Machine Learning (H)
Southern University of Science and Technology
Mengxuan Wu
12212006

Assignment 1

Mengxuan Wu

Question 1

The sum of squares error is

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

where t_n is the target value of x_n.
The partial derivative of $E(\mathbf{w})$ with respect to w_i is

$$\begin{aligned}
\frac{\partial E(\mathbf{w})}{\partial w_i} &= \frac{\partial}{\partial w_i}\frac{1}{2}\sum_{n=1}^{N}[y(x_n, \mathbf{w}) - t_n]^2 \\
&= \frac{1}{2}\sum_{n=1}^{N} 2[y(x_n, \mathbf{w}) - t_n]\frac{\partial y(x_n, \mathbf{w})}{\partial w_i} \\
&= \sum_{n=1}^{N}[y(x_n, \mathbf{w}) - t_n]\frac{\partial y(x_n, \mathbf{w})}{\partial w_i} \\
&= \sum_{n=1}^{N}[y(x_n, \mathbf{w}) - t_n]x_n^i
\end{aligned}$$

By setting the partial derivative to zero, we have

$$\begin{aligned}
\sum_{n=1}^{N}[y(x_n, \mathbf{w}) - t_n]x_n^i &= 0 \\
\sum_{n=1}^{N} y(x_n, \mathbf{w})x_n^i &= \sum_{n=1}^{N} t_n x_n^i \\
\sum_{n=1}^{N}\left(x_n^i \sum_{j=0}^{M} w_j x_n^j\right) &= \sum_{n=1}^{N} t_n x_n^i
\end{aligned}$$

Let X, T be two matrices, where $X_{ij} = x_i^{j-1}$, $T_i = t_i$. Then the equation can be written as

$$X^T X\mathbf{w} = X^T T$$

Hence, the optimal $\mathbf{w}^*$ that minimizes the sum of squares error is

$$\mathbf{w}^* = (X^T X)^{-1}X^T T$$

Question 2

a

The probability of selecting an apple is

$$\begin{aligned} P(\text{apple}) &= P(\text{apple}|r)P(r) + P(\text{apple}|b)P(b) + P(\text{apple}|g)P(g) \\ &= 0.3 \times 0.2 + 0.5 \times 0.2 + 0.3 \times 0.6 \\ &= 0.34 \end{aligned}$$

b

If the fruit is an orange, the probability of it being from the green box is

$$\begin{aligned} P(g|\text{orange}) &= \frac{P(\text{orange}|g)P(g)}{P(\text{orange})} \\ &= \frac{P(\text{orange}|g)P(g)}{P(\text{orange}|r)P(r) + P(\text{orange}|b)P(b) + P(\text{orange}|g)P(g)} \\ &= \frac{0.18}{0.36} \\ &= 0.5 \end{aligned}$$

Question 3

a

For continuous random variables x and z

$$\begin{aligned} \mathbb{E}[x+z] &= \iint (x+z)p(x,z)dxdz \\ &= \iint xp(x,z)dxdz + \iint zp(x,z)dxdz \\ &= \int x \left(\int p(x,z)dz \right) dx + \int z \left(\int p(x,z)dx \right) dz \\ &= \int xp(x)dx + \int zp(z)dz \\ &= \mathbb{E}[x] + \mathbb{E}[z] \end{aligned}$$

For discrete random variables x and z

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

b

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

Since x and z are statistically independent, $\mathbb{E}[xz] = \mathbb{E}[x]\mathbb{E}[z]$, the equation becomes

$$
\begin{aligned}
\mathrm{var}[x+z] &= \mathbb{E}[x^2] + \mathbb{E}[z^2] - \mathbb{E}[x]^2 - \mathbb{E}[z]^2 \\
&= \mathrm{var}[x] + \mathrm{var}[z]
\end{aligned}
$$

Question 4

a

The log likelihood function for the Poisson distribution is

$$\log p(\mathcal{D}|\lambda) = \sum_{n=1}^{N} \log \frac{\lambda^{X_n} e^{-\lambda}}{X_n!} = \sum_{n=1}^{N} (-\lambda + X_n \log \lambda - \log X_n!)$$

The derivative of the log likelihood function with respect to λ is

$$\frac{\partial}{\partial \lambda} \log p(\mathcal{D}|\lambda) = \sum_{n=1}^{N} (-1 + \frac{X_n}{\lambda})$$

By setting the derivative to zero, we have

$$
\begin{aligned}
\sum_{n=1}^{N} (-1 + \frac{X_n}{\lambda}) &= 0 \\
\sum_{n=1}^{N} X_n &= \sum_{n=1}^{N} \lambda \\
\lambda &= \frac{1}{N} \sum_{n=1}^{N} X_n
\end{aligned}
$$

Hence, the sample mean is the maximum likelihood estimate of $\hat{\lambda}$.

b

The log likelihood function for the exponential distribution is

$$\log p(\mathcal{D}|\lambda) = \sum_{n=1}^{N} \log \frac{1}{\lambda} e^{-\frac{X_n}{\lambda}} = \sum_{n=1}^{N} (-\log \lambda - \frac{X_n}{\lambda})$$

The derivative of the log likelihood function with respect to λ is

$$\frac{\partial}{\partial \lambda} \log p(\mathcal{D}|\lambda) = \sum_{n=1}^{N} (\frac{X_n}{\lambda^2} - \frac{1}{\lambda})$$

By setting the derivative to zero, we have

$$\begin{aligned}
\sum_{n=1}^{N}(\frac{X_n}{\lambda^2} - \frac{1}{\lambda}) = 0 \\
\sum_{n=1}^{N} X_n = \sum_{n=1}^{N} \lambda \\
\lambda = \frac{1}{N}\sum_{n=1}^{N} X_n
\end{aligned}$$

Hence, the sample mean is the maximum likelihood estimate of $\hat{\lambda}$.

Question 5

a

The probability of classifying correctly is

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

The probability of classifying incorrectly is

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

b

The derivative of expected loss with respect to function $y(x)$ is

$$\begin{aligned}
\frac{\partial \mathbb{E}[L(t, y(x))]}{\partial y(x)} &= \frac{\partial}{\partial y(x)} \iint ||t - y(x)||^2 p(x, t)dxdt \\
&= 2\int (y(x) - t)p(x, t)dt
\end{aligned}$$

By setting the derivative to zero, we have

$$\begin{aligned}
\int (y(x) - t)p(x, t)dt = 0 \\
\int y(x)p(x, t)dt = \int tp(x, t)dt \\
y(x)\int p(x, t)dt = \int tp(x, t)dt \\
y(x)p(x) = \int tp(x, t)dt \\
y(x) = \int tp(t|x)dt \\
y(x) = \mathbb{E}_t[t|x]
\end{aligned}$$

Question 6

a

The entropy for $X \sim \text{Gaussian}(\mu, \sigma^2)$ is

$$\begin{aligned} H(X) &= -\int_{-\infty}^{\infty} p(x) \log p(x) dx \\ &= -\mathbb{E}[\log p(x)] \\ &= -\mathbb{E}[-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2}] \\ &= \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\mathbb{E}[(x-\mu)^2] \end{aligned}$$

For Gaussian distribution, we have $\mathbb{E}[(x-\mu)^2] = \sigma^2$, hence

$$H(X) = \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2}$$

b

For continuous random variables x and y, the mutual information is

$$\begin{aligned} I[x,y] &= \iint p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dxdy \\ &= \iint p(x,y) \log \frac{p(x,y)}{p(x)} dxdy - \iint p(x,y) \log p(y) dxdy \\ &= \iint p(x,y) \log \frac{p(x,y)}{p(x)} dxdy - \int \left(\int p(x,y) dx \right) \log p(y) dy \\ &= \iint p(x,y) \log \frac{p(x,y)}{p(x)} dxdy - \int p(y) \log p(y) dy \\ &= H(y) - H(y|x) \end{aligned}$$

Since x and y are interchangeable, we have $I[x,y] = H(x) - H(x|y) = H(y) - H(y|x)$. For discrete random variables x and y, the mutual information is

$$\begin{aligned} I[x,y] &= \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\ &= \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)} - \sum_x \sum_y p(x,y) \log p(y) \\ &= \sum_x \sum_y p(x,y) \log p(y|x) - \sum_y \left(\sum_x p(x,y) \right) \log p(y) \\ &= \sum_x \sum_y p(x,y) \log p(y|x) - \sum_y p(y) \log p(y) \\ &= H(y) - H(y|x) \end{aligned}$$

Similarly, we have $I[x,y] = H(x) - H(x|y) = H(y) - H(y|x)$.