CS405 Homework #1

12110644 周思呈

Question 1

Consider the polynomial function:

$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x + \ldots + w_M x^M = \sum_{i=0}^{M} w_i x^i$

Calculate the coefficients $\mathbf{w} = w_i$ that minimize its sum-of-squares error function. Here a suffix i denotes the index of a component, whereas $(x)^i$ denotes x raised to the power of i.

Let $\mathbf{x} = [x^0, x^1, \ldots, x^m]^T$, $\mathbf{w} = [w_0, w_1, \ldots, w_M]$ so $y(x, \mathbf{w}) = \mathbf{wx} = \mathbf{x^T w^T}$.

Assume our target label value is t_n, so we have to minimize

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 \tag{1}$$

where we add coefficient $\frac{1}{2}$ to simplify the derivation process. Let $\mathbf{t} = [t_0, t_1, \ldots, t_n]$, $\mathbf{X} = [\mathbf{x_0}, \mathbf{x_1}, \ldots, \mathbf{x_n}]$.

$$\begin{aligned} \frac{\partial E(\mathbf{w})}{\partial \mathbf{w}} &= \sum_{n=1}^{N} (y_n \frac{\partial y}{\partial \mathbf{w}} - t_n \frac{\partial y}{\partial \mathbf{w}}) \\ &= \sum_{n=1}^{N} \frac{\partial (\mathbf{x^T w^T wx} - 2t_n \mathbf{wx})}{\partial \mathbf{w}} \\ &= \sum_{n=1}^{N} (2\mathbf{wxx^T} - 2t_n \mathbf{x^T}) \\ &= 2(\mathbf{wXX^T} - \mathbf{tX^T}) = 0 \end{aligned} \tag{2}$$

So $\mathbf{w} = \mathbf{tX^T}(\mathbf{XX^T})^{-1}$.

Question 2

Suppose that we have three colored boxes $r(red)$, $b(blue)$, and $g(green)$. Box r contains 3 apples, 4 oranges, and 3 limes, box b contains 1 apple, 1 orange, and 0 limes, and box g contains 3 apples, 3 oranges, and 4 limes. If a box is chosen at random with probabilities $p(r) = 0.2, p(b) = 0.2, p(g) = 0.6$, and a piece of fruit is removed from the box (with equal probability of selecting any of the items in the box), then what is the probability of selecting an apple? If we observe that the selected fruit is in fact an orange, what is the probability that it came from the green box?

Let $p(a)$ indicates selecting an apple, then

$$\begin{aligned} p(a) &= p(a|r)p(r) + p(a|g)p(g) + p(a|b)p(b) \\ &= \frac{3}{3+4+3} \times 0.2 + \frac{3}{3+3+4} \times 0.6 + \frac{1}{1+1} \times 0.2 \\ &= 0.34 \end{aligned} \tag{3}$$

So the probability of selecting an apple is 0.34.

Let $p(o)$ indicates selecting an orange, then according to the Bayes' Law,

$$\begin{aligned} p(g|o) &= \frac{p(g)p(o|g)}{p(o)} \\ &= \frac{0.6 \times 0.3}{0.36} \\ &= \frac{1}{2} \end{aligned} \tag{4}$$

If the selected fruit is an orange, the possibility it came from the green box is 0.5.

Question 3

Given two statistically independent variables x and z, show that the mean and variance of their sum satisfies

$$\begin{aligned}\mathbb{E}[x+z] &= \mathbb{E}[x] + \mathbb{E}[z] \\ \mathrm{var}[x+z] &= \mathrm{var}[x] + \mathrm{var}[z]\end{aligned} \tag{5}$$

Suppose X and Z are continuous.

$$\begin{aligned}E(X+Z) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+z)p(x,z)dxdz \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xp(x,z)dxdz + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}zp(x,z)dxdz \\ &= \int_{-\infty}^{\infty}xp_X(x)dx + \int_{-\infty}^{\infty}zp_Z(z)dz \\ &= E(X) + E(Z)\end{aligned} \tag{6}$$

It is similar in the discrete case.

$$\begin{aligned}E[X+Z] &= \sum_{i,j}(x_i+z_j)P\{X=x_i, Z=z_j\} \\ &= \sum_{i,j}x_iP\{X=x_i, Z=z_j\} + \sum_{i,j}z_jP\{X=x_i, Z=z_j\} \\ &= \sum_{i}x_i\sum_{j}P\{X=x_i, Z=z_j\} + \sum_{j}z_j\sum_{i}P\{X=x_i, Z=z_j\} \\ &= \sum_{i}x_iP\{X=x_i\} + \sum_{j}z_jP\{Z=z_j\} \\ &= E[X] + E[Z]\end{aligned} \tag{7}$$

Variance of variable X is defined as

$$D(X) = E((X-E(X))^2) \tag{8}$$

And it has a property that

$$\begin{aligned}D(X) &= E\left\{[X-E(X)]^2\right\} = E\left\{X^2 - 2XE(X) + [E(X)]^2\right. \\ &= E\left(X^2\right) - 2E(X)E(X) + [E(X)]^2 \\ &= E\left(X^2\right) - [E(X)]^2\end{aligned} \tag{9}$$

Apply it to what we need to prove

$$\begin{aligned}D(X+Z) &= E[(X+Z)^2] - E^2(X+Z) \\ &= E(X^2+2XZ+Z^2) - (E(X)+E(Z))^2 \\ &= E(X^2) + 2E(XZ) + E(Z^2) - E^2(X) - 2E(X)E(Z) - E^2(z) \\ &= D(X) + D(Z) + 2Cov(X,Z)\end{aligned} \tag{10}$$

Because X and Z are independent, we have $Cov(X,Z)=0$, so $D(X+Z) = D(X) + D(Z)$.

Question 4

In probability theory and statistics, the Poisson distribution, is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant rate and independently of the time since the last event. If X is Poisson distributed, i.e. $X \sim Possion(\lambda)$, its probability mass function takes the following form:

$P(X|\lambda) = \frac{\lambda^X e^{-\lambda}}{X!}$

It can be shown that if $\mathbb{E}(X) = \lambda$. Assume now we have n data points from $Possion(\lambda) : \mathcal{D} = \{X_1, X_2, \ldots, X_n\}$. Show that the sample mean $\widehat{\lambda} = \frac{1}{n}\sum_{i=1}^{n}X_i$ is the maximum likelihood estimate(MLE) of λ.

If X is exponential distribution and its distribution density function is $f(x) = \frac{1}{\lambda}e^{-\frac{x}{\lambda}}$ for $x > 0$ and $f(x) = 0$ for $x \leq 0$. Show that the sample mean $\widehat{\lambda}\frac{1}{n}\sum_{i=1}^{n}X_i$ is the maximum likelihood estimate(MLE) of λ.

The likelihood function is

$$lik(\theta) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \tag{11}$$

If the random variables are independently and identically distributed(IID), we have

$$lik(\theta) = P(X_1 = x_1)P(X_2 = x_2)\ldots P(X_n = x_n) \tag{12}$$

Consider random variables in Poisson distribution

$$L(x|\lambda) = \frac{\lambda^{\sum_{i=1}^n x_i} e^{-n\lambda}}{\prod_{i=1}^n x_i!} \tag{13}$$

We want to find out what will λ be when L is the smallest, so we take the log of L and take the partial derivative with respect to λ

$$\frac{\partial L(x|\lambda)}{\partial \lambda} = \sum_{i=1}^n x_i \frac{1}{\lambda} - n = 0 \tag{14}$$

So we have $\widehat{\lambda} = \frac{1}{n}\sum_{i=1}^n X_i$.

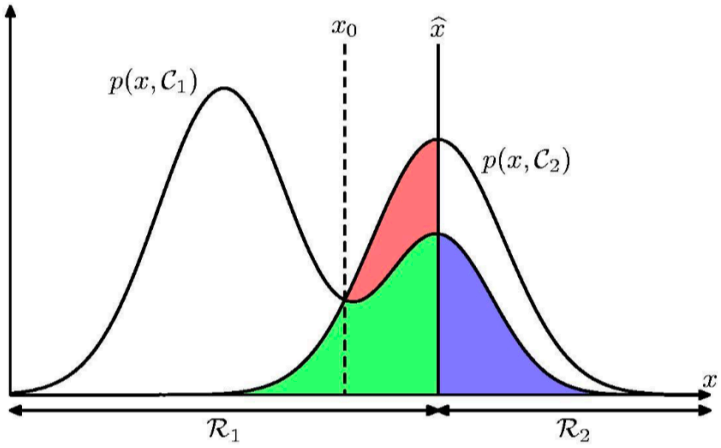
If X is exponential distribution, consider when $x > 0$

$$\begin{aligned} L(x|\lambda) &= (\frac{1}{\lambda})^n e^{-\frac{1}{\lambda}\sum_{i=1}^n x_i} \\ \frac{\partial L(x|\lambda)}{\partial \lambda} &= -n\lambda^{-1} + \sum_{i=1}^n \lambda^{-2} = 0 \end{aligned} \tag{15}$$

So we have $\widehat{\lambda} = \frac{1}{n}\sum_{i=1}^n X_i$.

Question 5

(a) **Write down the probability of classifying correctly $p(correct)$ and the probability of misclassification $p(mistake)$ according to the following chart.**



$$\begin{aligned} p(\text{ mistake }) &= p\left(\boldsymbol{x} \in \mathcal{R}_1, \mathcal{C}_2\right) + p\left(\boldsymbol{x} \in \mathcal{R}_2, \mathcal{C}_1\right) \\ &= \int_{\mathcal{R}_1} p\left(\boldsymbol{x}, \mathcal{C}_2\right) \mathrm{d}\boldsymbol{x} + \int_{\mathcal{R}_2} p\left(\boldsymbol{x}, \mathcal{C}_1\right) \mathrm{d}\boldsymbol{x} \end{aligned} \tag{16}$$

$$p(\text{ correct }) = \sum_{k=1}^K p\left(\boldsymbol{x} \in \mathcal{R}_k, \mathcal{C}_k\right) = \sum_{k=1}^K \int_{\mathcal{R}_k} p\left(\boldsymbol{x}, \mathcal{C}_k\right) \mathrm{d}\boldsymbol{x} \tag{17}$$

(b) **For multiple target variables described by vector $\mathbf{t}$, the expected squared loss function is given by**

$\mathbb{E}[L(\mathbf{t}, \mathbf{y}(\mathbf{x}))] = \int\int \|\mathbf{y}(\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{x}, \mathbf{t}) \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{t}$

Show that the function $\mathbf{y}(\mathbf{x})$ for which this expected loss is minimized given by $\mathbf{y}(\mathbf{x}) = \mathbb{E}_{\mathbf{t}}[\mathbf{t}|\mathbf{x}]$.

> **Hints**
>
> For a single target variable t, the loss is given by
>
> $\mathbb{E}[L] = \int\int \{y(\mathbf{x}) - t\}^2 p(\mathbf{x}, t) \mathrm{d}\mathbf{x}\mathrm{d}\mathrm{t}$
>
> The result is as follows
>
> $y(\mathbf{x}) = \frac{\int tp(\mathbf{x},t)\mathrm{d}\mathrm{t}}{p(\mathbf{x})} = \int tp(t|\mathbf{x})\mathrm{d}\mathrm{t} = \mathbb{E}_t[t|\mathbf{x}]$

$$\frac{\delta \mathbb{E}(L)}{\delta \mathbf{y}(\mathbf{x})} = 2[\int (y_1(\mathbf{x}) - t)p(\mathbf{x}, t)dt, \ldots, \int (y_n(\mathbf{x}) - t)p(\mathbf{x}, t)dt] = 0$$
$$y_i(\mathbf{x}) = \frac{\int tp(\mathbf{x}, t)dt}{p(\mathbf{x})} = E[t|\mathbf{x}] \tag{18}$$

So we have $\mathbf{y(x)} = \mathbb{E}_\mathbf{t}[\mathbf{t}|\mathbf{x}]$.

Question 6

(a) **We defined the entropy based on a discrete random variable $\mathbf{X}$ as**

$\mathbf{H}[\mathbf{X}] = -\sum_i p(x_i)\mathrm{ln}p(x_i)$

Now consider the case that $\mathbf{X}$ is a continuous random variable with the probability density function $p(x)$. The entropy is defined as

$\mathbf{H}[\mathbf{X}] = -\int p(x)\mathrm{ln}p(x)dx$

Assume that $\mathbf{X}$ follows Gaussian distribution with the mean μ and variance σ, i.e.

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

Please derive its entropy $\mathbf{H}[\mathbf{X}]$.

$$\begin{aligned}
H[x] &= -\int \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \ln \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx \\
&= -\frac{1}{(2\pi\sigma^2)^{1/2}} \int \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \left(-\ln(\sqrt{2\pi}\sigma) - \frac{(x-\mu)^2}{2\sigma^2}\right) dx \\
&= -\frac{1}{(2\pi\sigma^2)^{1/2}} \cdot -\ln(\sqrt{2\pi}\sigma) \int \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx + \frac{1}{(2\pi\sigma^2)^{1/2}} \int \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \frac{(x-\mu)^2}{2\sigma^2} dx \\
&= \frac{\ln(\sqrt{2\pi}\sigma)}{(2\pi\sigma^2)^{1/2}} \int \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx + \frac{1}{(2\pi\sigma^2)^{1/2}} \int \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \frac{(x-\mu)^2}{2\sigma^2} dx \\
&= \frac{\ln(\sqrt{2\pi}\sigma)}{(2\pi\sigma^2)^{1/2}} \sqrt{2}\sigma \int \exp\left\{-\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)^2\right\} d\left(\frac{x-\mu}{\sqrt{2}\sigma}\right) + \frac{1}{(2\pi\sigma^2)^{1/2}} \sqrt{2}\sigma \int \exp\left\{-\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)^2\right\} \frac{(x-\mu)^2}{2\sigma^2} d\left(\frac{x-\mu}{\sqrt{2}\sigma}\right) \\
&= \frac{\ln(\sqrt{2\pi}\sigma)}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-y^2} dy + \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-y^2} y^2 dy \\
&= \ln(\sqrt{2\pi}\sigma) + \frac{1}{\sqrt{\pi}} \cdot -\frac{1}{2}\left(0 - \int_{-\infty}^{\infty} e^{-y^2} dy\right) \\
&= \ln(\sqrt{2\pi}\sigma) + \frac{1}{2} \\
&= \frac{1}{2}\left(\ln\left(2\pi\sigma^2\right) + 1\right)
\end{aligned}$$

(b) **Write down the mutual information $\mathbf{I(y, x)}$. Then show the following equation**

$\mathbf{I}[\mathbf{x}, \mathbf{y}] = \mathbf{H}[\mathbf{x}] - \mathbf{H}[\mathbf{x}|\mathbf{y}] = \mathbf{H}[\mathbf{y}] - \mathbf{H}[\mathbf{y}|\mathbf{x}]$

Mutual information is the relative entropy of the joint distribution and the product distribution. Consider the continuous situation

$$
\begin{aligned}
\mathbf{I}[\boldsymbol{x}, \boldsymbol{y}] &\equiv \mathrm{KL}(p(\boldsymbol{x}, \boldsymbol{y}) \| p(\boldsymbol{x}) p(\boldsymbol{y})) \\
&= -\iint p(\boldsymbol{x}, \boldsymbol{y}) \ln \left(\frac{p(\boldsymbol{x}) p(\boldsymbol{y})}{p(\boldsymbol{x}, \boldsymbol{y})} \right) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{y} \\
\mathbf{H}[\boldsymbol{x}] &= -\int p(\boldsymbol{x}) \ln p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\
\mathbf{H}[\boldsymbol{y}] &= -\int p(\boldsymbol{y}) \ln p(\boldsymbol{y}) \mathrm{d}\boldsymbol{y} \\
\mathbf{I}[\boldsymbol{x}, \boldsymbol{y}] &= \iint p(\boldsymbol{x}, \boldsymbol{y}) \ln \left(\frac{p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{y})} \right) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{y} - \int p(\boldsymbol{x}) \ln p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\
&= \iint p(\boldsymbol{y}) p(\boldsymbol{x}|\boldsymbol{y}) \ln \left(p(\boldsymbol{x}|\boldsymbol{y}) \right) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{y} + \mathbf{H}[\boldsymbol{x}] \\
&= -\mathbf{H}[\boldsymbol{x}|\boldsymbol{y}] + \mathbf{H}[\boldsymbol{x}]
\end{aligned} \tag{19}
$$

By analogy, we have $\mathbf{I}[\mathbf{x}, \mathbf{y}] = \mathbf{H}[\mathbf{y}] - \mathbf{H}[\mathbf{y}|\mathbf{x}]$.