10} possible samples in S, each event has probability $1/2^{10}$.

$$H(P) = \frac{1}{2^{10}} log2^{10} \times 2^{10} = log2^{10} = 10log2 \tag{3}$$

(d) $X = (X_1, ..., X_{10})$, where X_1 is a coin of bias 1/2 and $X_i = X_1 + i$:
For X_1, the probability of getting head(1)/tail(0) is 1/2. Once X_1 is determined, X_i where $i > 1$ would automatically be determined because $X_i = X_1 + i$. Hence, there are only two possible samples in S, each has probability 1/2.

$$H(P) = 2 \times \frac{1}{2} log2 = log2 \tag{4}$$

3. Entropy of continuous distributions

(a) one-dimensional Gaussian with mean μ and variance σ^2: we know that the p.d.f. is

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(x-\mu)^2}{2\sigma^2}) \tag{5}$$

Hence,

$$\begin{aligned} h(p) &= \int p(x) log \frac{1}{p(x)} dx = \int p(x) log \frac{1}{\frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(x-\mu)^2}{2\sigma^2})} dx \\ &= -(\int p(x) log \frac{1}{\sqrt{2\pi\sigma^2}} dx + log(e) \int p(x)(-\frac{(x-\mu)^2}{2\sigma^2}) dx) \\ &= -(-\frac{1}{2} log(2\pi\sigma^2) - log(e) \frac{\sigma^2}{2\sigma^2}) = \frac{1}{2} log(2\pi\sigma^2 e) \end{aligned} \tag{6}$$

(b) The differential entropy of the uniform distribution over [a,b] is

$$h(p) = \int p(x) log \frac{1}{p(x)} dx = \int_a^b \frac{1}{b-a} log(b-a) dx = log(b-a) \tag{7}$$

When b-a is small, we can derive $h(p) < 0$, yet entropy can scarcely be negative.

4.

The set S is $S = \{mysterious\ new\ language\ words\}$.
Feature selection:

$$T_1(x) = \mathbb{1}(x\ ends\ with\ a\ vowel)$$
$$T_2(x) = \mathbb{1}(x\ starts\ with\ 'z')$$
$$T_3(x) = \mathbb{1}(x\ has\ property:\ every\ consonant\ is\ followed\ by\ a\ vowel)$$

The functional form of this maximum entropy solution is:
Find $P = (p_x : x \in S)$,

$$maxH(P)\ s.t.$$
$$\sum_x p_x T_i(x) = b_i,\ \ i = 1,2,3\ and\ b_1 = 0.8, b_2 = 0.5, b_3 = 0.9$$
$$p_x \geq 0$$
$$\sum_x p_x = 1$$

5.

S = $\mathbb{R}^+$
$\pi = 1$
T(x) = x
Hence, we can derive

$$p_\theta(x) = \frac{1}{z_\theta} e^{\theta x} \tag{8}$$

where

$$z_\theta = \int_0^{+\infty} e^{\theta x} dx = \frac{1}{\theta} e^{\theta x} \mid_0^{+\infty} \tag{9}$$

From the equation above, we know that θ has to be negative such that z_θ can be finite. Therefore, $\Theta = \mathbb{R}^-$.

6.

$x = (1,2,3), |x| = \sqrt{14}$, hence the unit vector that has the same direction is $i = (1/\sqrt{14}, 2/\sqrt{14}, 3/\sqrt{14})$.

7.

Let i = (x,y) be the unit vector that is orthogonal to (1,1). We have

$$x^2 + y^2 = 1, x + y = 0 \tag{10}$$

Solving these two equation would yield $(x,y) = (1/\sqrt{2}, -1/\sqrt{2}),\ and\ (-1/\sqrt{2}, 1/\sqrt{2})$.

8.

For points in set where $x \cdot x = 25$, the squared value of their distance to the origin of this d-dimension space is 25. In other words, the length of a vector ending at such points is 5.

9.

$w = (2, -1, 6)$.

10.

A has dimensions 10×30, while B has dimensions 30×20.

11.

(a) The dimension of X is $n \times d$.
(b) The dimension of XX^T is $n \times n$.
(c) The (i, j) entry is $x^{(i)} \cdot x^{(j)}$.

12.

Suppose x is column vector, meaning the dimension of x is 10×1. Hence, the dimension of $x^T x x^T x x^T x$ is 1×1. Suppose $x = (x_1, x_2, ..., x_1 0)$, we know $x^T x = \sum_{i=1}^{10} x_i^2$, hence $x^T x x^T x x^T x = (\sum_{i=1}^{10} x_i^2)^3$.

13.

when $x = (1, 3, 5)^T$:

$$xx^T = \begin{bmatrix} 1 & 3 & 5 \\ 3 & 9 & 15 \\ 5 & 15 & 25 \end{bmatrix}, x^T x = 35 \tag{11}$$

14.

$$cos\theta = \frac{x^T y}{|x||y|} = \frac{2}{2 \times 2} = \frac{1}{2} \tag{12}$$

Hence, $\theta = cos^{-1} 0.5 = 60$ degrees.

15.

$$M = \begin{bmatrix} 3 & 1 & -2 \\ 1 & 0 & 0 \\ -2 & 0 & 6 \end{bmatrix} \tag{13}$$

16.

(a)(b)(c) are symmetric; (d) is not symmetric.

17.

(a) $|A| = 8!$
(b) $A^{-1} = diag(1, 1/2, 1/3, 1/4, 1/5/1, 6/1/7, 1/8)$.

18.

(a) Since u_i are orthogonal to each other and have unit length, we know that $u_i \cdot U_j = 0$ for $i \neq j$, and $u_i \cdot U_j = 1$ for $i = j$. Hence, $UU^T = I_{d\times d}$ where $I_{d\times d}$ is an identity matrix of size $d \times d$.
(b) The definition of inverse matrix is the following: $UU^{-1} = I$, from (a) we find that $UU^T = I$, by taking advantage of the fact that inverse matrix is unique, hence, $U^{-1} = U^T$.

19.

Since A is singular, we know that $det(A) = 0$. Hence, $z - 6 = 0 \Rightarrow z = 6$.