1.1. Curve Fitting

1. As we shall see in Chapter 3, the number of parameters is not necessarily the most appropriate measure of model complexity. 参数个数不足以衡量模型的复杂度
2. We shall see that the least squares approach to finding the model parameters represents a specific case of maximum likelihood (discussed in Section 1.2.5), and that the over-fitting problem can be understood as Section 3.4 a general property of maximum likelihood 最小二乘法可以看作MLE，过拟合是MLE的property

1.2. Probability Theory

> The Rules of Probability
>
> **sum rule** $$p(X) = \sum_{Y} p(X, Y) \tag{1.10}$$
>
> **product rule** $$p(X, Y) = p(Y|X)p(X). \tag{1.11}$$

Bayes' theorem

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} \tag{1.12}$$

$$p(X) = \sum_{Y} p(X|Y)p(Y). \tag{1.13}$$

We can view the denominator in Bayes' theorem as being the normalization constant required to ensure that the sum of the conditional probability on the left-hand side of (1.12) over all values of Y equals one.

If we had been asked which box had been chosen before being told the identity of the selected item of fruit, then the most complete information we have available is provided by the probability p(B). We call this the **prior probability** because it is the probability available before we observe the identity of the fruit.

Once we are told that the fruit is an orange, we can then use Bayes' theorem to compute the probability p(B|F), which we shall call the **posterior probability** because it is the probability obtained after we have observed F.

We note that if the joint distribution of two variables factorizes into the product of the marginals, so that p(X, Y) = p(X)p(Y), then X and Y are said to be **independent**. From the product rule, we see that p(Y |X) = p(Y), and so the conditional distribution of Y given X is indeed independent of the value of X.

1.2.1 Probability densities

The probability that x will lie in an interval (a, b) is then given by

$$p(x \in (a,b)) = \int_a^b p(x)\,\mathrm{d}x. \tag{1.24}$$

满足

$$p(x) \geqslant 0 \tag{1.25}$$

$$\int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1. \tag{1.26}$$

下面这里提到了一个jacobian factor，大概意思就是：
概率密度函数和普通的函数是不同的。
在pdf中，如果两个随机变量有非线性的函数关系，这两个函数的pdf不会保持这种关系。
进而可以推广出结论，pdf的最值与variable的选择有关

Under a nonlinear change of variable, a probability density transforms differently from a simple function, due to the Jacobian factor. For instance, if we consider a change of variables $x = g(y)$, then a function $f(x)$ becomes $\widetilde{f}(y) = f(g(y))$. Now consider a probability density $p_x(x)$ that corresponds to a density $p_y(y)$ with respect to the new variable y, where the suffices denote the fact that $p_x(x)$ and $p_y(y)$ are different densities. Observations falling in the range $(x, x + \delta x)$ will, for small values of δx, be transformed into the range $(y, y + \delta y)$ where $p_x(x)\delta x \simeq p_y(y)\delta y$, and hence

$$\begin{aligned} p_y(y) &= p_x(x)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right| \\ &= p_x(g(y))\,|g'(y)|\,. \end{aligned} \tag{1.27}$$

One consequence of this property is that the concept of the maximum of a probability density is dependent on the choice of variable.

下面是对于“pdf的最值与variable的选择无关”的说明：
由下面的推导，可知对于普通的非线性函数，最值的非线性关系是保持的：

Consider first the way a function $f(x)$ behaves when we change to a new variable y where the two variables are related by $x = g(y)$. This defines a new function of y given by

$$\widetilde{f}(y) = f(g(y)). \tag{2}$$

Suppose $f(x)$ has a mode (i.e. a maximum) at $\widehat{x}$ so that $f'(\widehat{x}) = 0$. The corresponding mode of $\widetilde{f}(y)$ will occur for a value $\widehat{y}$ obtained by differentiating both sides of (2) with respect to y

$$\widetilde{f}'(\widehat{y}) = f'(g(\widehat{y}))g'(\widehat{y}) = 0. \tag{3}$$

Assuming $g'(\widehat{y}) \neq 0$ at the mode, then $f'(g(\widehat{y})) = 0$. However, we know that $f'(\widehat{x}) = 0$, and so we see that the locations of the mode expressed in terms of each of the variables x and y are related by $\widehat{x} = g(\widehat{y})$, as one would expect. Thus, finding a mode with respect to the variable x is completely equivalent to first transforming to the variable y, then finding a mode with respect to y, and then transforming back to x.

但是
由于有这个式子：

$$\begin{aligned} p_y(y) &= p_x(x)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right| \\ &= p_x(g(y))\,|g'(y)|\,. \end{aligned} \tag{1.27}$$

Now consider the behaviour of a probability density $p_x(x)$ under the change of variables $x = g(y)$, where the density with respect to the new variable is $p_y(y)$ and is given by ((1.27)). Let us write $g'(y) = s|g'(y)|$ where $s \in \{-1, +1\}$. Then ((1.27)) can be written

$$p_y(y) = p_x(g(y))sg'(y).$$

Differentiating both sides with respect to y then gives

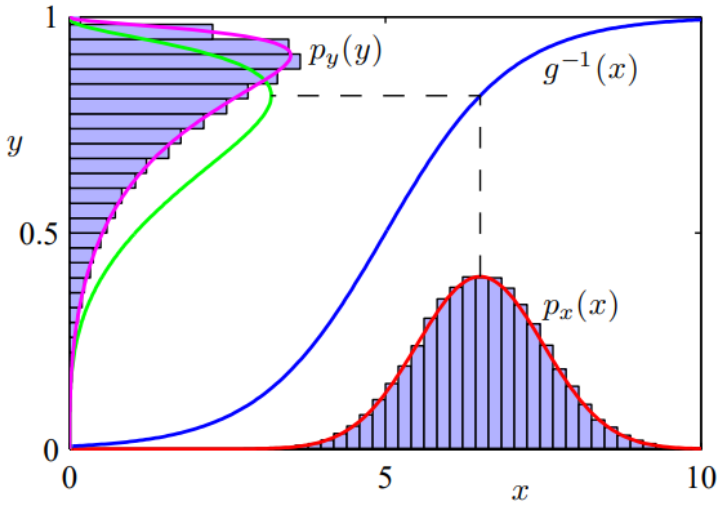
$$p'_y(y) = sp'_x(g(y))\{g'(y)\}^2 + sp_x(g(y))g''(y). \tag{4}$$

Due to the presence of the second term on the right hand side of (4) the relationship $\widehat{x} = g(\widehat{y})$ no longer holds. Thus the value of x obtained by maximizing $p_x(x)$ will not be the value obtained by transforming to $p_y(y)$ then maximizing with respect to y and then transforming back to x. This causes modes of densities to be dependent on the choice of variables. In the case of linear transformation, the second term on

在(4)中，因为有第二项，所以左边等于0时 $p'_x(g(y))$ 不一定等于0，因此两个极值不一定同时达到。

需要注意的是，当 $x = g(y)$ 为线性变化时，$g''(y) = 0$ ，因此上面式子里的第二项就没有了，此时关系保持。

Figure 1 Example of the transformation of the mode of a density under a non-linear change of variables, illustrating the different behaviour compared to a simple function. See the text for details.



上图中，从x 变换到 y经历了一个非线性变换。如果不考虑jacobian factor，应该是红线转移到绿线，最值保持函数关系。但是因为有jacobian factor，实际上转移到了紫色的线，最值并不符合函数关系

---

cumulative distribution function:

$$P(z) = \int_{-\infty}^{z} p(x)\,\mathrm{d}x \tag{1.28}$$

The sum and product rules

$$p(x) = \int p(x, y)\,\mathrm{d}y \tag{1.31}$$

$$p(x, y) = p(y|x)p(x). \tag{1.32}$$

1.2.2 Expectations and covariances

Expectation of f(x):

$$\mathbb{E}[f] = \sum_{x} p(x)f(x) \tag{1.33}$$

$$\mathbb{E}[f] = \int p(x)f(x)\,\mathrm{d}x. \tag{1.34}$$

可以用有限的N次sample近似求expectation：

$$\mathbb{E}[f] \simeq \frac{1}{N}\sum_{n=1}^{N} f(x_n). \tag{1.35}$$

We shall make extensive use of this result when we discuss sampling methods in Chapter 11. The approximation in (1.35) becomes **exact** in the limit $N \to \infty$.

用下标表示which variable is being averaged over，
在 $\mathbb{E}_x f(x,y)$ 中，是对 x 取平均， $\mathbb{E}_x f(x,y)$ will be a function of y .

Conditional Expectation

$$\mathbb{E}_x[f|y] = \sum_x p(x|y) f(x) \tag{1.37}$$

Variance:

$$\mathrm{var}[f] = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right] \tag{1.38}$$

$$\mathrm{var}[f] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2. \tag{1.39}$$

$$\begin{aligned} \mathrm{cov}[x,y] &= \mathbb{E}_{x,y}\left[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}\right] \\ &= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y] \end{aligned} \tag{1.41}$$

$$\begin{aligned} \mathrm{cov}[\mathbf{x},\mathbf{y}] &= \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\{\mathbf{x} - \mathbb{E}[\mathbf{x}]\}\{\mathbf{y}^{\mathrm{T}} - \mathbb{E}[\mathbf{y}^{\mathrm{T}}]\}\right] \\ &= \mathbb{E}_{\mathbf{x},\mathbf{y}}[\mathbf{x}\mathbf{y}^{\mathrm{T}}] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}^{\mathrm{T}}]. \end{aligned} \tag{1.42}$$

If we consider the covariance of the components of a vector x with each other, then we use a slightly simpler notation $cov[\mathbf{x}] \equiv cov[\mathbf{x},\mathbf{x}]$.

1.2.3 Bayesian probabilities

对于不可重复的实验，如冰川会不会融化，我们不能通过频率来描述uncertainty，这时候要通过probability来描述。此时每当我们掌握了一些新的证据，都会对原有的估计加以修正，这就是bayesian的思路。

prior, posterior, likelihood之间的关系，这个应该看了好多遍了：

by incorporating the evidence provided by the observed data. As we shall see in detail later, we can adopt a similar approach when making inferences about quantities such as the parameters $\mathbf{w}$ in the polynomial curve fitting example. We capture our assumptions about $\mathbf{w}$, before observing the data, in the form of a prior probability distribution $p(\mathbf{w})$. The effect of the observed data $\mathcal{D} = \{t_1, \ldots, t_N\}$ is expressed through the conditional probability $p(\mathcal{D}|\mathbf{w})$, and we shall see later, in Section 1.2.5, how this can be represented explicitly. Bayes' theorem, which takes the form

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})} \tag{1.43}$$

then allows us to evaluate the uncertainty in $\mathbf{w}$ *after* we have observed $\mathcal{D}$ in the form of the posterior probability $p(\mathbf{w}|\mathcal{D})$.

$$\text{posterior} \propto \text{likelihood} \times \text{prior} \tag{1.44}$$

of these quantities are viewed as functions of $\mathbf{w}$. The denominator in

分母是normalization term，因为如果左右两边同时对w积分：

$$p(\mathcal{D}) = \int p(\mathcal{D}|\mathbf{w})p(\mathbf{w})\,\mathrm{d}\mathbf{w}. \tag{1.45}$$

In a frequentist setting, w is considered to be a **fixed** parameter, whose value is determined by some form of 'estimator', and error bars on this estimate are obtained by considering the distribution of possible data sets $\mathcal{D}$.
By contrast, from the Bayesian viewpoint there is only a single data set $\mathcal{D}$ (namely the one that is actually observed), and the **uncertainty in the parameters is expressed through a probability distribution over** w**.**

1.2.4 The Gaussian distribution

$$\mathcal{N}\left(x|\mu,\sigma^2\right) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{1.46}$$

The square root of the variance σ, is called the *standard deviation*
β = 1/σ2, is called the *precision*

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) x\,\mathrm{d}x = \mu. \tag{1.49}$$

$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) x^2 \, \mathrm{d}x = \mu^2 + \sigma^2. \tag{1.50}$$

$$\mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2 \tag{1.51}$$

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \tag{1.52}$$

where the D-dimensional vector $\boldsymbol{\mu}$ is called the mean, the $D \times D$ matrix $\boldsymbol{\Sigma}$ is called the covariance, and $|\boldsymbol{\Sigma}|$ denotes the determinant of $\boldsymbol{\Sigma}$. We shall make use of the multivariate Gaussian distribution briefly in this chapter, although its properties will be studied in detail in Section 2.3.

One common criterion for determining the parameters in a probability distribution using an observed data set is to find the parameter values that maximize the likelihood function.
This might seem like a strange criterion because, from our foregoing discussion of probability theory, **it would seem more natural to maximize the probability of the parameters given the data, not the probability of the data given the parameters.**

MLE得到的μ_{ML}就是sample mean，σ_{ML}就是sample variance

$$\ln p\left(\mathbf{x}|\mu,\sigma^2\right) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi). \tag{1.54}$$

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n \tag{1.55}$$

$$\sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2 \tag{1.56}$$

maximum likelihood approach systematically underestimates the variance of the distribution.

$$\mathbb{E}[\mu_{\mathrm{ML}}] = \mu \tag{1.57}$$

$$\mathbb{E}[\sigma^2_{\mathrm{ML}}] = \left(\frac{N-1}{N}\right)\sigma^2 \tag{1.58}$$

1.2.5 Curve fitting re-visited

we shall assume that, given the value of x, the corresponding value of t has a Gaussian distribution with a mean equal to the value y(x, w) of the polynomial curve

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}\left(t|y(x, \mathbf{w}), \beta^{-1}\right) \tag{1.60}$$

写出likelihood：

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}\left(t_n|y(x_n, \mathbf{w}), \beta^{-1}\right). \tag{1.61}$$

求log，后两项与w无关

$$\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = -\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi). \tag{1.62}$$

the sum-of-squares error function has arisen as a consequence of **maximizing likelihood under the assumption of a Gaussian noise distribution**

同时，对β求导，可以得出

the Gaussian conditional distribution. Maximizing (1.62) with respect to β gives

$$\frac{1}{\beta_{\mathrm{ML}}} = \frac{1}{N}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}_{\mathrm{ML}}) - t_n\}^2. \tag{1.63}$$

此时就可以用这个分布进行预测了

$$p(t|x, \mathbf{w}_{\mathrm{ML}}, \beta_{\mathrm{ML}}) = \mathcal{N}\left(t|y(x, \mathbf{w}_{\mathrm{ML}}), \beta^{-1}_{\mathrm{ML}}\right). \tag{1.64}$$

---

introduce a prior distribution over the polynomial coefficients w

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right\} \tag{1.65}$$

此时likelihood：

$$p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha). \tag{1.66}$$

可以看到相当于加入了正则项：

$$\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}. \tag{1.67}$$

This technique is called **maximum posterior**, or simply **MAP**.

1.2.6 Bayesian curve fitting

虽然前面求出了w的后验，但这不能算是完成的bayesian treatment。we should consistently apply the sum and product rules of probability, which requires, as we shall see shortly, that we integrate over all values of w。

We therefore wish to evaluate the predictive distribution $p(t|x, \mathbf{x}, \mathbf{t})$(这里设$\alpha$和$\beta$已知)

$$p(t|x, \mathbf{x}, \mathbf{t}) = \int p(t|x, \mathbf{w})p(\mathbf{w}|\mathbf{x}, \mathbf{t})\,\mathrm{d}\mathbf{w}. \tag{1.68}$$

其中

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}\left(t|y(x, \mathbf{w}), \beta^{-1}\right) \tag{1.60}$$

Here p(w|x, t) is the posterior distribution，通过$\frac{prior \times likelihood}{normalization\ term}$得到，section3.3中可知，对于curve fitting，posterior也是一个gaussian

此外更进一步。预测结果也是一个gaussian：

$$p(t|x, \mathbf{x}, \mathbf{t}) = \mathcal{N}\left(t|m(x), s^2(x)\right) \tag{1.69}$$

where the mean and variance are given by

$$m(x) = \beta \boldsymbol{\phi}(x)^{\mathrm{T}} \mathbf{S} \sum_{n=1}^{N} \boldsymbol{\phi}(x_n) t_n \tag{1.70}$$

$$s^2(x) = \beta^{-1} + \boldsymbol{\phi}(x)^{\mathrm{T}} \mathbf{S} \boldsymbol{\phi}(x). \tag{1.71}$$

Here the matrix $\mathbf{S}$ is given by

$$\mathbf{S}^{-1} = \alpha \mathbf{I} + \beta \sum_{n=1}^{N} \boldsymbol{\phi}(x_n) \boldsymbol{\phi}(x)^{\mathrm{T}} \tag{1.72}$$

where $\mathbf{I}$ is the unit matrix, and we have defined the vector $\boldsymbol{\phi}(x)$ with elements $\phi_i(x) = x^i$ for $i = 0, \ldots, M$.

The first term in (1.71) represents the uncertainty in the predicted value of t due to the noise on the target variables and was expressed already in the maximum likelihood predictive distribution (1.64) through β_{ML}^{-1}. However, the second term arises from the uncertainty in the parameters w and is a consequence of the Bayesian treatment.

1.3. Model Selection

Akaike information criterion, or AIC chooses the model for which the quantity

$$\ln p(\mathcal{D}|\mathbf{w}_{\mathrm{ML}}) - M \tag{1.73}$$

is largest. Here p(D|wML) is the best-fit log likelihood, and **M is the number of adjustable parameters in the model.**

section 4.4.1中要讲到 Bayesian information criterion, or BIC
Such criteria do not take account of the uncertainty in the model parameters,
however, and in practice they tend to favour overly simple models. We therefore turn in Section 3.4 to a fully Bayesian approach where we shall see how complexity penalties arise in a natural and principled way.

1.5. Decision Theory

作用：Here we turn to a discussion of decision theory that, **when combined with probability theory, allows us to make optimal decisions in situations involving uncertainty** such as those encountered in pattern recognition.

Determination of p(x, t) from a set of training data is an example of **inference** and is typically a very difficult problem whose solution forms the subject of much of this book

The optimal solution is the one which minimizes the loss function. However, the loss function depends on the true class, which is unknown. For a given input vector $\mathbf{x}$, our uncertainty in the true class is expressed through the joint probability distribution $p(\mathbf{x}, \mathcal{C}_k)$ and so we seek instead to minimize the average loss, where the average is computed with respect to this distribution, which is given by

$$\mathbb{E}[L] = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k) \, \mathrm{d}\mathbf{x}. \tag{1.80}$$

Each $\mathbf{x}$ can be assigned independently to one of the decision regions $\mathcal{R}_j$. Our goal is to choose the regions $\mathcal{R}_j$ in order to minimize the expected loss (1.80), which implies that for each $\mathbf{x}$ we should minimize $\sum_k L_{kj} p(\mathbf{x}, \mathcal{C}_k)$. As before, we can use

1.5.3 The reject option

In some applications, it will be appropriate to avoid making decisions on the difficult cases in anticipation of a lower error rate on those examples for which a classification decision is made. This is known as the **reject option**.

1.5.4 Inference and decision

We have broken the classification problem down into two separate stages

inference stage in which we use training data to learn a model for p(Ck|x)
decision stage in which we use these posterior probabilities to make optimal class assignments

如果直接learn a function，把输入映射到决策中，就是discriminant function

下面这里讲了三种模型：generative, discrimitive和直接映射到0和1

(a) First solve the inference problem of determining the class-conditional densities $p(\mathbf{x}|\mathcal{C}_k)$ for each class $\mathcal{C}_k$ individually. Also separately infer the prior class probabilities $p(\mathcal{C}_k)$. Then use Bayes' theorem in the form

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})} \tag{1.82}$$

to find the posterior class probabilities $p(\mathcal{C}_k|\mathbf{x})$. As usual, the denominator in Bayes' theorem can be found in terms of the quantities appearing in the numerator, because

$$p(\mathbf{x}) = \sum_k p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k). \tag{1.83}$$

Equivalently, we can model the joint distribution $p(\mathbf{x}, \mathcal{C}_k)$ directly and then normalize to obtain the posterior probabilities. Having found the posterior probabilities, we use decision theory to determine class membership for each new input $\mathbf{x}$. Approaches that explicitly or implicitly model the distribution of inputs as well as outputs are known as *generative models*, because by sampling from them it is possible to generate synthetic data points in the input space.

(b) First solve the inference problem of determining the posterior class probabilities $p(\mathcal{C}_k|\mathbf{x})$, and then subsequently use decision theory to assign each new $\mathbf{x}$ to one of the classes. Approaches that model the posterior probabilities directly are called *discriminative models*.

(c) Find a function $f(\mathbf{x})$, called a discriminant function, which maps each input $\mathbf{x}$ directly onto a class label. For instance, in the case of two-class problems, $f(\cdot)$ might be binary valued and such that $f = 0$ represents class $\mathcal{C}_1$ and $f = 1$ represents class $\mathcal{C}_2$. In this case, probabilities play no role.

1.5.5 Loss functions for regression

之前说的都是分类问题，对于回归问题，我们想要minimize的loss的期望可以表示为：

$$\mathbb{E}[L] = \iint L(t, y(\mathbf{x}))p(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t. \tag{1.86}$$

比如说square loss：

$$\mathbb{E}[L] = \iint \{y(\mathbf{x}) - t\}^2 p(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t. \tag{1.87}$$

下面这一段是对上面这个式子求导，要用到variational calculus（对于函数求导），需要看一下appendix D

Our goal is to choose $y(\mathbf{x})$ so as to minimize $\mathbb{E}[L]$. If we assume a completely flexible function $y(\mathbf{x})$, we can do this formally using the calculus of variations to give

$$\frac{\delta \mathbb{E}[L]}{\delta y(\mathbf{x})} = 2 \int \{y(\mathbf{x}) - t\} p(\mathbf{x}, t)\, \mathrm{d}t = 0. \tag{1.88}$$

Solving for $y(\mathbf{x})$, and using the sum and product rules of probability, we obtain

$$y(\mathbf{x}) = \frac{\int t p(\mathbf{x}, t)\, \mathrm{d}t}{p(\mathbf{x})} = \int t p(t|\mathbf{x})\, \mathrm{d}t = \mathbb{E}_t[t|\mathbf{x}] \tag{1.89}$$

which is the conditional average of t conditioned on x and is known as the **regression function**

可以看到regression function $\mathbb{E}_t[t|x]$，可以minimize square loss的期望，是由$p(t|x)$的均值得到的

因为$\mathbb{E}_t[t|x]$是最优解， $y(x) - \mathbb{E}[t|x]$的期望是0，因此交叉项消失

$$\begin{aligned}\{y(\mathbf{x}) - t\}^2 &= \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}] + \mathbb{E}[t|\mathbf{x}] - t\}^2 \\ &= \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 + 2\{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}\{\mathbb{E}[t|\mathbf{x}] - t\} + \{\mathbb{E}[t|\mathbf{x}] - t\}^2\end{aligned}$$

where, to keep the notation uncluttered, we use $\mathbb{E}[t|\mathbf{x}]$ to denote $\mathbb{E}_t[t|\mathbf{x}]$. Substituting into the loss function and performing the integral over t, we see that the cross-term vanishes and we obtain an expression for the loss function in the form

$$\mathbb{E}[L] = \int \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 p(\mathbf{x})\, \mathrm{d}\mathbf{x} + \int \{\mathbb{E}[t|\mathbf{x}] - t\}^2 p(\mathbf{x})\, \mathrm{d}\mathbf{x}. \tag{1.90}$$

而由(1.90)的前半部分也可以得出之前的结论，即最优解是由conditional mean $\mathbb{E}_t[t|x]$得出的

第二项则是target的variance，因此预测出的target的不同可以看作噪声。而因为这项与y(x)无关，因此这个方差是去不掉的

类比分类问题，回归问题也可以分为三类：

(a) First solve the inference problem of determining the joint density $p(\mathbf{x}, t)$. Then normalize to find the conditional density $p(t|\mathbf{x})$, and finally marginalize to find the conditional mean given by (1.89).

(b) First solve the inference problem of determining the conditional density $p(t|\mathbf{x})$, and then subsequently marginalize to find the conditional mean given by (1.89).

(c) Find a regression function $y(\mathbf{x})$ directly from the training data.

1.6. Information Theory

引出$h(x)$:
我们用$h(x)$描述degree of surprise，显然$h(x)$与$p(x)$相关
当x和y独立时，我们观察x的surprise+观察y的surprise应该等于同时观察x和y的surprise，即
$h(x,y)=h(x)+h(y)$
而又有$p(x,y)=p(x)\cdot(y)$
因此可以定义information

$$h(x)=-\log_2 p(x) \tag{1.92}$$

传输一个变量x的$h(x)$的期望，就是x的entropy

$$\mathrm{H}[x]=-\sum_x p(x)\log_2 p(x). \tag{1.93}$$

This important quantity is called the *entropy* of the random variable x. Note that $\lim_{p\to 0} p\ln p=0$ and so we shall take $p(x)\ln p(x)=0$ whenever we encounter a value for x such that $p(x)=0$.

之后的讨论中，entropy的底数为e

---

multiplicity:
把N个相同的物体分到n个箱子中的分法:
首先选第一个物体有N种选法，第二个物体有N-1种选法。一共有$N!$种选法
而n个箱子内部本身是无序的，因此最终结果为:

$$W=\frac{N!}{\prod_i n_i!} \tag{1.94}$$

which is called the **multiplicity**

而entropy则是 **logarithm of the multiplicity scaled by an appropriate constant**

$$\mathrm{H} = \frac{1}{N}\ln W = \frac{1}{N}\ln N! - \frac{1}{N}\sum_i \ln n_i!. \tag{1.95}$$

We now consider the limit $N \to \infty$, in which the fractions n_i/N are held fixed, and apply Stirling's approximation

$$\ln N! \simeq N\ln N - N \tag{1.96}$$

which gives

$$\mathrm{H} = -\lim_{N\to\infty}\sum_i \left(\frac{n_i}{N}\right)\ln\left(\frac{n_i}{N}\right) = -\sum_i p_i \ln p_i \tag{1.97}$$

离散值的熵：

$$\mathrm{H}_\Delta = -\sum_i p(x_i)\Delta \ln\left(p(x_i)\Delta\right) = -\sum_i p(x_i)\Delta \ln p(x_i) - \ln\Delta \tag{1.102}$$

对于连续值，differential entropy：

$$\lim_{\Delta\to 0}\left\{\sum_i p(x_i)\Delta \ln p(x_i)\right\} = -\int p(x)\ln p(x)\,\mathrm{d}x \tag{1.103}$$

where the quantity on the right-hand side is called the *differential entropy*. We see

对于离散变量，用拉格朗日maximize extropy得到uniform distribution

对于连续变量，用拉格朗日maximize extropy得到Gaussian distribution

---

conditional entropy：
如果对于变量x和y，我们先观察到了y，那么观察x得到的信息量为−ln p(y|x)，x此时的条件熵为：

$$\mathrm{H}[\mathbf{y}|\mathbf{x}] = -\iint p(\mathbf{y},\mathbf{x})\ln p(\mathbf{y}|\mathbf{x})\,\mathrm{d}\mathbf{y}\,\mathrm{d}\mathbf{x} \tag{1.111}$$

条件熵满足

$$\mathrm{H}[\mathbf{x},\mathbf{y}] = \mathrm{H}[\mathbf{y}|\mathbf{x}] + \mathrm{H}[\mathbf{x}] \tag{1.112}$$

where $H[x, y]$ is the differential entropy of $p(x, y)$ and $H[x]$ is the differential entropy of the marginal

distribution $p(x)$

1.6.1 Relative entropy and mutual information

Consider some unknown distribution $p(x)$, and suppose that we have modelled this using an approximating distribution $q(x)$.
If we use $q(x)$ to construct a coding scheme for the purpose of transmitting values of x to a receiver, then the average **additional** amount of information (in nats) required to specify the value of x (assuming we choose an efficient coding scheme) as a result of using $q(x)$ instead of the true distribution $p(x)$ is given by

$$\begin{aligned} \mathrm{KL}(p\|q) &= -\int p(\mathbf{x}) \ln q(\mathbf{x})\,\mathrm{d}\mathbf{x} - \left(-\int p(\mathbf{x}) \ln p(\mathbf{x})\,\mathrm{d}\mathbf{x}\right) \\ &= -\int p(\mathbf{x}) \ln \left\{\frac{q(\mathbf{x})}{p(\mathbf{x})}\right\} \mathrm{d}\mathbf{x}. \end{aligned} \tag{1.113}$$

This is known as the relative entropy or Kullback-Leibler divergence,or KL divergence
KL(p||q) >=0 with equality if, and only if, p(x)= q(x).

后面要用到jensen不等式，因此先说明凸函数的定义：

$$f(\lambda a + (1-\lambda)b) \leqslant \lambda f(a) + (1-\lambda) f(b). \tag{1.114}$$

This is equivalent to the requirement that the second derivative of the function be everywhere positive

Using the technique of proof by induction, we can show from (1.114) that a convex function $f(x)$ satisfies

$$f\left(\sum_{i=1}^{M} \lambda_i x_i\right) \leqslant \sum_{i=1}^{M} \lambda_i f(x_i) \tag{1.115}$$

where $\lambda_i \geqslant 0$ and $\sum_i \lambda_i = 1$, for any set of points $\{x_i\}$. The result (1.115) is known as *Jensen's inequality*. If we interpret the λ_i as the probability distribution over a discrete variable x taking the values $\{x_i\}$, then (1.115) can be written

$$f\left(\mathbb{E}[x]\right) \leqslant \mathbb{E}[f(x)] \tag{1.116}$$

where $\mathbb{E}[\cdot]$ denotes the expectation. For continuous variables, Jensen's inequality takes the form

$$f\left(\int \mathbf{x} p(\mathbf{x})\,\mathrm{d}\mathbf{x}\right) \leqslant \int f(\mathbf{x}) p(\mathbf{x})\,\mathrm{d}\mathbf{x}. \tag{1.117}$$

we can interpret the Kullback-Leibler divergence as a measure of the dissimilarity of the two distributions p(x) and q(x).

KL divergence实际上求不出来，因为我们不知道真正的p(x).
但是我们知道由p(x)产生的N个x
因此p(x)可以用$p(\mathrm{x})$近似

$$\mathrm{KL}(p\|q) \simeq \sum_{n=1}^{N} \{-\ln q(\mathbf{x}_n|\boldsymbol{\theta}) + \ln p(\mathbf{x}_n)\}\,. \tag{1.119}$$

The second term on the right-hand side of (1.119) is independent of $\boldsymbol{\theta}$, and the first term is the negative log likelihood function for $\boldsymbol{\theta}$ under the distribution $q(\mathbf{x}|\boldsymbol{\theta})$ evaluated using the training set. Thus we see that minimizing this Kullback-Leibler divergence is equivalent to maximizing the likelihood function.

Thus we see that **minimizing this Kullback-Leibler divergence is equivalent to maximizing the likelihood function**

Mutual information

对于joint distribution中的一组变量x和y，如果它们不独立，就不能表示为p(x, y)= p(x)p(y)
但是，我们可以用KL divergence衡量它们之间“有多么不独立”

$$\begin{aligned} \mathrm{I}[\mathbf{x},\mathbf{y}] &\equiv \mathrm{KL}(p(\mathbf{x},\mathbf{y})\|p(\mathbf{x})p(\mathbf{y})) \\ &= -\iint p(\mathbf{x},\mathbf{y}) \ln\left(\frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x},\mathbf{y})}\right) \mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} \end{aligned} \tag{1.120}$$

which is called the *mutual information* between the variables $\mathbf{x}$ and $\mathbf{y}$. From the

I(x, y) >= 0 with equality if, and only if, x and y are independent

mutual information is related to the conditional entropy through

$$\mathrm{I}[\mathbf{x},\mathbf{y}] = \mathrm{H}[\mathbf{x}] - \mathrm{H}[\mathbf{x}|\mathbf{y}] = \mathrm{H}[\mathbf{y}] - \mathrm{H}[\mathbf{y}|\mathbf{x}]. \tag{1.121}$$

推导:

$$\begin{aligned} I(\mathbf{x};\mathbf{y}) &= -\iint p(\mathbf{x},\mathbf{y}) \ln p(\mathbf{y})\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} + \iint p(\mathbf{x},\mathbf{y}) \ln p(\mathbf{y}|\mathbf{x})\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} \\ &= -\int p(\mathbf{y}) \ln p(\mathbf{y})\,\mathrm{d}\mathbf{y} + \iint p(\mathbf{x},\mathbf{y}) \ln p(\mathbf{y}|\mathbf{x})\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} \\ &= H(\mathbf{y}) - H(\mathbf{y}|\mathbf{x}). \end{aligned} \tag{58}$$

Thus we can view the mutual information as the reduction in the uncertainty about x by virtue of being told the value of y (or vice versa).
我们可以把mutual information看作是观察到y之后，对于x的uncertainty减少了多少（通过entropy和conditional entropy来衡量）
或者我们可以把p(x)看作先验，p(x|y)看作是后验。mutual information表示观测到y之后the reduction in uncertainty about x