CS329 Homework #1

Course: Machine Learning(H)(CS329) - Instructor: Qi Hao

Name: Jianan Xie(谢嘉楠)

SID: 12110714

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.

Ans:

The sum-of-squares error function of this polynomial fitting function is $E(w) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, w) - y_n\}^2$, and here we denote that $\mathbf{w} = [w_0, w_1, \ldots, w_M]^T$, $\mathbf{x_i} = [1, x_i, x_i^2, \ldots, x_i^M]^T$, $\mathbf{y} = [y_1, y_2, \ldots, y_N]^T$ and $\mathbf{X} = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{bmatrix}$. Thus, we rewrite $E(w)$ as following format:

$$\begin{aligned} E(w) &= \frac{1}{2}\sum_{n=1}^{N}\{w^T x_n - y_n\}^2 = \frac{1}{2}\sum_{n=1}^{N}\{y_n - x_n^T w\}^2 \\ &= \frac{1}{2}\begin{bmatrix} y_1 - x_1^T w & \ldots & y_N - x_N^T w \end{bmatrix}\begin{bmatrix} y_1 - x_1^T w \\ y_2 - x_2^T w \\ \vdots \\ y_N - x_N^T w \end{bmatrix} \\ &= \frac{1}{2}(\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w}) \end{aligned}$$

then we need to find the $\mathbf{w}$ to minimize $E(w)$, that is when $\frac{\partial E(w)}{\partial w} = 0$. We should know some matrix differential formula.:

$$\frac{\partial a^T x}{\partial x} = \frac{\partial x^T a}{\partial x} = a \tag{1}$$

$$\frac{\partial x^T A x}{\partial x} = (A + A^T)x \tag{2}$$

next we find the $\hat{\mathbf{w}}$ to make $\frac{\partial E(\hat{w})}{\partial \hat{w}} = 0$

$$\begin{aligned} \frac{\partial E(w)}{\partial w} &= \frac{1}{2}\frac{\partial(\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w})}{\partial w} \\ &= \frac{1}{2}\frac{\partial(\mathbf{y^T}\mathbf{y} - \mathbf{w^T}\mathbf{X^T}\mathbf{y} - \mathbf{y^T}\mathbf{X}\mathbf{w} + \mathbf{w^T}\mathbf{X^T}\mathbf{X}\mathbf{w})}{\partial w} \\ &= \frac{1}{2}(\frac{\partial \mathbf{y^T}\mathbf{y}}{\partial w} - \frac{\partial \mathbf{w^T}\mathbf{X^T}\mathbf{y}}{\partial w} - \frac{\partial \mathbf{y^T}\mathbf{X}\mathbf{w}}{\partial w} + \frac{\partial \mathbf{w^T}\mathbf{X^T}\mathbf{X}\mathbf{w}}{\partial w}) \\ &= \frac{1}{2}(0 - \mathbf{X^T}\mathbf{y} - \mathbf{X^T}\mathbf{y} + (\mathbf{X^T}\mathbf{X} + \mathbf{X^T}\mathbf{X})\mathbf{w}) \qquad (\text{becuase of } (1) \text{ and } (2)) \\ &= \mathbf{X^T}\mathbf{X}\mathbf{w} - \mathbf{X^T}\mathbf{y} \\ &= 0 \end{aligned}$$

So, $\mathbf{X^T X\hat{w} - X^T y} = 0$,then we get $\mathbf{\hat{w}} = \mathbf{(X^T X)^{-1} X^T y}$, if $\mathbf{X^T X}$ is invertible. In other cases, we should choose a suitble $\mathbf{\hat{w}}$,s.t. $\mathbf{X^T X\hat{w} - X^T y} = 0$.

Question 2

Suppose that we have three colored boxes $r(\mathrm{red})$, $b(\mathrm{blue})$, and $g(\mathrm{green})$. 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?

Ans:

$P(apple|r) = 3/(3+4+3) = \frac{3}{10} = 0.3$,

$P(apple|b) = 1/(1+1+0) = \frac{1}{2} = 0.5$,

$P(apple|g) = 3/(3+3+4) = \frac{3}{10} = 0.3.$

so,

$P(apple) = P(apple|r)P(r) + P(apple|b)P(b) + P(apple|g)P(g) = 0.3 \times 0.2 + 0.5 \times 0.2 + 0.3 \times 0.6 = 0.34$

$P(orange|r) = 4/(3+4+3) = \frac{4}{10} = 0.4,$

$P(orange|b) = 1/(1+1+0) = \frac{1}{2} = 0.5,$

$P(orange|g) = 3/(3+3+4) = \frac{3}{10} = 0.3$

so, $P(g|orange) = \frac{P(orange|g)P(g)}{P(orange|r)P(r)+P(orange|b)P(b)+P(orange|g)P(g)} = \frac{0.3\times0.6}{0.4\times0.2+0.5\times0.2+0.3\times0.6} = 0.5.$

Question 3

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

$$\mathbb{E}[x + z] = \mathbb{E}[x] + \mathbb{E}[z]$$

$$\mathrm{var}[x + z] = \mathrm{var}[x] + \mathrm{var}[z]$$

Ans:

- If **x** and **z** are both discrete statistically independent variables.

$$\begin{aligned} E[x+z] &= \sum_x \sum_z (x+z)p(x,z) \\ &= \sum_x \sum_z (x+z)p(x)p(z) \\ &= \sum_x \sum_z (xp(x)p(z) + zp(z)p(x)) \\ &= \sum_x xp(x) + E[z]\sum_x p(x) \\ &= E[x] + E[z] \end{aligned}$$

- If **x** and **z** are both continuous statistically independent variables.

$$
\begin{aligned}
E[x+z] &= \int_x \int_z (x+z) f(x,z) dz dx \\
&= \int_x \int_z (x+z) f(x) f(z) dz dx \\
&= \int_x (x f(x) dx \int_z f(z) dz + f(x) dx \int_z z f(z) dz) \\
&= \int_x (x f(x) dx + E[z] f(x) dx) \\
&= \int_x x f(x) dx + E[z] \int_x f(x) dx \\
&= E[x] + E[z]
\end{aligned}
$$

Next, we prove $var[x+z] = var[x] + var[z]$

$$
\begin{aligned}
var[x+z] &= E\{[(x+z) - E[x+z]]^2\} \\
&= E[(x+z)^2] - (E[x+z])^2 \\
&= E[x^2 + 2xz + z^2] - (E[x] + E[z])^2 \\
&= E[x^2] + 2E[x]E[z] + E[z^2] - (E[x])^2 - 2E[x]E[z] - (E[z])^2 \\
&= E[x^2] - (E[x])^2 + E[z^2] - (E[z])^2 \\
&= var[x] + var[z]
\end{aligned}
$$

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.

Ans:

$X \sim Possion(\lambda)$: $L(\lambda; D) = \prod_{i=1}^n P(X_i|\lambda) = \frac{\lambda^{\sum_{i=1}^n X_i} e^{-n\lambda}}{\prod_{i=1}^n X_i!}$ and
$\ln L(\lambda; D) = \sum_{i=1}^n X_i \ln \lambda - n\lambda - \sum_{i=1}^n \ln X_i!$

$$
\begin{aligned}
\frac{\partial \ln L(\lambda; D)}{\partial \lambda} &= \frac{\partial}{\partial \lambda}(\sum_{i=1}^n X_i \ln \lambda - n\lambda - \sum_{i=1}^n \ln X_i!) \\
&= \sum_{i=1}^n X_i \frac{1}{\lambda} - n \\
&= 0
\end{aligned}
$$

Then we get the MLE of λ is $\hat{\lambda} = \frac{1}{n}\sum_{i=1}^n X_i$

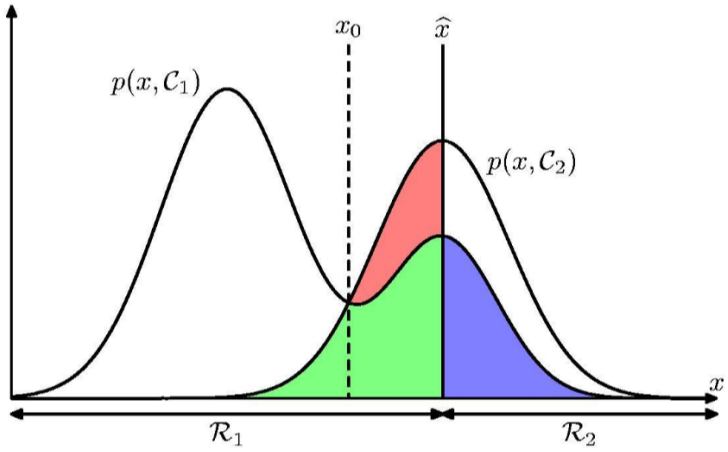
$X \sim P(\lambda)$: $L(\lambda; D) = \prod_{i=1}^n f(X_i) = \frac{1}{\lambda^n} e^{-\frac{\sum_{i=1}^n X_i}{\lambda}}$ and $\ln L(\lambda; D) = -\frac{\sum_{i=1}^n X_i}{\lambda} - n \ln \lambda$

$$
\begin{aligned}
\frac{\partial \ln L(\lambda; D)}{\partial \lambda} &= \frac{\partial}{\partial \lambda}(-\frac{\sum_{i=1}^n X_i}{\lambda} - n \ln \lambda) \\
&= \frac{\sum_{i=1}^n X_i}{\lambda^2} - \frac{n}{\lambda} \\
&= 0
\end{aligned}
$$

Then we get the MLE of λ is $\hat{\lambda} = \frac{1}{n}\sum_{i=1}^n X_i$

Question 5

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



Ans:

$p(correct) = \int_{R_1} p(x, C_1)dx + \int_{R_2} p(x, C_2)dx, p(mistake) = \int_{R_2} p(x, C_1)dx + \int_{R_1} p(x, C_2)dx.$

(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
>
> $\mathbb{E}[L] = \int \int \{y(\mathbf{x}) - t\}^2 p(\mathbf{x}, t) \mathrm{d}\mathbf{x}\mathrm{d}t$
>
> The result is as follows
>
> $y(\mathbf{x}) = \frac{\int t p(\mathbf{x}, t) \mathrm{d}t}{p(\mathbf{x})} = \int t p(t|\mathbf{x}) \mathrm{d}t = \mathbb{E}_t[t|\mathbf{x}]$

Ans:

Our goal is to find a $\mathbf{y}(\mathbf{x})$ so as to minimize $\mathbf{E}[\mathbf{L}(\mathbf{t}, \mathbf{y}(\mathbf{x}))]$, we do this formally using the calculus of variations to give

$$\frac{\delta E[L(\mathbf{t}, \mathbf{y}(\mathbf{x}))]}{\delta \mathbf{y}(\mathbf{x})} = \int 2(\mathbf{y}(\mathbf{x}) - \mathbf{t}) p(\mathbf{t}, \mathbf{x}) d\mathbf{t} = 0$$

Solving the $\mathbf{y}(\mathbf{x})$:

$$\mathbf{y}(\mathbf{x}) = \frac{\int \mathbf{t} p(\mathbf{t}, \mathbf{x}) d\mathbf{t}}{\int p(\mathbf{t}, \mathbf{x}) d\mathbf{t}} = \frac{\int \mathbf{t} p(\mathbf{t}, \mathbf{x}) d\mathbf{t}}{p(\mathbf{x})} = \int \mathbf{t} p(\mathbf{t}|\mathbf{x}) d\mathbf{t} = E_{\mathbf{t}}[\mathbf{t}|\mathbf{x}]$$

Question 6

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

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

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

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

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 $\mathbf{H}[\mathbf{X}]$.

Ans:

$$\begin{aligned}
\mathbf{H}[\mathbf{X}] &= -\int p(x)\ln p(x)dx \\
&= -\int p(x)[-\frac{(x-\mu)^2}{2\sigma^2} - \ln\sqrt{2\pi}\sigma]dx \\
&= \frac{1}{2\sigma^2}\int (x-\mu)^2 p(x)dx + \frac{1}{2}\ln(2\pi\sigma^2)\int p(x)dx \\
&= \frac{1}{2\sigma^2}\times\sigma^2 + \frac{1}{2}\ln(2\pi\sigma^2) && \text{(definition of variance)} \\
&= \frac{1}{2} + \frac{1}{2}\ln(2\pi\sigma^2)
\end{aligned}$$

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

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

Ans:

$$\mathbf{I}[\mathbf{x},\mathbf{y}] \equiv KL(p(x,y)||p(x)p(y)) = -\int\int p(x,y)\ln(\frac{p(x)p(y)}{p(x,y)})dxdy$$

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

$$\begin{aligned}
\mathbf{I}[\mathbf{x},\mathbf{y}] &= -\int\int p(x,y)\ln(\frac{p(x)p(y)}{p(x,y)})dxdy \\
&= -\int\int p(x,y)\ln(\frac{p(x)}{p(x|y)})dxdy \\
&= -\int\int p(x,y)\ln p(x)dxdy + \int\int p(x,y)\ln p(x|y)dxdy \\
&= -\int\int \ln p(x)p(x,y)dydx + \int\int p(x,y)\ln p(x|y)dxdy \\
&= -\int p(x)\ln p(x)dx + \int\int p(x,y)\ln p(x|y)dxdy \\
&= \mathbf{H}[\mathbf{x}] - \mathbf{H}[\mathbf{x}|\mathbf{y}]
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{I}[\mathbf{x},\mathbf{y}] &= -\int\int p(x,y)\ln(\frac{p(x)p(y)}{p(x,y)})dxdy \\
&= -\int\int p(x,y)\ln(\frac{p(y)}{p(y|x)})dxdy \\
&= -\int\int p(x,y)\ln p(y)dxdy + \int\int p(x,y)\ln p(y|x)dxdy \\
&= -\int\int \ln p(y)p(x,y)dxdy + \int\int p(x,y)\ln p(y|x)dxdy \\
&= -\int p(y)\ln p(y)dy + \int\int p(x,y)\ln p(y|x)dxdy \\
&= \mathbf{H}[\mathbf{y}] - \mathbf{H}[\mathbf{y}|\mathbf{x}]
\end{aligned}$$