习题 1

1.1　(⋆) 思考一下 (1.2) 中给出的平方和误差函数，其中函数 $y(x, \mathbf{w})$ 为 (1.1) 中的多项式。证明使得误差函数最小的系数 $\mathbf{w} = \{w_i\}$ 是如下线性方程组的解：

$$\sum_{j=0}^{M} A_{ij} w_j = T_i \tag{0.1}$$

其中，

$$A_{ij} = \sum_{n=1}^{N} (x_n)^{i+j}, T_i = \sum_{n=1}^{N} (x_n)^i t_n \tag{0.2}$$

以上等式中，i 和 j 作为下标出现时表示向量分量的索引序数，而 $(x)^i$ 这样的写法表示 x 的 i 次方。

解： 根据 (1.1)，

$$y(x_n, \mathbf{w}) = \sum_{j=0}^{M} w_j x_n^j$$

于是将其带入 (1.2) 中，有：

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \left\{ \sum_{j=0}^{M} w_j x_n^j - t_n \right\}^2$$

对 w_i 求导，则

$$\begin{aligned} \frac{\partial E(\mathbf{w})}{\partial w_i} &= \frac{1}{2} \sum_{n=1}^{N} 2 \left\{ \sum_{j=0}^{M} w_j x_n^j - t_n \right\} \frac{\partial \left(\sum_{j=0}^{M} w_j x_n^j - t_n \right)}{\partial w_i} \\ &= \sum_{n=1}^{N} \left\{ \sum_{j=0}^{M} w_j x_n^j - t_n \right\} x_n^i \\ &= \sum_{n=1}^{N} \sum_{j=0}^{M} \left\{ w_j x_n^{i+j} - x_n^i t_n \right\} \end{aligned}$$

令 $\partial E(\mathbf{w}) / \partial w_i = 0$，稍作调整：

$$\sum_{j=0}^{M} \sum_{n=1}^{N} x_n^{i+j} w_j = \sum_{n=1}^{N} x_n^i t_n$$

即题目中欲证明的等式，证毕。

1.2　(⋆) 写出一组形如 (1.122) 的线性方程组，要求使得 (1.4) 中的正则化平方和误差函数最小化的系数 w_i 满足该方程组。

解： 本题的步骤与 1.1 几乎完全一致，只是最后结果多了一项，当且仅当 $i = j$ 时，

$$\sum_{j=0}^{M} \left[\sum_{n=1}^{N} (x_n)^{i+j} + \lambda \right] w_j = \sum_{n=1}^{N} (x_n)^i t_n$$

否则

$$\sum_{j=0}^{M} \left[\sum_{n=1}^{N} (x_n)^{i+j} \right] w_j = \sum_{n=1}^{N} (x_n)^i t_n$$

故可将最后结果写成

$$\sum_{j=0}^{M} \widetilde{A}_{ij} w_j = T_i$$

其中

$$\widetilde{A}_{ij} = A_{ij} + \lambda I_{ij}$$

I 为单位矩阵。

1.3 ($\star\ \star$) 假设现在有 3 个不同颜色的箱子，分别为 r(红色)，b(蓝色) 和 g(绿色)。箱子 r 里有 3 个苹果，4 个橙子和 3 个酸橙；箱子 b 里有 1 个苹果和 1 个橙子，没有酸橙；箱子 g 里有 3 个苹果，3 个橙子和 4 个酸橙。现在要在其中选择一个箱子，然后随机从中拿走一个水果，假如选择到这几个箱子的概率分别是 $p(r) = 0.2$，$p(b) = 0.2$ 和 $p(g) = 0.6$，拿到箱子中每个水果的可能性都相等，那么拿到一个苹果的可能性是多少？假设我们已经得知拿到的水果是一个橙子，那么它来自于绿色箱子的概率是多少？

解：

$$p(apple) = 0.2 \times \frac{3}{3+4+3} + 0.2 \times \frac{1}{1+1} + 0.6 \times \frac{3}{3+3+4} = 0.34$$

$$p(g|orange) = \frac{p(orange|g)p(g)}{p(orange)} = \frac{0.6 \times \frac{3}{10}}{0.2 \times \frac{4}{10} + 0.2 \times \frac{1}{2} + 0.6 \times \frac{3}{10}} = \frac{1}{2}$$

1.4 ($\star\ \star$) 假设有一个定义在连续变量 x 上的概率密度函数 $p_x(x)$，并且进行了一个非线性变换 $x = g(y)$，那么概率密度的变换是根据公式 (1.27) 得到的。对 (1.27) 微分，证明由于 Jacobian 因子的关系，使得 y 的概率密度取得最大值的 $\hat{y}$ 与使得 x 的概率密度取得最大值的 $\hat{x}$ 通常不会是一个简单的函数关系 $\hat{x} = g(\hat{y})$。这表明了一个事实：与简单的函数相比，概率密度函数的最大值与变量的选择是相关的。请证明在线性变换下，概率密度函数最大值取值位置的变换关系与变量变换的关系是相同的。

解：

1.5 ($\star$) 使用定义 (1.38) 证明 $\mathrm{var}[f(x)]$ 满足 (1.39)。

解：

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

中间的步骤中，由于 $\mathbb{E}[f(x)])^2$ 为常数，所以可以直接提出，后面的按照顺序计算即可。

1.6 ($\star$) 证明：如果两个变量 x 和 y 是相互独立的，那么它们的协方差为 0。

解： 即要证明在 x 与 y 相互独立时 $\mathrm{cov}[x, y] = 0$。根据定义，$\mathrm{cov}[x, y] = \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y]$。

对于离散变量，

$$\begin{aligned}
\mathbb{E}[xy] &= \sum_{x,y} p(x, y)xy \\
&= \sum_{x} xp(x) \sum_{y} yp(y) \\
&= \mathbb{E}[x]\mathbb{E}[y]
\end{aligned}$$

于是 $\mathrm{cov}[x, y] = 0$。

对于连续变量，

$$\begin{aligned}\mathbb{E}[xy] &= \int_{x,y} p(x, y)xy \, \mathrm{d}x\mathrm{d}y \\ &= \int_x xp(x) \, \mathrm{d}x \int_y yp(y) \, \mathrm{d}y \\ &= \mathbb{E}[x]\mathbb{E}[y]\end{aligned}$$

同样有 $\mathrm{cov}[x, y] = 0$。

1.7 (⋆ ⋆) 在这个练习中，我们将会证明一元高斯分布满足归一化条件 (1.48)。为了证明这个问题，以下积分

$$I = \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}x^2\right) \mathrm{d}x \tag{0.3}$$

可以通过写成其平方形式来计算：

$$I^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}x^2 - \frac{1}{2\sigma^2}y^2\right) \mathrm{d}x \, \mathrm{d}y \tag{0.4}$$

现在再将笛卡尔坐标 (x, y) 转化为极坐标 (r, θ)，并用 $u = r^2$ 进行替换。证明：通过对 θ 和 u 积分，然后在等式两边同时取平方根，可以得到：

$$I = (2\pi\sigma^2)^{1/2} \tag{0.5}$$

最后再利用这个结果证明高斯分布 $\mathcal{N}(x|\mu, \sigma^2)$ 满足归一化条件。

1.8 (⋆ ⋆) 通过对变量进行变换，证明一元高斯分布 (1.46) 满足 (1.49)。然后，通过对归一化条件

$$\int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2)\mathrm{d}x = 1 \tag{0.6}$$

的等式两侧关于 σ^2 求微分，证明高斯分布满足 (1.50)。最后证明 (1.51) 成立。

1.9 (⋆) 证明高斯分布 (1.46) 的模 (即最大值) 在 μ 处取得。类似地，证明多元高斯分布 (1.52) 的模在 $\boldsymbol{\mu}$ 处取得。

1.10 (⋆) 假设两个变量 x 和 z 统计学独立。证明两个变量加和的均值和方差分别满足

$$\mathbb{E}[x + z] = \mathbb{E}[x] + \mathbb{E}[z] \tag{0.7}$$

$$\mathrm{var}[x + z] = \mathrm{var}[x] + \mathrm{var}[z] \tag{0.8}$$

1.11 (⋆) 分别令对数似然函数 (1.54) 关于 μ 和 σ^2 的导数等于 0，证明 (1.55) 和 (1.56)。

1.12 (⋆ ⋆) 利用 (1.49) 和 (1.50)，证明

$$\mathbb{E}[x_n x_m] = \mu^2 + I_{nm}\sigma^2 \tag{0.9}$$

其中 x_n 和 x_m 分别表示从均值为 μ，方差为 σ^2 的高斯分布中抽取的数据样本，I_{nm} 在 $n = m$ 时等于 1，否则等于 0。证明 (1.57) 和 (1.58)。

1.13 (⋆) 假设利用 (1.56) 估计高斯分布的方差，但使用均值的最大似然估计 μ_{ML} 替换了真实的均值 μ。证明这个估计的期望等于真实的方差 σ^2。

1.14 ($\star\star$) 证明元素为 w_{ij} 的任意方阵可以写成 $w_{ij} = w_{ij}^{\mathrm{S}} + w_{ij}^{\mathrm{A}}$ 的形式，其中 w_{ij}^{S} 和 w_{ij}^{A} 分别为对称矩阵和反对称矩阵，即对于一切 i 和 j，$w_{ij}^{\mathrm{S}} w_{ji}^{\mathrm{S}}$，$w_{ij}^{\mathrm{A}} = -w_{ji}^{\mathrm{A}}$。现在考虑在 D 维空间中高阶多项式的二次项

$$\sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij} x_i x_j \tag{0.10}$$

证明：

$$\sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij} x_i x_j = \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}^{\mathrm{S}} x_i x_j \tag{0.11}$$

也就是说反对称矩阵不会产生什么影响。所以我们可以看出，在一般情况下，矩阵系数 w_{ij} 可以选择成对称的，同时这个 D^2 矩阵中，并非所有的元素都是相互独立地选取的。证明矩阵 $w_{ij}^{\mathrm{S}} w_{ji}^{\mathrm{S}}$ 中独立参数的个数为 $D(D+1)/2$。

1.15 ($\star\star\star$) 在本练习以及接下来的练习中，我们将要研究多项式里相互独立的参数数量会随着多项式阶数 M 和输入空间维度 D 的变化而变化。我们从写出 D 维空间多项式的 M 阶项开始：

$$\sum_{i_1=1}^{D}\sum_{i_2=1}^{D}\cdots\sum_{i_M=1}^{D} w_{i_1 i_2 \ldots i_M} x_{i_1} x_{i_2} \ldots x_{i_M} \tag{0.12}$$

系数 $w_{i_1 i_2 \ldots i_M}$ 包含了 D^M 个元素，但由于 $x_{i_1} x_{i_2} \ldots x_{i_M}$ 中互换对称性的存在，相互独立的参数数量要远小于 D^M。首先证明系数的冗余性可以通过将 M 阶项写成如下的形式进行消除：

$$\sum_{i_1=1}^{D}\sum_{i_2=1}^{i_1}\cdots\sum_{i_M=1}^{i_{M-1}} \widetilde{w}_{i_1 i_2 \ldots i_M} x_{i_1} x_{i_2} \ldots x_{i_M} \tag{0.13}$$

注意，系数 $\widetilde{w}$ 与系数 w 之间的关系并不需要精确地表示出来。使用这个结果，证明 M 阶情况下的独立参数数量 $n(D, M)$ 满足如下递归关系：

$$n(D, M) = \sum_{i=1}^{D} n(i, M-1) \tag{0.14}$$

然后利用归纳法证明如下结果成立：

$$\sum_{i=1}^{D} \frac{(i+M-2)!}{(i-1)!(M-1)!} = \frac{(D+M-1)!}{(D-1)!M!} \tag{0.15}$$

这个证明可以通过先证明 $D=1$ 时，对于任意的 M 以上等式均成立，因为 $0!=1$，然后假设在维数为 D 时结论成立，再继续证明维数为 $D+1$ 时结论同样成立。最后，利用以上两个结果，根据归纳法共同证明：

$$n(D, M) = \frac{(D+M-1)!}{(D-1)!M!} \tag{0.16}$$

为了证明这一点，可以先证明 $M=2$ 时，对于任意的 $D \geqslant 1$ 等式均成立，可以与习题 1.14 的结果进行对比来得出这个结论。然后利用 (1.135) 和 (1.136)，证明如果阶数为 $M-1$ 时结论成立，那么在阶数为 M 时同样成立。

1.16 ($\star\star\star$) 在习题 1.15 中，我们证明了 D 维多项式的 M^{th} 阶项中独立参数的数量满足等式 (1.135)。现在我们来求取阶数小于等于 M 的所有项中独立参数的总数 $N(D, M)$。首先证明 $N(D, M)$ 满足：

$$N(D, M) = \sum_{m=0}^{M} n(D, m) \tag{0.17}$$

其中 $n(D,m)$ 为 m 阶项中独立参数的数量。现在使用 (1.137) 来利用归纳法证明：

$$N(D,M)=\frac{(D+M)!}{D!M!} \tag{0.18}$$

可以首先证明 $M=0$ 时对于任意 $D\geqslant 1$ 均成立，然后假设阶数为 M 时成立，再证明阶数为 $M+!$ 时结论成立。最后，对于较大的 n，利用斯特林近似：

$$n!\approx n^n e^{-n} \tag{0.19}$$

证明在 $D\gg M$ 时，$N(D,M)$ 的增长呈 D^M 型，而对于 $M\gg D$，$N(D,M)$ 的增长呈 M^D 型。现在研究一下 D 维空间中的立方多项式 $(M=3)$，并对于以下两种情况计算独立参数的总数：(i) $D=10$ (ii)$D=100$，这两种情况分别对应了小规模和中等规模的机器学习应用问题。

1.17 $(\star\star)$ Gamma 函数的定义为

$$\Gamma(x)\equiv\int_0^\infty u^{x-1}e^{-u}\,\mathrm{d}u \tag{0.20}$$

利用分部积分法，证明 $\Gamma(x+1)=x\Gamma(x)$。并证明 $\Gamma(1)=1$，以及当 x 为整数时 $\Gamma(x+1)=x!$。

1.18 $(\star\star)$ 我们可以利用结论 (1.126) 来推导 D 维空间中单位半径的球体的表面积 S_D 和体积 V_D。为了做到这一点，考虑如下的通过笛卡尔坐标与极坐标的变换关系得到的内容：

$$\prod_{i=1}^{D}\int_{-\infty}^{\infty}e^{-x_i^2}\,\mathrm{d}x_i=S_D\int_0^\infty e^{-r^2}r^{D-1}\,\mathrm{d}r \tag{0.21}$$

利用 (1.141) 给出的 Gamma 函数的定义和 (1.126)，计算等式两侧，可以证明

$$S_D=\frac{2\pi^{D/2}}{\Gamma(D/2)} \tag{0.22}$$

接下来，通过对半径从 0 到 1 进行积分，证明 D 维空间中的单位球体体积为

$$V_D=\frac{S_D}{D} \tag{0.23}$$

最后，利用 $\Gamma(1)=1$ 和 $\Gamma(3/2)=\sqrt{\pi}/2$，来证明 $D=2$ 和 $D=3$ 时 (1.143) 和 (1.144) 成立。

1.19 $(\star\star)$ 接下来研究一个 D 维空间中半径为 a 的球体，与它在一起的还有一个同心的边长为 $2a$ 的超立方体，所以球体与超立方体的每一个面都相切与该面的中心点。利用 (1.18) 中的结果，证明球体和立方体的体积比为

$$\frac{\text{volume of sphere}}{\text{volume of cube}}=\frac{\pi^{D/2}}{D2^{D-1}\Gamma(D/2)} \tag{0.24}$$

然后在 $x\gg 1$ 的情况下，根据斯特林公式：

$$\Gamma(x+1)\approx(2\pi)^{1/2}e^{-x}x^{x+1/2} \tag{0.25}$$

从而证明当 $D\to\infty$ 时，(1.145) 中的比值趋近于 0。同时证明，超立方体的中心点到它的一个顶点之间的距离与中心点到它的一个面的垂直距离的比值为 $\sqrt{D}$，而且在 $D\to\infty$ 时，这个比值趋近于 ∞。从这个结果我们可以看出，在高维的空间中，一个立方体的大部分体积都集中在边角，也就是说它们看起来是非常“尖锐”的！

1.20 $(\star\star)$ 在这个练习中，我们会研究高维空间中的高斯分布的一些性质。假设一个 D 维空间中的高斯分布为

$$p(\mathbf{x})=\frac{1}{(2\pi\sigma^2)^{D/2}}exp\left(-\frac{\|\mathbf{x}\|^2}{2\sigma^2}\right) \tag{0.26}$$

我们希望求出极坐标下关于半径的概率密度，而且是方向变量已经被积分后消失了的那种。为了做到这一点，证明在半径为 r，厚度为 ϵ 的薄壳上，在 $\epsilon \ll 1$ 时的概率密度的积分为 $p(r)\epsilon$，其中

$$p(r) = \frac{S_D r^{D-1}}{(2\pi\sigma^2)^{D/2}} \exp\left(\frac{r^2}{2\sigma^2}\right) \tag{0.27}$$

其中 S_D 为 D 维空间中单位球体的表面积。证明对于较大的 D，函数 $p(r)$ 有单一的驻点 $\hat{r} \approx \sqrt{D}\sigma$。通过计算在 $\sigma \ll \hat{r}$ 时的 $p(\hat{r}+\sigma)$，证明当 D 很大时，

$$p(\hat{r}+\epsilon) = p(\hat{r}) \exp\left(-\frac{\epsilon^2}{\sigma^2}\right) \tag{0.28}$$

这表明 $\hat{r}$ 是半径的概率密度的最大值，而且 $p(r)$ 在远离 $\hat{r}$ 处的最大值时呈长度放缩为 σ 的指数型下降。我们已经看到对于较大的 D，$\sigma \ll \hat{r}$，所以大部分的概率质量都集中在较大半径处的薄壳内。最后，证明概率密度 $p(\mathbf{x})$ 在原点处的值大于半径 $\hat{r}$ 处的值，相差的因子为 $\exp(D/2)$。于是我们可以看到高维高斯分布的概率质量大部分都位于不同于概率密度最高的区域的位置上。高维空间中概率分布的这条性质会在后续章节中模型参数的贝叶斯推断里发挥重要作用。

1.21 $(\star\ \star)$ 研究两个非负数 a 和 b，并证明如果 $a \leqslant b$，则 $a \leqslant (ab)^{1/2}$。使用这个结论证明，如果一个二分类问题的决策域使得误分类的概率最小，那么这个概率满足：

$$p(\text{mistake}) \leqslant \int \{p(\mathbf{x}, \boldsymbol{\mathcal{C}_1})p(\mathbf{x}, \mathcal{C}_2)\}^{1/2}\ \mathrm{d}\mathbf{x} \tag{0.29}$$

1.22 $(\star)$ 给定一个元素为 L_{kj} 的损失矩阵，对于任意的 $\mathbf{x}$，期望风险会在我们选择使得 (1.81) 最小化的分类时取得最小值。证明当损失矩阵为 $L_{kj} = 1 - I_{kj}$ 时，可以使得分类标准退化为选择具有最大后验概率的标准，其中 I_{kj} 为单位矩阵中的元素。这种形式的损失矩阵，它的意义是什么？

1.23 $(\star)$ 推导对于一般的损失矩阵和先验分类概率，期望损失最小化的准则。

1.24 $(\star\ \star)$ 现在研究一个分类问题，