CS329 Machine Learning

Homework #1

Fan Site

fanst2021@mail.sustech.edu.cn

Question 1

Consider the polynomial function:

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

Calculate the coefficients $\boldsymbol{w} = \{w_i\}$ 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.

Solution: least squares

The sum-of-squares error function is defined as:

$$E(\boldsymbol{w}) = \sum_{j=1}^{N} \left[y(x_j, \boldsymbol{w}) - t_j\right]^2$$

where:

- N is the number of data points.
- (x_j, t_j) are the input-output pairs in the dataset.
- $y(x_j, \boldsymbol{w})$ is the given polynomial function.

To minimize the error function $E(\boldsymbol{w})$ with respect to the coefficients $\boldsymbol{w} = \{w_i\}$, we'll take the derivative of the error function with respect to each coefficient w_i and set each derivative equal to zero to find the minimum. The derivative of $E(\boldsymbol{w})$ with respect to w_i is:

$$\frac{\partial E}{\partial w_i} = -2 \sum_{j=1}^{N} x_j^i \left(t_j - \sum_{k=0}^{M} w_k x_j^k\right)$$

Setting this derivative equal to zero for each i:

$$-2 \sum_{j=1}^{N} x_j^i \left(t_j - \sum_{k=0}^{M} w_k x_j^k\right) = 0$$

Now, we can solve for the coefficients w_i:

$$\sum_{j=1}^{N} x_j^i \sum_{k=0}^{M} w_k x_j^k = \sum_{j=1}^{N} x_j^i t_j$$

This equation forms a system of linear equations for the coefficients $\boldsymbol{w}$.

$$\boldsymbol{X}\boldsymbol{w} = \boldsymbol{T}$$

Where:

- $\boldsymbol{X}$ is the design matrix whose elements are x_j^i for all i and j.
- $\boldsymbol{T}$ is the target vector whose elements are the corresponding t_j values.

Hence $\boldsymbol{w} = \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{T}$

With this $\boldsymbol{w}$, the sum-of-squares error is minimized.

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?

Solution: Bayes

Let A_i be the event "Choosing the box with color i", where $i \in \{r, g, b\}$

So that $P(A_r) = p(r) = 0.2$, $P(A_g) = p(g) = 0.6$, $P(A_b) = p(b) = 0.2$.

Let B_i be the event "The selected item is fruit i", where $i \in \{a, o, l\}$, meaning apple, orange and lime.

Using the Law of Total Probability

$$\begin{aligned} P(B_a) &= P(B_a|A_r) * P(A_r) + P(B_a|A_g) * P(A_r) + P(B_a|A_b) * P(A_b) \\ &= \frac{3}{10} * 0.2 + \frac{3}{10} * 0.6 + \frac{1}{2} * 0.2 \\ &= 0.34 \end{aligned}$$

The probability of selecting an apple is 0.34.

$$\begin{aligned} P(B_o) &= P(B_o|A_r) * P(A_r) + P(B_o|A_g) * P(A_r) + P(B_o|A_b) * P(A_b) \\ &= \frac{4}{10} * 0.2 + \frac{3}{10} * 0.6 + \frac{1}{2} * 0.2 \\ &= 0.36 \end{aligned}$$

Using the Bayes' Theorem

$$P(A_g|B_o) = \frac{P(B_o|A_g) * P(A_g)}{P(B_o)} = \frac{0.3 * 0.6}{0.36} = 0.50$$

Known that an orange is selected, the probability that it's from green box is 0.50.

Question 3

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

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

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

Solution

Assume that X and Z are continuous random variables

$$\begin{aligned} \mathbb{E}(X + Z) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x + z) p(x, z) dx dz \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x p(x, z) dx dz + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} z p(x, z) dx dz \\ &= \int_{-\infty}^{\infty} x p_X(x) dx + \int_{-\infty}^{\infty} z p_Z(z) dz \\ &= \mathbb{E}(X) + \mathbb{E}(Z) \end{aligned}$$

Similarly, for discrete case

$$\begin{aligned}\mathbb{E}(X+Z) &= \sum_x \sum_z (x+z)p(x,z) \\ &= \sum_x \sum_z xp(x,z) + \sum_x \sum_z zp(x,z) \\ &= \sum_x xp_X(x) + \sum_z xp_Z(z) \\ &= \mathbb{E}(X) + \mathbb{E}(Z)\end{aligned}$$

Then prove that for independent variables X and Z, "the variance of the sum equals to sum of the variance"

$$\begin{aligned}\operatorname{var}[X+Z] &= \mathbb{E}\left((X+Z)^2\right) - (\mathbb{E}(X+Z))^2 \\ &= \mathbb{E}(X^2+2XZ+Z^2) - \left((\mathbb{E}(X))^2 + 2\mathbb{E}(X)\mathbb{E}(Z) + (\mathbb{E}(Z))^2\right) \\ &= \mathbb{E}(X^2) - (\mathbb{E}(X))^2 + \mathbb{E}(Z^2) - (\mathbb{E}(Z))^2 + 2(\mathbb{E}(XZ) - \mathbb{E}(X)\mathbb{E}(Z)) \\ &= \operatorname{var}[X] + \operatorname{var}[Z]\end{aligned}$$

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. $X \sim \text{Possion}(\lambda)$, 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 $\text{Possion}(\lambda) : \mathcal{D} = \{X_1, X_2, ..., X_n\}$. Show that the sample mean $\hat{\lambda} = \frac{1}{n}\sum_{i=1}^n X_i$ 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 $\hat{\lambda} = \frac{1}{n}\sum_{i=1}^n X_i$ is the maximum likelihood estimate(MLE) of λ.

Solution

Possion distribution

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

$$\text{Likelihood} \quad L(\lambda) = \prod_{i=1}^{n} P(X_i|\lambda)$$

$$\ln L(\lambda) = \ln\left(\prod_{i=1}^{n} \frac{\lambda_i^X e^{-\lambda}}{X_i!}\right) = \ln\left(\prod_{i=1}^{n} \frac{1}{X_i!}\right) + \ln\lambda \sum_{i=1}^{n} X_i - n\lambda$$

$$\text{Let} \quad \frac{\partial \ln L(\lambda)}{\partial\lambda} = \sum_{i=1}^{n} \frac{X_i}{\lambda} - n = 0$$

$$\text{The MLE is } \hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Exponential distribution

$$P(X|\lambda) = \frac{1}{\lambda} e^{-\frac{X}{\lambda}}$$

$$\text{Likelihood} \quad L(\lambda) = \prod_{i=1}^{n} P(X_i|\lambda)$$

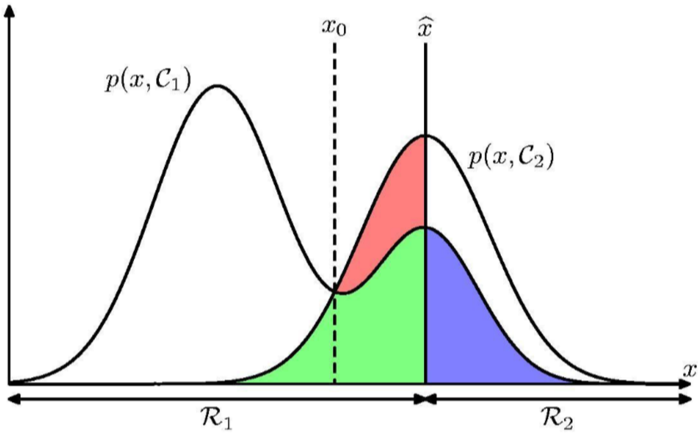
$$\ln L(\lambda) = \ln\left(\prod_{i=1}^{n} \frac{1}{\lambda} e^{-\frac{X_i}{\lambda}}\right) = -n\ln\lambda - \sum_{i=1}^{n} \frac{X_i}{\lambda}$$

$$\text{Let} \quad \frac{\partial \ln L(\lambda)}{\partial\lambda} = -\frac{n}{\lambda} + \sum_{i=1}^{n} \frac{X_i}{\lambda^2} = 0$$

$$\text{The MLE is } \hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Question 5

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



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

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

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

(a) Solution

$$p(\text{correct}) = p(x \in \mathcal{R}_1, \mathcal{C}_1) + p(x \in \mathcal{R}_2, \mathcal{C}_2) = \int_{\mathcal{R}_1} p(x, \mathcal{C}_1)dx + \int_{\mathcal{R}_2} p(x, \mathcal{C}_2)dx$$

$$p(\text{mistake}) = p(x \in \mathcal{R}_1, \mathcal{C}_2) + p(x \in \mathcal{R}_2, \mathcal{C}_1) = \int_{\mathcal{R}_1} p(x, \mathcal{C}_2)dx + \int_{\mathcal{R}_2} p(x, \mathcal{C}_1)dx$$

(b) Solution

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

$$\text{Let} \quad \frac{\partial \mathbb{E}[L(\boldsymbol{t}, \boldsymbol{y}(\boldsymbol{x}))]}{\partial \boldsymbol{y}(\boldsymbol{x})} = 2\int (\boldsymbol{y}(\boldsymbol{x}) - \boldsymbol{t})p(\boldsymbol{x}, \boldsymbol{t})d\boldsymbol{t} = 0$$

$$\text{Isolating } \boldsymbol{y}(\boldsymbol{x}), \quad \boldsymbol{y}(\boldsymbol{x}) \int p(\boldsymbol{x}, \boldsymbol{t})d\boldsymbol{t} = \int \boldsymbol{t}p(\boldsymbol{x}, \boldsymbol{t})d\boldsymbol{t}$$

$$\boldsymbol{y}(\boldsymbol{x}) = \frac{\int tp(\boldsymbol{x}, \boldsymbol{t})d\boldsymbol{t}}{\int p(\boldsymbol{x}, \boldsymbol{t})d\boldsymbol{t}} = \frac{\int tp(\boldsymbol{x}, \boldsymbol{t})d\boldsymbol{t}}{p(\boldsymbol{x})} = \int \boldsymbol{t}p(\boldsymbol{t}|\boldsymbol{x})d\boldsymbol{t} = \mathbb{E}_{\boldsymbol{t}}[\boldsymbol{t}|\boldsymbol{x}]$$

Question 6

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

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

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

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

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

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

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

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

$$\mathbf{I}[\mathbf{X}, \mathbf{Y}] = \mathbf{H}[\mathbf{X}] - \mathbf{H}[\mathbf{X}|\mathbf{Y}] = \mathbf{H}[\mathbf{Y}] - \mathbf{H}[\mathbf{Y}|\mathbf{X}]$$

(a) Solution

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

$$\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$$

$$\text{Let} \quad t = \frac{x-\mu}{\sqrt{2}\sigma}, \frac{dt}{dx} = \frac{1}{\sqrt{2}\sigma}$$

$$
\begin{aligned}
\mathbf{H}[\mathbf{X}] &= -\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \ln\left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right) dx \\
&= -\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-t^2} \ln\left(\frac{1}{\sqrt{2\pi}\sigma} e^{-t^2}\right) dx \\
&= -\frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{-t^2} \left(-\ln\left(\sqrt{2\pi}\sigma\right) - t^2\right) dx \\
&= \frac{\ln\left(\sqrt{2\pi}\sigma\right)}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{-t^2} dx + \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} t^2 e^{-t^2} dx \\
&= \frac{\ln\left(\sqrt{2\pi}\sigma\right)}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-t^2} dt + \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} t^2 e^{-t^2} dt \\
&= \ln\left(\sqrt{2\pi}\sigma\right) + \frac{1}{\sqrt{\pi}} \cdot -\frac{1}{2} \left(0 - \int_{-\infty}^{\infty} e^{-t^2} dt\right) \\
&= \ln\left(\sqrt{2\pi}\sigma\right) + \frac{1}{2}
\end{aligned}
$$

(b) Solution

Discrete Distributions

$$
\mathbf{I}[\mathbf{X},\mathbf{Y}] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_{X,Y}(x,y) \ln\left(\frac{P_{X,Y}(x,y)}{P_X(x) P_Y(y)}\right)
$$

$$
\begin{aligned}
\mathbf{I}[\mathbf{X},\mathbf{Y}] &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_{X,Y}(x,y) \ln\left(\frac{P_{X,Y}(x,y)}{P_Y(y)}\right) - \sum_{x \in \mathcal{X}} P_X(x) \ln(P_X(x)) \\
&= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_Y(y) P(x|y) \ln P(x|y) + \mathbf{H}(\mathbf{X}) \\
&= \mathbf{H}[\mathbf{X}] - \mathbf{H}[\mathbf{X}|\mathbf{Y}].
\end{aligned}
$$

where $P_{X,Y}$ is the joint probability mass function of X and Y, and P_X and P_Y are the marginal probability mass functions of X and Y respectively. Similarly, $\mathbf{I}[\mathbf{X},\mathbf{Y}] = \mathbf{H}[\mathbf{Y}] - \mathbf{H}[\mathbf{Y}|\mathbf{X}]$.

Continuous Distributions

$$
\mathbf{I}[\mathbf{x},\mathbf{y}] = \int_{\mathcal{X}} \int_{\mathcal{Y}} P_{X,Y}(x,y) \ln\left(\frac{P_{X,Y}(x,y)}{P_X(x) P_Y(y)}\right) dy dx
$$

$$
\begin{aligned}
\mathbf{I}[\mathbf{X},\mathbf{Y}] &= \int_{\mathcal{X}} \int_{\mathcal{Y}} P_{X,Y}(x,y) \ln\left(\frac{P_{X,Y}(x,y)}{P_Y(y)}\right) dy dx - \int_{\mathcal{X}} P_X(x) \ln(P_X(x)) dx \\
&= \int_{\mathcal{X}} \int_{\mathcal{Y}} P_Y(y) P(x|y) \ln P(x|y) dy dx + \mathbf{H}(\mathbf{X}) \\
&= \mathbf{H}[\mathbf{X}] - \mathbf{H}[\mathbf{X}|\mathbf{Y}].
\end{aligned}
$$

where $P_{X,Y}$ is the joint probability density function of X and Y, and P_X and P_Y are the marginal probability density functions of X and Y respectively. Similarly, $\mathbf{I}[\mathbf{X},\mathbf{Y}] = \mathbf{H}[\mathbf{Y}] - \mathbf{H}[\mathbf{Y}|\mathbf{X}]$.