The foundation of DL

I haven't thought of a title yet

School of Computer Science
WuHan University

Franz Hu

What is the first

This is my second time to write notes all in English.

Contents

1 Mathematical Basics **1**
1.1 Information Theory 1

1 Mathematical Basics

1.1 Information Theory

The quantification of information content is contingent upon the probability distribution. We seek a measure that expresses the information content in a manner that is inversely related to its probability. Furthermore, the information gain from observing two independent events should equal the sum of the information gains obtained from observing each event individually. That is:

$$h(x, y) = h(x) + h(y)$$

Meanwhile, their joint probability is given by the product of their individual probabilities:

$$p(x, y) = p(x)p(y)$$

From these premises, we can deduce that the relationship between information content and probability is expressed as:

$$h(x) = -\log_2 p(x)$$

The negative sign ensures the non-negativity of information.

Definition 1.1.1. *Therefore, the average information content of a transmission process is obtained by taking the expectation with respect to $p(x)$:*

$$H[x] = -\sum_x p(x) \log_2 p(x)$$

This quantity is referred to as the entropy of the random variable x.

For continuous variable x, we define:

$$H[x] = -\int p(x) \ln p(x)\, dx$$

as the differential entropy of continuous variables.

Theorem 1.1.1. *Noiseless coding theorem: Entropy serves as a lower bound on the number of bits required to transmit the state of a random variable.*

Indeed, this calculates the same fundamental concept as Huffman's minimum path encoding. Both entropy and Huffman coding address the theoretical limit and practical achievement of efficient information representation.

For discrete distributions, the maximum entropy configuration occurs when the variable's probabilities are uniformly distributed among all possible states. For continuous variables, we proceed with the following analysis:

$$\int_{-\infty}^{\infty} p(x)\, dx = 1$$
$$\int_{-\infty}^{\infty} xp(x)\, dx = \mu$$
$$\int_{-\infty}^{\infty} (x-\mu)^2 p(x)\, dx = \sigma^2$$

We can use the method of Lagrange multipliers to compute the constrained maximum.

$$\mathcal{L}(x) = -\int_{-\infty}^{\infty} p(x) \ln p(x)\, dx + \lambda_1 (\int_{-\infty}^{\infty} p(x)\, dx - 1) + \lambda_2 (\int_{-\infty}^{\infty} xp(x)\, dx - \mu) + \lambda_3 (\int_{-\infty}^{\infty} (x-\mu)^2 p(x)\, dx - \sigma^2)$$

$$\begin{aligned} \frac{\delta \mathcal{L}}{\delta p} &= \frac{\delta}{\delta p}[-p \ln p + \lambda_1 p + \lambda_2 xp + \lambda_3 (x-\mu)^2 p] \\ &= -\ln p - 1 + \lambda_1 + \lambda_2 x + \lambda_3 (x-\mu)^2 = 0 \end{aligned}$$

Key Rule: For a functional $\mathcal{L} = \int F(p,x)dx$, the variational derivative $\frac{\delta \mathcal{L}}{\delta p}$ equals the partial derivative of the integrand F with respect to p (i.e., $\frac{\partial F}{\partial p}$).

$$\text{Let } \frac{\partial}{\partial p(x)} F(p) = -\ln p(x) - 1 + \lambda_1 + \lambda_2 x + \lambda_3 (x-\mu)^2 = 0$$

$$\begin{aligned} p(x) &= e^{\left(-1+\lambda_1+\lambda_2 x+\lambda_3(x-\mu)^2\right)} \\ &= e^{-1+\lambda_1} \cdot e^{\lambda_2 x+\lambda_3(x-\mu)^2} = Ce^{\lambda_2 x+\lambda_3(x-\mu)^2} \\ &= Ce^{\lambda_3\left(x^2-2\left(\mu-\frac{\lambda_2}{2\lambda_3}\right)x+\mu^2\right)} = Ce^{\lambda_3\left(x-\mu+\frac{\lambda_2}{2\lambda_3}\right)^2} \end{aligned}$$

Since $p(x) > 0$, so $C > 0$.$p(x)$ is symmetric about $\mu - \frac{\lambda_2}{2\lambda_3}$, so $\mathbb{E}[p(x)] = \mu - \frac{\lambda_2}{2\lambda_3} = \mu$. It follows that:

$$\lambda_2 = 0$$

The form thus becomes:

$$p(x) = Ce^{\lambda_3(x-\mu)^2}$$

Finally, substituting in the constraint conditions gives the answer:

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

Here we supplement some necessary knowledge of functional analysis. A **function** represents a mapping from numbers to numbers, while a **functional** represents a mapping from functions to numbers.

For instance, consider the functional $J(y) = \int_{x_1}^{x_2} \sqrt{1+(y_x)^2}\, dx$. The goal is to find a suitable function $y(x)$ such that $J(y)$ attains its minimum.

Generally, we define a functional concerned with y (which is twice differentiable on the interval $[a,b]$):

$$J(y) = \int_a^b F(x, y, y_x)\, dx$$

Given that $y(a)$ and $y(b)$ are known, and that F is twice differentiable with respect to all its arguments, what condition must the function $y(x)$ satisfy in this general case for the functional to attain a minimum value?

For the general case described above, when the functional $J(y)$ attains an extremum, the function $y(x)$ must satisfy the Euler-Lagrange equation:

Theorem 1.1.2. ***Euler-Lagrange Equation:*** *Let $J(y)$ be a functional defined by*

$$J(y) = \int_a^b F(x, y, y_x) dx$$

where F is twice continuously differentiable in all its arguments. If $y = u(x)$ yields an extremum of $J(y)$ among all functions satisfying the boundary conditions $y(a) = y_0$ and $y(b) = y_1$, then $u(x)$ must satisfy the following necessary condition:

$$\frac{\partial F}{\partial y} - \frac{d}{dx}\left(\frac{\partial F}{\partial y_x}\right) = 0$$

Proof. We assume that the function yielding an extremum for the functional $J(y)$ is $y = u(x)$. On the interval $[x_1 = a, x_2 = b]$, we define a family of functions:

$$y(x) = u(x) + \epsilon\eta(x)$$

where ϵ is a real parameter and $\eta(x)$ is defined as the difference between any other function $y(x)$ connecting points P_1 and P_2 and the function $u(x)$. Consequently, $\eta(x)$ must satisfy the boundary conditions $\eta(a) = \eta(b) = 0$, ensuring $y(x)$ also satisfies these conditions and represents an admissible path. For any fixed $\eta(x)$, the functional $J(y) = J(u + \epsilon\eta)$ becomes a function of the parameter ϵ only, denoted $J(\epsilon)$. Since $y(x) = u(x)$ when $\epsilon = 0$ (where $J(y)$ attains its extremum), it follows that:

$$\left.\frac{dJ(\epsilon)}{d\epsilon}\right|_{\epsilon=0} = 0$$

To satisfy this condition, we first compute $\frac{dJ(\epsilon)}{d\epsilon}$ for an arbitrary but fixed $\eta(x)$:

$$\begin{aligned}
\frac{dJ(\epsilon)}{d\epsilon} &= \frac{d}{d\epsilon}\int_a^b F(x, y(x), y_x(x))dx \\
&= \int_a^b \frac{\partial F(x, y(x), y_x(x))}{\partial \epsilon}dx \\
&= \int_a^b \left(\frac{\partial F}{\partial y}\frac{\partial y}{\partial \epsilon} + \frac{\partial F}{\partial y_x}\frac{\partial y_x}{\partial \epsilon}\right)dx \\
&= \int_a^b \frac{\partial F}{\partial y}\eta(x)dx + \int_a^b \frac{\partial F}{\partial y_x}\eta'(x)dx \\
&= \int_a^b \frac{\partial F}{\partial y}\eta(x)dx + \int_a^b \frac{\partial F}{\partial y_x}d\eta(x) \\
&= \int_a^b \frac{\partial F}{\partial y}\eta(x)dx + \left[\eta(x)\frac{\partial F}{\partial y_x}\right]_a^b - \int_a^b \frac{d}{dx}\left(\frac{\partial F}{\partial y_x}\right)\eta(x)dx \quad \text{(Integration by Parts)} \\
&= \int_a^b \left[\frac{\partial F}{\partial y} - \frac{d}{dx}\left(\frac{\partial F}{\partial y_x}\right)\right]\eta(x)dx \quad \text{(since } \eta(a) = \eta(b) = 0\text{)}
\end{aligned}$$

Evaluating this derivative at $\epsilon = 0$ (where $y = u$ and $y_x = u_x$) yields:

$$\left.\frac{dJ(\epsilon)}{d\epsilon}\right|_{\epsilon=0} = \int_a^b \left[\frac{\partial F}{\partial u} - \frac{d}{dx}\left(\frac{\partial F}{\partial u_x}\right)\right]\eta(x)dx = 0$$

Given that $\eta(x)$ is arbitrary and the expression $\frac{\partial F}{\partial u} - \frac{d}{dx}\left(\frac{\partial F}{\partial u_x}\right)$ is a fixed function for given F, the integral can only be zero for all $\eta(x)$ if:

$$\frac{\partial F}{\partial u} - \frac{d}{dx}\left(\frac{\partial F}{\partial u_x}\right) = 0$$

This completes the proof. The function $u(x)$ that extremizes the functional $J(y)$ must satisfy this Euler-Lagrange equation.

□

Therefore, the distribution that maximizes the differential entropy is the **Gaussian distribution**. If we compute the differential entropy of the Gaussian distribution, we obtain:

$$H[x] = \frac{1}{2}(1 + \ln(2\pi\sigma^2))$$

It is noteworthy that, unlike discrete entropy, differential entropy can be negative. $H(x) < 0$, when $\sigma^2 < \frac{1}{2\pi e}$.

To describe the average additional information required to encode the values of x, we define the **relative entropy**, also called Kullback and Leibler divergence:

$$\mathrm{KL}(p \parallel q) = -\int p(x)\ln q(x)d\,x - (-\int p(x)\ln p(x)d\,x) = -\int p(x)\ln\frac{q(x)}{p(x)}d\,x = -\mathbb{E}_p[\ln\frac{q(x)}{p(x)}]$$

Theorem 1.1.3. ***Jensen's inequality***
For any set of points $\{x_i\}$, the convex function $f(x)$ must satisfy:

$$f(\sum_{i=1}^{M} \lambda_i x_i) \leqslant \sum_{i=1}^{M} \lambda_i f(x_i)$$

where $\lambda_i \geqslant 0$ and $\sum_{i=1}^{M} \lambda_i = 1$.

$$\mathrm{KL}(p \parallel q) = -\int p(x) \ln \frac{q(x)}{p(x)} d\,x \geqslant -\ln \int q(x) d\,x = 0$$

Suppose the data is generated from an unknown distribution $p(x)$ that we wish to model. We can attempt to approximate this distribution using a parametric distribution $q(x \mid \boldsymbol{\theta})$ controlled by a set of adjustable parameters $\boldsymbol{\theta}$. One method to determine $\boldsymbol{\theta}$ is to minimize the Kullback-Leibler divergence between $p(x)$ and $q(x \mid \boldsymbol{\theta})$. However, since we do not know $p(x)$, this cannot be done directly. Nevertheless, assuming we have observed a finite set of training points—x_n generated from $p(x)$ for $n = 1, \ldots, N$—we can approximate the expectation with respect to $p(x)$ via a finite sum over these training points:

$$\mathrm{KL}(p \| q) \approx \frac{1}{N} \sum_{n=1}^{N} \left(-\ln q\left(x_n \mid \boldsymbol{\theta}\right) + \ln p\left(x_n\right)\right)$$

The second term on the right-hand side of the equation is independent of $\boldsymbol{\theta}$. The first term is the **negative log-likelihood function** of the distribution $q(x \mid \boldsymbol{\theta})$ evaluated using the training set. Therefore, it can be seen that minimizing this Kullback-Leibler divergence is equivalent to maximizing the log-likelihood function.

Definition 1.1.2. *Suppose x is given, the average additional information to know y is:*

$$H[y|x] = -\iint p(y,x) \ln p(y|x) d\,y d\,X$$

as the conditional entropy of y given x.

When two variables x and y are independent, their joint distribution can be factorized into the product of their marginal distributions, i.e., $p(x, y) = p(x)p(y)$. If these two variables are not independent, one can assess how "close" they are to independence by considering the Kullback-Leibler divergence between the joint distribution and the product of the marginal distributions, which is given by:

Definition 1.1.3.

$$I[x,y] \equiv \mathrm{KL}(p(x,y) \| p(x)p(y)) = -\iint p(x,y) \ln \left(\frac{p(x)p(y)}{p(x,y)}\right) dx\, dy$$

*$I[x, y]$ is called the **mutual information** between variables x and y. From the properties of the Kullback-Leibler divergence, we see that $I[x, y] \geqslant 0$, with equality holding if and only if x and y are independent. Using the sum and product rules of probability, we can see that the mutual information is related to the conditional entropy.*

$$I[x,y] = H[x] - H[x|y] = H[y] - H[y|x] \tag{2.110}$$

Therefore, the mutual information represents the reduction in uncertainty about x due to being informed of the value of y (or vice versa). From a Bayesian perspective, $p(x)$ can be viewed as the prior distribution of x, and $p(x|y)$ as the posterior distribution after observing new data y. In summary, the mutual information represents the reduction in uncertainty about x resulting from the new observation y.

The Bayesian approach to parameter estimation incorporates prior knowledge through the **maximum a posteriori** (MAP) framework. We express our belief about the parameters **w** before observing the data by specifying a **prior distribution** $p(\mathbf{w})$. A common and convenient choice is the Gaussian prior, which encodes a preference for smaller parameter values:

$$w_j \sim \mathcal{N}(0, \lambda^{-1}), \quad \text{for all } j.$$

This distributional assumption reflects the belief that the model weights should not be too large, with the hyperparameter λ controlling the strength of this belief.

The core objective is to find the parameter values **w** that maximize the **posterior probability** $p(\mathbf{w} \mid X, Y)$. According to Bayes' theorem, the posterior is proportional to the product of the likelihood and the prior:

$$p(\mathbf{w} \mid X, Y) \propto p(Y \mid X, \mathbf{w})\, p(\mathbf{w}).$$

Maximizing the posterior probability is equivalent to minimizing its negative logarithm:

$$-\ln p(\mathbf{w} \mid X, Y) = -\ln p(Y \mid X, \mathbf{w}) - \ln p(\mathbf{w}) + \text{constant}.$$

Substituting the Gaussian likelihood for the regression model, $p(Y \mid X, \mathbf{w}) = \prod_{i=1}^{N} \mathcal{N}(y_i \mid \mathbf{w}^T \mathbf{x}_i, \sigma^2)$, and the Gaussian prior, $p(\mathbf{w}) = \prod_{j=1}^{D} \mathcal{N}(w_j \mid 0, \lambda^{-1})$, yields the following expression:

$$-\ln p(\mathbf{w} \mid X, Y) = \frac{1}{2\sigma^2} \sum_{i=1}^{N} \left(y_i - \mathbf{w}^T \mathbf{x}_i\right)^2 + \frac{\lambda}{2} \sum_{j=1}^{D} w_j^2 + \text{constant}.$$

The right-hand side of this equation is precisely the standard **regularized loss function** used in machine learning, composed of a mean squared error loss and an L2 regularization term:

$$J(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \mathbf{w}^T \mathbf{x}_i\right)^2 + \frac{\lambda}{2} \sum_{j=1}^{D} w_j^2.$$

From this Bayesian perspective, the regularization term $\frac{\lambda}{2} \sum_j w_j^2$ is not an ad-hoc penalty but arises naturally from the negative log of the Gaussian prior. It penalizes large weights, thereby encouraging a simpler, smoother model that is less prone to overfitting.