机器学习与深度学习习题集答案-1

本文是机器学习和深度学习习题集（上）的答案，免费提供给大家，也是《机器学习-原理、算法与应用》一书的配套产品。此习题集可用于高校的机器学习与深度学习教学，以及在职人员面试准备时使用。

第 2 章 数学知识

1.计算下面函数的一阶导数和二阶导数

$$f\left(x\right)=x\ln x-\frac{1+e^{2x}}{1-e^{2x}}$$

根据基本函数，复合函数，四则运算的求导公式有

$$\begin{aligned}f^{'}\left(x\right)&=x^{'}\ln x+x\left(\ln x\right)^{'}-\frac{\left(1+e^{2x}\right)^{'}\left(1-e^{2x}\right)-\left(1+e^{2x}\right)\left(1-e^{2x}\right)^{'}}{\left(1-e^{2x}\right)^{2}}\\&=\ln x+1-\frac{\left(1+e^{2x}\right)\left(2x\right)^{'}\left(1-e^{2x}\right)+\left(1+e^{2x}\right)\left(1-e^{2x}\right)\left(2x\right)^{'}}{\left(1-e^{2x}\right)^{2}}\\&=\ln x+1-\frac{\left(1+e^{2x}\right)\left(2x\right)^{'}\left(1-e^{2x}\right)+\left(1+e^{2x}\right)\left(1-e^{2x}\right)\left(2x\right)^{'}}{\left(1-e^{2x}\right)^{2}}\\&=\ln x+1-\frac{4\left(1+e^{2x}\right)\left(1-e^{2x}\right)}{\left(1-e^{2x}\right)^{2}}\\&=\ln x+1-4\frac{1+e^{2x}}{1-e^{2x}}\end{aligned}$$

2.计算下面两个向量的内积

$$\mathbf{x}=\begin{bmatrix}1 & 2 & 3\end{bmatrix}$$
$$\mathbf{y}=\begin{bmatrix}-1 & 5 & 10\end{bmatrix}$$

根据内积的定义

$$\mathbf{x}^{\mathrm{T}}\mathbf{y}=1\times\left(-1\right)+2\times5+3\times10=39$$

3.计算下面向量的 1 范数和 2 范数

$$\mathbf{x}=\begin{bmatrix}1 & -2 & 3\end{bmatrix}$$

其 1 范数为

$$\|\mathbf{x}\|_1=|1|+|-2|+|3|=6$$

其 2 范数为

$$\|\mathbf{x}\|_2=\sqrt{1^2+(-2)^2+3^2}=14$$

4.计算下面两个矩阵的乘积：

$$\mathbf{A}=\begin{bmatrix}1 & 2 & 1\\ 0 & 1 & 1\end{bmatrix}$$

$$\mathbf{B}=\begin{bmatrix}1 & 2 & 1 & 1\\ 5 & 4 & 0 & 1\\ 7 & 6 & 1 & 0\end{bmatrix}$$

根据矩阵乘法的定义有

$$\mathbf{AB}=\begin{bmatrix}1\times1+2\times5+1\times7 & 1\times2+2\times4+1\times6 & 1\times1+2\times0+1\times1 & 1\times1+2\times1+1\times0\\ 0\times1+1\times5+1\times7 & 0\times2+1\times4+1\times6 & 0\times1+1\times0+1\times1 & 0\times1+1\times1+1\times0\end{bmatrix}$$

$$=\begin{bmatrix}18 & 16 & 2 & 3\\ 12 & 10 & 1 & 1\end{bmatrix}$$

5.计算下面多元函数的偏导数

$$f(x_1,x_2,x_3)=\ln\left(1+e^{-2x_1+3x_2-4x_3}\right)$$

对各个变量的偏导数分别为

$$\begin{aligned}\frac{\partial f}{\partial x_1}&=\frac{1}{1+e^{-2x_1+3x_2-4x_3}}\frac{\partial f}{\partial x_1}\left(1+e^{-2x_1+3x_2-4x_3}\right)\\&=\frac{e^{-2x_1+3x_2-4x_3}}{1+e^{-2x_1+3x_2-4x_3}}\frac{\partial f}{\partial x_1}\left(-2x_1+3x_2-4x_3\right)\\&=-\frac{2e^{-2x_1+3x_2-4x_3}}{1+e^{-2x_1+3x_2-4x_3}}\\\frac{\partial f}{\partial x_2}&=\frac{1}{1+e^{-2x_1+3x_2-4x_3}}\frac{\partial f}{\partial x_2}\left(1+e^{-2x_1+3x_2-4x_3}\right)\\&=\frac{e^{-2x_1+3x_2-4x_3}}{1+e^{-2x_1+3x_2-4x_3}}\frac{\partial f}{\partial x_2}\left(-2x_1+3x_2-4x_3\right)\\&=\frac{3e^{-2x_1+3x_2-4x_3}}{1+e^{-2x_1+3x_2-4x_3}}\\\frac{\partial f}{\partial x_3}&=\frac{1}{1+e^{-2x_1+3x_2-4x_3}}\frac{\partial f}{\partial x_3}\left(1+e^{-2x_1+3x_2-4x_3}\right)\\&=\frac{e^{-2x_1+3x_2-4x_3}}{1+e^{-2x_1+3x_2-4x_3}}\frac{\partial f}{\partial x_3}\left(-2x_1+3x_2-4x_3\right)\\&=-\frac{4e^{-2x_1+3x_2-4x_3}}{1+e^{-2x_1+3x_2-4x_3}}\end{aligned}$$

6.计算下面多元函数的梯度

$$f\left(x_1,x_2,x_3\right)=\ln\left(1+e^{-2x_1+3x_2-4x_3}\right)$$

根据上一题的结果

$$\nabla f=\begin{bmatrix}\frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \frac{\partial f}{\partial x_3}\end{bmatrix}=\begin{bmatrix}-\frac{2e^{-2x_1+3x_2-4x_3}}{1+e^{-2x_1+3x_2-4x_3}}\\\frac{3e^{-2x_1+3x_2-4x_3}}{1+e^{-2x_1+3x_2-4x_3}}\\-\frac{4e^{-2x_1+3x_2-4x_3}}{1+e^{-2x_1+3x_2-4x_3}}\end{bmatrix}$$

7.计算下面多元函数的雅克比矩阵

$$\begin{aligned}y_1&=x_1^2-\ln x_2+e^{x_1x_3}\\y_2&=\sin\left(x_1x_2\right)+x_3^2\end{aligned}$$

根据雅克比矩阵的定义有

$$\begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \frac{\partial y_1}{\partial x_3} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \frac{\partial y_2}{\partial x_3} \end{bmatrix} = \begin{bmatrix} 2x_1 + x_3 e^{x_1 x_3} & -\frac{1}{x_2} & x_1 e^{x_1 x_3} \\ x_2 \cos(x_1 x_2) & x_2 \cos(x_1 x_2) & 2x_3 \end{bmatrix}$$

8.计算下面多元函数的 Hessian 矩阵

$$f(x_1, x_2, x_3) = x_1^2 - \ln x_2 + e^{x_1 x_3}$$

首先计算一阶偏导数

$$\frac{\partial f}{\partial x_1} = 2x_1 + x_3 e^{x_1 x_3}$$
$$\frac{\partial f}{\partial x_2} = -\frac{1}{x_2}$$
$$\frac{\partial f}{\partial x_3} = x_1 e^{x_1 x_3}$$

然后计算各二阶偏导数

$$\frac{\partial^2 f}{\partial x_1^2} = \frac{\partial f}{\partial x_1}\left(2x_1 + x_3 e^{x_1 x_3}\right) = 2 + x_3^2 e^{x_1 x_3}$$
$$\frac{\partial^2 f}{\partial x_1 \partial x_2} = \frac{\partial f}{\partial x_2}\left(2x_1 + x_3 e^{x_1 x_3}\right) = 0$$
$$\frac{\partial^2 f}{\partial x_1 \partial x_3} = \frac{\partial f}{\partial x_2}\left(2x_1 + x_3 e^{x_1 x_3}\right) = e^{x_1 x_3} + x_1 x_3 e^{x_3}$$
$$\frac{\partial^2 f}{\partial x_2^2} = \frac{\partial f}{\partial x_2}\left(-\frac{1}{x_2}\right) = \frac{1}{x_2^2}$$
$$\frac{\partial^2 f}{\partial x_2 \partial x_3} = \frac{\partial f}{\partial x_3}\left(-\frac{1}{x_2}\right) = 0$$
$$\frac{\partial^2 f}{\partial x_3^2} = \frac{\partial f}{\partial x_3}\left(x_1 e^{x_1 x_3}\right) = x_1^2 e^{x_1 x_3}$$

根据 Hessian 矩阵的定义有

$$\begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} & \frac{\partial^2 f}{\partial x \partial z} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} & \frac{\partial^2 f}{\partial y \partial z} \\ \frac{\partial^2 f}{\partial z \partial x} & \frac{\partial^2 f}{\partial z \partial y} & \frac{\partial^2 f}{\partial z^2} \end{bmatrix} = \begin{bmatrix} 2 + x_3^2 e^{x_1 x_3} & 0 & e^{x_1 x_3} + x_1 x_3 e^{x_3} \\ 0 & \frac{1}{x_2^2} & 0 \\ e^{x_1 x_3} + x_1 x_3 e^{x_3} & 0 & x_1^2 e^{x_1 x_3} \end{bmatrix}$$

9.计算下面函数的所有极值点，并指明是极大值还是极小值：

$$f\left(x\right)=x^3+2x^2-5x+10$$

函数的一阶导数为

$$f^{'}\left(x\right)=3x^2+4x-5$$

解方程

$$3x^2+4x-5=0$$

可以得到驻点

$$x_1=\frac{-2+\sqrt{19}}{3}$$
$$x_2=\frac{-2-\sqrt{19}}{3}$$

函数的二阶导数为

$$f^{''}\left(x\right)=6x+4$$

在驻点处的二阶导数为

$$f^{''}\left(\frac{-2+\sqrt{19}}{3}\right)=6\times\frac{-2+\sqrt{19}}{3}+4=2\sqrt{19}>0$$
$$f^{''}\left(\frac{-2-\sqrt{19}}{3}\right)=6\times\frac{-2-\sqrt{19}}{3}+4=-2\sqrt{19}<0$$

因此$\frac{-2+\sqrt{19}}{3}$是极小值点，$\frac{-2-\sqrt{19}}{3}$是极大值点。

10.推导多元函数梯度下降法的迭代公式。

根据多元函数泰勒公式，如果忽略一次以上的项，函数在 **x** 点处可以展开为

$$f\left(\mathbf{x}+\Delta\mathbf{x}\right)=f\left(\mathbf{x}\right)+\left(\nabla f\left(\mathbf{x}\right)\right)^{\mathrm{T}}\Delta\mathbf{x}+o\left(\left\|\Delta\mathbf{x}\right\|\right)$$

对上式变形，函数的增量与自变量增量、函数梯度的关系为

$$f\left(\mathbf{x}+\Delta\mathbf{x}\right)-f\left(\mathbf{x}\right)=\left(\nabla f\left(\mathbf{x}\right)\right)^{\mathrm{T}}\Delta\mathbf{x}+o\left(\left\|\Delta\mathbf{x}\right\|\right)$$

如果令$\Delta\mathbf{x}=-\nabla f\left(\mathbf{x}\right)$则有

$$f\left(\mathbf{x}+\Delta\mathbf{x}\right)-f\left(\mathbf{x}\right)\approx-\left(\nabla f\left(\mathbf{x}\right)\right)^{\mathrm{T}}\nabla f\left(\mathbf{x}\right)\leq 0$$

即函数值减小。即有

$$f\left(\mathbf{x}+\Delta\mathbf{x}\right)\leq f\left(\mathbf{x}\right)$$

梯度下降法每次的迭代增量为

$$\Delta\mathbf{x}=-\alpha\nabla f\left(\mathbf{x}\right)$$

其中α为人工设定的接近于的正数，称为步长或学习率。其作用是保证$\mathbf{x}+\Delta\mathbf{x}$在$\mathbf{x}$的邻域内，从而可以忽略泰勒公式中的$o\left(\left\|\Delta\mathbf{x}\right\|\right)$项。

使用该增量则有

$$\left(\nabla f\left(\mathbf{x}\right)\right)^{\mathrm{T}}\Delta\mathbf{x}=-\alpha\left(\nabla f\left(\mathbf{x}\right)\right)^{\mathrm{T}}\left(\nabla f\left(\mathbf{x}\right)\right)\leq 0$$

函数值下降。从初始点$\mathbf{x}_0$开始，反复使用如下迭代公式

$$\mathbf{x}_{k+1}=\mathbf{x}_k-\alpha\nabla f\left(\mathbf{x}_k\right)\quad（1）$$

只要没有到达梯度为$\mathbf{0}$的点，函数值会沿序列$\mathbf{x}_k$递减，最终收敛到梯度为$\mathbf{0}$的点。从$\mathbf{x}_0$出发，用式 1 进行迭代，会形成一个函数值递减的序列$\left\{\mathbf{x}_i\right\}$

$$f\left(\mathbf{x}_0\right)\geq f\left(\mathbf{x}_1\right)\geq f\left(\mathbf{x}_2\right)\geq\ldots\geq f\left(\mathbf{x}_k\right)$$

11.梯度下降法为什么要在迭代公式中使用步长系数？

其作用是保证$\mathbf{x}+\Delta\mathbf{x}$在$\mathbf{x}$的邻域内，即控制增量的步长，从而可以忽略泰勒公式中的$o\left(\left\|\Delta\mathbf{x}\right\|\right)$项。否则不能保证每次迭代时函数值下降。

12.梯度下降法如何判断是否收敛？

迭代终止的条件是函数的梯度值为$\mathbf{0}$（实际实现时是接近于$\mathbf{0}$即可），此时认为已经达到极值点。可以通过判定梯度的二范数是否充分接近于 0 而实现。

13.推导多元函数牛顿法的迭代公式。

根据 Fermat 定理，函数在点$\mathbf{x}$处取得极值的必要条件是梯度为$\mathbf{0}$

$$\nabla f\left(\mathbf{x}\right)=\mathbf{0}$$

对于一般的函数，直接求解此方程组存在困难。对目标函数在$\mathbf{x}_0$处作二阶泰勒展开

$$f\left(\mathbf{x}\right)=f\left(\mathbf{x}_0\right)+\nabla f\left(\mathbf{x}_0\right)^{\mathrm{T}}\left(\mathbf{x}-\mathbf{x}_0\right)+\frac{1}{2}\left(\mathbf{x}-\mathbf{x}_0\right)^{\mathrm{T}}\nabla^2 f\left(\mathbf{x}_0\right)\left(\mathbf{x}-\mathbf{x}_0\right)+o\left(\left\|\mathbf{x}-\mathbf{x}_0\right\|^2\right)$$

忽略二次以上的项，将目标函数近似成二次函数，等式两边同时对$\mathbf{x}$求梯度，可得

$$\nabla f\left(\mathbf{x}\right)\approx\nabla f\left(\mathbf{x}_0\right)+\nabla^2 f\left(\mathbf{x}_0\right)\left(\mathbf{x}-\mathbf{x}_0\right)$$

其中$\nabla^2 f\left(\mathbf{x}_0\right)$为在$\mathbf{x}_0$处的 Hessian 矩阵。令函数的梯度为$\mathbf{0}$，有

$$\nabla f\left(\mathbf{x}_0\right)+\nabla^2 f\left(\mathbf{x}_0\right)\left(\mathbf{x}-\mathbf{x}_0\right)=\mathbf{0}$$

解这个线性方程组可以得到

$$\mathbf{x}=\mathbf{x}_0-\left(\nabla^2 f\left(\mathbf{x}_0\right)\right)^{-1}\nabla f\left(\mathbf{x}_0\right) \quad （1）$$

如果将梯度向量简写为$\mathbf{g}$，Hessian 矩阵简记为$\mathbf{H}$，式 1 可以简写为

$$\mathbf{x}=\mathbf{x}_0-\mathbf{H}^{-1}\mathbf{g} \quad （2）$$

在泰勒公式中忽略了高阶项将函数做了近似，因此这个解不一定是目标函数的驻点，需要反复用式 2 进行迭代。从初始点$\mathbf{x}_0$处开始，计算函数在当前点处的 Hessian 矩阵和梯度向量，然后用下面的公式进行迭代

$$\mathbf{x}_{k+1}=\mathbf{x}_k-\alpha\mathbf{H}_k^{-1}\mathbf{g}_k \quad （3）$$

直至收敛到驻点处。迭代终止的条件是梯度的模接近于$\mathbf{0}$，或达到指定的迭代次数。其中α是人工设置的学习率。需要学习率的原因与梯度下降法相同，是为了保证能够忽略泰勒公式中的高阶无穷小项。

14.如果步长系数充分小，牛顿法在每次迭代时能保证函数值下降吗？

不能。

15.梯度下降法和牛顿法能保证找到函数的极小值点吗，为什么？

不能，可能收敛到鞍点，不是极值点。

16.解释一元函数极值判别法则。

假设x_0为函数的驻点，可分为以下三种情况。

情况一：在该点处的二阶导数大于 0，则为函数的极小值点；

情况二：在该点处的二阶导数小于 0，则为极大值点；

情况三：在该点处的二阶导数等于 0，则情况不定，可能是极值点，也可能不是极值点。

17.解释多元函数极值判别法则。

假设多元函数在点M的梯度为$\mathbf{0}$，即M是函数的驻点。其 Hessian 矩阵有如下几种情况。

情况一：Hessian 矩阵正定，函数在该点有极小值。

情况二：Hessian 矩阵负定，函数在该点有极大值。

情况三：Hessian 矩阵不定，则不是极值点，称为鞍点。

Hessian 矩阵正定类似于一元函数的二阶导数大于 0，负定则类似于一元函数的二阶导数小于 0。

18.什么是鞍点？

Hessian 矩阵不定的点称为鞍点，它不是函数的极值点。

19.解释什么是局部极小值，什么是全局极小值。

局部极值点。假设 $\mathbf{x}^*$ 是一个可行解，如果对可行域内所有点 $\mathbf{x}$ 都有 $f\left(\mathbf{x}^*\right) \le f\left(\mathbf{x}\right)$，则称 $\mathbf{x}^*$ 为全局极小值。

全局极值点。对于可行解 $\mathbf{x}^*$，如果存在其 δ 邻域，使得该邻域内的所有点即所有满足 $\left\|\mathbf{x}-\mathbf{x}^*\right\| \le \delta$ 的点 $\mathbf{x}$，都有 $f\left(\mathbf{x}^*\right) \le f\left(\mathbf{x}\right)$，则称 $\mathbf{x}^*$ 为局部极小值。

20.用拉格朗日乘数法求解如下极值问题

$$\min f\left(x_1, x_2\right) = x_1^2 + x_2^2$$
$$x_1 x_2 = 3$$

构造拉格朗日乘子函数

$$L\left(x_1, x_2, \lambda\right) = x_1^2 + x_2^2 + \lambda\left(x_1 x_2 - 3\right)$$

对所有变量求偏导数，并令其为 0

$$\frac{\partial L}{\partial x_1} = 2x_1 + \lambda x_2 = 0$$
$$\frac{\partial L}{\partial x_2} = 2x_2 + \lambda x_1 = 0$$
$$\frac{\partial L}{\partial \lambda} = x_1 x_2 - 3 = 0$$

可以解得

$$x_1 = x_2 = \sqrt{3}$$

21.什么是凸集？

对于 n 维空间中的点集 C，如果对该集合中的任意两点 $\mathbf{x}$ 和 $\mathbf{y}$，以及实数 $0 \le \theta \le 1$，都有

$$\theta\mathbf{x} + \left(1-\theta\right)\mathbf{y} \in C$$

则称该集合称为凸集。凸集的形状是凸的，没有凹进去的地方。

22.什么是凸函数，如何判断一个一元函数是不是凸函数，如何判断一个多元函数是不是凸函数？

对于函数 $f(\mathbf{x})$，对于其定义域内的任意两点 $\mathbf{x}$ 和 $\mathbf{y}$，以及任意的实数 $0 \le \theta \le 1$，都有

$$f\left(\theta\mathbf{x}+(1-\theta)\mathbf{y}\right) \le \theta f(\mathbf{x}) + (1-\theta) f(\mathbf{y})$$

则函数 $f(\mathbf{x})$ 为凸函数。

一元函数是凸函数的二阶判定规则为其二阶导数大于等于 0，即

$$f''(x) \ge 0$$

对于多元函数则根据 Hessian 矩阵判定。如果函数的 Hessian 矩阵半正定则函数是凸函数。

22.什么是凸优化？

如果一个最优化问题的可行域是凸集且目标函数是凸函数，则该问题为凸优化问题。

23.证明凸优化问题的局部最优解一定是全局最优解。

假设 $\mathbf{x}$ 是一个局部最优解但不是全局最优解，即存在一个可行解 $\mathbf{y}$

$$f(\mathbf{x}) > f(\mathbf{y})$$

根据局部最优解的定义，不存在满足 $\|\mathbf{x}-\mathbf{z}\|_2 \le \delta$ 并且 $f(\mathbf{z}) < f(\mathbf{x})$ 的点。选择一个点

$$\mathbf{z} = \theta\mathbf{y} + (1-\theta)\mathbf{x}$$

其中

$$\theta = \frac{\delta}{2\|\mathbf{x}-\mathbf{y}\|_2}$$

则有

$$\begin{aligned}\|\mathbf{x}-\mathbf{z}\|_2 &= \left\|\mathbf{x} - \left(\frac{\delta}{2\|\mathbf{x}-\mathbf{y}\|_2}\mathbf{y} + \left(1-\frac{\delta}{2\|\mathbf{x}-\mathbf{y}\|_2}\right)\mathbf{x}\right)\right\|_2 \\ &= \left\|\frac{\delta}{2\|\mathbf{x}-\mathbf{y}\|_2}(\mathbf{x}-\mathbf{y})\right\|_2 \\ &= \frac{\delta}{2} \le \delta\end{aligned}$$

即该点在 $\mathbf{x}$ 的 δ 邻域内。另外有

$$f(\mathbf{z}) = f\left(\theta\mathbf{y}+(1-\theta)\mathbf{x}\right) \le \theta f(\mathbf{y}) + (1-\theta) f(\mathbf{x}) < f(\mathbf{x})$$

这与 $\mathbf{x}$ 是局部最优解矛盾。如果一个局部最优解不是全局最优解，在它的任何邻域内还

可以找到函数值比该点更小的点，这与该点是局部最优解矛盾。

24.对于如下最优化问题：

$$\begin{aligned}&\min f\left(x_1,x_2\right)=x_1^2+x_2\\&4-2x_1-x_2\le 0\\&1-x_2\le 0\end{aligned}$$

构造广义拉格朗日乘子函数，将该问题转化为对偶问题。

首先构造拉格朗日乘子函数

$$L\left(\mathbf{x},\boldsymbol{\alpha}\right)=x_1^2+x_2+\alpha_1\left(4-2x_1-x_2\right)+\alpha_2\left(1-x_2\right)$$

对偶问题为先控制原始优化变量 $\mathbf{x}$，然拉格朗日乘子函数取极小值，然后控制拉格朗日乘子变量，让拉格朗日乘子函数取极大值。

首先对 x_1 求偏导数，并令其为 0，可以解得

$$\frac{\partial L}{\partial x_1}=2x_1-2\alpha_1=0\Rightarrow x_1=\alpha_1$$

然后对 x_2 求偏导数，并令其为 0，可以解得

$$\frac{\partial L}{\partial x_2}=1-\alpha_1-\alpha_2=0$$

将这些解代入拉格朗日乘子函数，可以得到

$$\begin{aligned}L\left(\mathbf{x},\boldsymbol{\alpha}\right)&=\alpha_1^2+x_2+\alpha_1\left(4-2\alpha_1-x_2\right)+\alpha_2\left(1-x_2\right)\\&=-\alpha_1^2+4\alpha_1+\alpha_2+x_2\left(1-\alpha_1-\alpha_2\right)\end{aligned}$$

由于 $1-\alpha_1-\alpha_2=0$，拉格朗日乘子函数可以简化为

$$L\left(\mathbf{x},\boldsymbol{\alpha}\right)=-\alpha_1^2+4\alpha_1+\alpha_2$$

由于拉格朗日乘子变量必须大于等于 0，因此对偶问题为

$$\begin{aligned}&\max\nolimits_{\boldsymbol{\alpha}}-\alpha_1^2+4\alpha_1+\alpha_2\\&\alpha_1\ge 0\\&\alpha_2\ge 0\\&1-\alpha_1-\alpha_2=0\end{aligned}$$

25.一维正态分布的概率密度函数为

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

给定一组样本$x_1,...,x_l$。用最大似然估计求解正态分布的均值和方差。

对于正态分布$N\left(\mu,\sigma^2\right)$，有样本集$x_1,...,x_n$。该样本集的似然函数为

$$\begin{aligned}L(\mu,\sigma)&=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\\&=\left(2\pi\sigma^2\right)^{-\frac{n}{2}}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)\end{aligned}$$

对数似然函数为

$$\ln L(\mu,\sigma)=-\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln\left(\sigma^2\right)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

对μ和σ求偏导数，得到似然方程组为

$$\begin{cases}\dfrac{\partial\ln L(\mu,\sigma)}{\partial\mu}=\dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^{n}(x_i-\mu)=0\\[2ex]\dfrac{\partial\ln L(\mu,\sigma)}{\partial\sigma^2}=-\dfrac{n}{2\sigma^2}+\dfrac{n}{2\sigma^4}\displaystyle\sum_{i=1}^{n}(x_i-\mu)^2=0\end{cases}$$

解得

$$\mu=\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$$

$$\sigma^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$$

26.如何判断一个矩阵是否为正定矩阵？

如果对于任意非**0**向量**x**都有

$$\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}\geq 0$$

则称**A**为半正定矩阵。如果将上面的不等式严格成立，称为正定矩阵。判定矩阵正定可以根据上面的定义。另外一种常用的方式是判定其所有特征值是否都为正，如果为正，则为正定矩阵。

27.解释最速下降法的原理。

梯度下降法中步长是固定的，最速下降法是对梯度下降法的改进，它动态确定步长值。最速下降法同样沿着梯度相反的方向进行迭代，但每次需要计算最佳步长α。

定义最速下降法的搜索方向为

$$\mathbf{d}_k = -\nabla f\left(\mathbf{x}_k\right)$$

在该方向上寻找使得函数值最小的步长，通过求解如下一元函数优化问题实现

$$\alpha_k = \arg\min_{\alpha} f\left(\mathbf{x}_k + \alpha \mathbf{d}_k\right)$$

优化变量是α。实现时有两种方案。第一种方案是将α的取值离散化，即取典型值$\alpha_1,...,\alpha_k$，分别计算取这些值的目标函数值然后确定最优值。或直接求解上面目标函数的驻点，对于有些情况可得到解析解。

28.解释坐标下降法的原理。

坐标下降法是一种分治法。对于多元函数的优化问题，坐标下降法每次只对一个变量进行优化，依次优化每一个变量，直至收敛。假设要求解的优化问题为

$$\min f\left(\mathbf{x}\right), \mathbf{x} = \left(x_1, x_2, ..., x_n\right)$$

算法在每次迭代时依次选择$x_1,...,x_n$进行优化，每次求解单个变量的优化问题。

29.一维正态分布的概率密度函数为

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

按照定义计算其数学期望与方差。

首先计算数学期望。使用换元法，令

$$z = \frac{x-\mu}{\sigma}$$

则有

$$x = \mu + \sigma z$$

根据数学期望的定义，有

$$
\begin{aligned}
\mathrm{E}[X] &= \int_{-\infty}^{+\infty} x \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \\
&= \int_{-\infty}^{+\infty} (\sigma z + \mu) \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{z^2}{2}} d(\sigma z + \mu) \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} (\sigma z + \mu) e^{-\frac{z^2}{2}} dz \\
&= \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} z e^{-\frac{z^2}{2}} dz + \frac{\mu}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{z^2}{2}} dz \\
&= 0 + \frac{\mu}{\sqrt{2\pi}} \sqrt{2\pi} \\
&= \mu
\end{aligned}
$$

上式第 5 步成立是因为 $ze^{-\frac{z^2}{2}}$ 是奇函数，它在 $(-\infty,+\infty)$ 内的积分为 0。第 6 步利用了下面的结论

$$
\int_{-\infty}^{+\infty} e^{-\frac{z^2}{2}} dz = \sqrt{2\pi}
$$

下面计算方差，同样令 $z = \frac{x-\mu}{\sigma}$，则有

$$
\begin{aligned}
\mathrm{var}[X] &= \int_{-\infty}^{+\infty} (x-\mu)^2 \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \\
&= \int_{-\infty}^{+\infty} \sigma^2 z^2 \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{z^2}{2}} d(\sigma z + \mu) \\
&= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} z^2 e^{-\frac{z^2}{2}} dz \\
&= -\frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} z de^{-\frac{z^2}{2}} \\
&= -\frac{\sigma^2}{\sqrt{2\pi}} \left(ze^{-\frac{z^2}{2}} \Big|_{-\infty}^{+\infty} - \int_{-\infty}^{+\infty} e^{-\frac{z^2}{2}} dz \right) \\
&= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{z^2}{2}} dz \\
&= \sigma^2
\end{aligned}
$$

上式第 5 步利用了分布积分法。第 6 步成立是因为

$$
\lim_{x \to +\infty} ze^{-\frac{z^2}{2}} = 0
$$

$$
\lim_{x \to -\infty} ze^{-\frac{z^2}{2}} = 0
$$

30.两个离散型概率分布的 KL 散度定义为：

$$D_{\mathrm{KL}}\left(P\|Q\right)=\sum_{x}P\left(x\right)\ln\frac{P\left(x\right)}{Q\left(x\right)}$$

（1）证明下面的不等式，当$x>0$时：

$$\ln x\le x-1$$

（2）利用该不等式证明 KL 散度非负，即

$$D_{\mathrm{KL}}\left(p\|q\right)\ge 0$$

首先证明（2），根据定义有

$$\begin{aligned}D_{\mathrm{KL}}\left(p\|q\right)&=-\sum_{x}p\left(x\right)\ln\frac{q\left(x\right)}{p\left(x\right)}\\&\ge-\sum_{x}p\left(x\right)\left(\frac{q\left(x\right)}{p\left(x\right)}-1\right)\\&=-\sum_{x}q\left(x\right)+\sum_{x}p\left(x\right)=0\end{aligned}$$

接下来证明（1）。利用导数可以证明某些不等式，其思路是证明在某一点处是函数的极大值或极小值点。下面证明当$x>0$时下面不等式成立

$$\ln x<x-1$$

构造函数

$$f\left(x\right)=x-1-\ln x$$

其导数为

$$f^{'}\left(x\right)=1-\frac{1}{x}$$

当$x<1$时有$f^{'}\left(x\right)<0$，函数单调减，在$x>1$时有$f^{'}\left(x\right)>0$，函数单调增。1 是该函数的极小值点，且$f\left(1\right)=0$，因此不等式成立。

31.对于离散型概率分布，证明当其为均匀分布时熵有最大值。

对于离散型随机变量，熵是如下多元函数

$$H\left(p\right)=-\sum_{i=1}^{n}x_i\ln x_i$$

其中x_i为随机变量取第i个值的概率。由于是概率分布，因此有如下约束

$$\sum_{i=1}^{n} x_i = 1$$
$$x_i \geq 0$$

对数函数的定义域非负，因此可以去掉上面的不等式约束。构造拉格朗日乘子函数

$$L(\mathbf{x}, \lambda) = -\sum_{i=1}^{n} x_i \ln x_i + \lambda \left(\sum_{i=1}^{n} x_i - 1 \right)$$

对 x_i 和乘子变量偏导数并令其为 0，可以得到下面的方程组

$$\frac{\partial L}{\partial x_i} = -\ln x_i - 1 + \lambda = 0$$
$$\sum_{i=1}^{n} x_i = 1$$

可以解得 $x_i = 1/n$ 。此时熵的值为

$$H(p) = -\sum_{i=1}^{n} \frac{1}{n} \ln \frac{1}{n}$$
$$= \ln n$$

进一步可以证明该值是极大值。熵函数的二阶偏导数为

$$\frac{\partial^2 H}{\partial x_i^2} = -1/x_i$$
$$\frac{\partial^2 H}{\partial x_i \partial x_j} = 0, j \neq i$$

它的 Hessian 矩阵是如下的对角阵

$$\begin{bmatrix} -1/x_1 & \ldots & 0 \\ \ldots & .. & \ldots \\ 0 & \ldots & -1/x_n \end{bmatrix}$$

由于 $x_i > 0$ ，Hessian 矩阵负定，因此熵函数是凹函数。故 $x_i = 1/n$ 时熵有极大值。

32.对于连续型概率分布，已知其数学期望为 μ ，方差为 σ^2 。用变分法证明当此分布为正态分布时熵有最大值。

给定数学期望与方差即有如下等式约束

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

为了保证 $p(x)$ 是一个概率密度函数，还有如下约束

$$\int_{-\infty}^{+\infty} p(x)dx = 1$$

熵对应的泛函为

$$F[p] = -\int_{-\infty}^{+\infty} p(x)\log p(x)dx$$

这是一个带等式约束的泛函极值问题。构造拉格朗日乘子泛函

$$F[p,\alpha,\beta,\gamma] = -\int_{-\infty}^{+\infty} p(x)\ln p(x)dx + \alpha\left(\int_{-\infty}^{+\infty} p(x)dx - 1\right) + \beta\left(\int_{-\infty}^{+\infty} xp(x)dx - \mu\right)$$
$$+ \gamma\left(\int_{-\infty}^{+\infty} (x-\mu)^2 p(x)dx - \sigma^2\right)$$

根据欧拉-拉格朗日方程，由于泛函的核没有 $p(x)$ 的导数项，对 $p(x)$ 有如下微分方程

$$\frac{\delta F}{\delta p} = -\left(1+\ln p(x)\right) + \alpha + \beta x + \gamma(x-\mu)^2 = 0 \qquad (1)$$

对乘子变量求偏导数可以得到

$$\begin{aligned}
&\frac{\delta F}{\delta \alpha} = \int_{-\infty}^{+\infty} p(x)dx - 1 = 0 \\
&\frac{\delta F}{\delta \beta} = \int_{-\infty}^{+\infty} xp(x)dx - \mu = 0 \qquad (2) \\
&\frac{\delta F}{\delta \gamma} = \int_{-\infty}^{+\infty} (x-\mu)^2 p(x)dx - \sigma^2 = 0
\end{aligned}$$

根据式 1 可以解得

$$p(x) = \exp\left(\gamma(x-\mu)^2 + \beta x + \alpha - 1\right)$$

将其代入式 2 可以解得

$$\begin{aligned}
&\alpha = 1 - \ln\left(2\pi\sigma^2\right)^{\frac{1}{2}} \\
&\beta = 0 \\
&\gamma = -\frac{1}{2\sigma^2}
\end{aligned}$$

最终解得

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

此即正态分布的概率密度函数。

33.对于两个离散型概率分布，证明当二者相等时交叉熵有极小值。

假设第一个概率分布已知，即对于概率分布 $p(x)$，随机变量 X 即取第 i 个值的概率为常数 a_i，假设对于概率分布 $q(x)$，随机变量 X 取第 i 值的概率为 x_i。此时交叉熵为如下形式的多元函数

$$H(\mathbf{x}) = -\sum_{i=1}^{n} a_i \ln x_i$$

概率分布有如下约束条件

$$\sum_{i=1}^{n} x_i = 1$$

构造拉格朗日乘子函数

$$L(\mathbf{x}, \lambda) = -\sum_{i=1}^{n} a_i \ln x_i + \lambda \left(\sum_{i=1}^{n} x_i - 1 \right)$$

对所有变量求偏导数，并令偏导数为 0，可以得到下面的方程组

$$\begin{aligned} -\frac{a_i}{x_i} + \lambda &= 0 \\ \sum_{i=1}^{n} x_i &= 1 \end{aligned} \quad （1）$$

由于 a_i 是一个概率分布，因此有

$$\sum_{i=1}^{n} a_i = 1 \quad （2）$$

联立方程 1 与式 2 可以解得

$$\lambda = 1$$
$$x_i = a_i$$

因此在两个概率分布相等的时候交叉熵有极值。接下来证明这个极值是极小值。交叉熵函数的 Hessian 矩阵为

$$\begin{bmatrix} a_1 / x_1^2 & \dots & 0 \\ \dots & \dots & \dots \\ 0 & \dots & a_n / x_n^2 \end{bmatrix}$$

该矩阵正定，因此交叉熵函数是凸函数，上面的极值是极小值。

34.为什么在实际的机器学习应用中经常假设样本数据服从正态分布？

主要原因是：

1. 由中心极限定理保证，多个独立同分布的随机变量之和服从正态分布。

2. 正态分布的各种积分，包括数学期望，方差，协方差，边缘分布，条件分布都易于计算，可以得到解析解。

3. 给定数学期望和方差，正态分布的熵最大。

35.什么是随机事件独立，什么是随机向量独立？

如果

$$p\left(B\middle|A\right)=p\left(B\right)$$

或

$$p\left(A\middle|B\right)=p\left(A\right)$$

则称随机事件 A 和 B 独立。随机事件独立意味着一个事件是否发生并不影响另外一个事件。如果随机事件 A 和 B 独立，条件概率的定义以及随机事件相互独立的定义有

$$p\left(A,B\right)=p\left(A\right)p\left(B\right)$$

将上面的定义进行推广，如果 n 个随机事件 $A_i,i=1,...,n$ 相互独立，则它们同时发生的概率等于它们各自发生的概率的乘积

$$p\left(A_1,...,A_n\right)=\prod_{i=1}^{n}p\left(A_i\right)$$

随机变量之间的独立性与随机事件类似。对于二维随机向量，如果满足

$$p\left(x,y\right)=p\left(x\right)p\left(y\right)$$

则称随机变量 x 和 y 相互独立，随机事件独立性的定义一致。推广到 n 维随机向量，如果满足

$$p\left(\mathbf{x}\right)=p\left(x_1\right)p\left(x_2\right)...p\left(x_n\right)$$

则称这些随机变量相互独立。

36.什么是弱对偶？什么是强对偶？

弱对偶定理：如果原问题和对偶问题都存在最优解，则对偶问题的最优值不大于原问题的最优值，即：

$$d^*=\max_{\boldsymbol{\lambda},\mathbf{v},\lambda_i\geq 0}\min_{\mathbf{x}}L\left(\mathbf{x},\boldsymbol{\lambda},\mathbf{v}\right)\leq\min_{\mathbf{x}}\max_{\boldsymbol{\lambda},\mathbf{v},\lambda_i\geq 0}L\left(\mathbf{x},\boldsymbol{\lambda},\mathbf{v}\right)=p^*$$

如果原问题对最优解的函数值与对偶问题的最优解的函数值相等，则为强对偶。

37.证明弱对偶定理。

对任意的 $\mathbf{x}$ 和 $\boldsymbol{\lambda}$、$\mathbf{v}$，根据定义，对于对偶问题有

$$\theta_D(\boldsymbol{\lambda},\mathbf{v})=\min_{\mathbf{x}} L(\mathbf{x},\boldsymbol{\lambda},\mathbf{v})\le L(\mathbf{x},\boldsymbol{\lambda},\mathbf{v})$$

对于原问题有

$$\theta_P(\mathbf{x})=\max_{\boldsymbol{\lambda},\mathbf{v},\lambda_i\ge 0} L(\mathbf{x},\boldsymbol{\lambda},\mathbf{v})\ge L(\mathbf{x},\boldsymbol{\lambda},\mathbf{v})$$

因此对任意的 $\mathbf{x}$ 和 $\boldsymbol{\lambda}$、$\mathbf{v}$ 有

$$\theta_P(\mathbf{x})\ge\theta_D(\boldsymbol{\lambda},\mathbf{v})$$

由于原问题和对偶问题的最优值存在，有

$$\min_{\mathbf{x}}\theta_P(\mathbf{x})\ge\max_{\boldsymbol{\lambda},\mathbf{v},\nu_i\ge 0}\theta_D(\boldsymbol{\lambda},\mathbf{v})$$

因此弱对偶成立。

38.简述 Slater 条件。

一个凸优化问题如果存在一个候选 $\mathbf{x}$ 使得所有不等式约束都是严格满足的，即对于所有的 i 都有 $g_i(\mathbf{x})<0$，不等式不取等号。则存在 $\mathbf{x}^*,\boldsymbol{\lambda}^*,\mathbf{v}^*$ 使得它们分别为原问题和对偶问题的最优解，并且：

$$p^*=d^*=L(\mathbf{x}^*,\boldsymbol{\lambda}^*,\mathbf{v}^*)$$

Slater 条件是强对偶成立的充分条件而不是必要条件。

39.简述 KKT 条件。

KKT（Karush-Kuhn-Tucker）条件用于求解带有等式和不等式约束的优化问题，是拉格朗日乘数法的推广，是取得极值的一阶必要条件。对于如下带有等式和不等式约束的优化问题

$$\begin{aligned}&\min f(\mathbf{x})\\&g_i(\mathbf{x})\le 0\quad \mathrm{i}=1,...,q\\&h_i(\mathbf{x})=0\qquad \mathrm{i}=1,...,p\end{aligned}$$

和拉格朗日对偶的做法类似，为其构造乘子函数

$$L(\mathbf{x},\boldsymbol{\lambda},\boldsymbol{\mu})=f(\mathbf{x})+\sum_{j=1}^{p}\lambda_j h_j(\mathbf{x})+\sum_{k=1}^{q}\mu_k g_k(\mathbf{x})$$

$\boldsymbol{\lambda}$ 和 $\boldsymbol{\mu}$ 称为 KKT 乘子。最优解 $\mathbf{x}^*$ 满足如下条件

$$\nabla_{\mathbf{x}} L\left(\mathbf{x}^*\right)=\mathbf{0}$$
$$\mu_k \geq 0$$
$$\mu_k g_k\left(\mathbf{x}^*\right)=0$$
$$h_j\left(\mathbf{x}^*\right)=0$$
$$g_k\left(\mathbf{x}^*\right) \leq 0$$

等式约束 $h_j\left(\mathbf{x}^*\right)=0$ 和不等式约束 $g_k\left(\mathbf{x}^*\right) \leq 0$ 是本身应该满足的约束，$\nabla_{\mathbf{x}} L\left(\mathbf{x}^*\right)=\mathbf{0}$ 和之前的拉格朗日乘数法一样。只多了关于 $g_i(\mathbf{x})$ 的方程

$$\mu_k g_k\left(\mathbf{x}^*\right)=0$$

这可以分两种情况讨论。如果

$$g_k\left(\mathbf{x}^*\right)<0$$

要满足 $\mu_k g_k\left(\mathbf{x}^*\right)=0$ 的条件，则有 $\mu_k=0$。此时极值在不等式约束围成的区域内部取得。如果

$$g_k\left(\mathbf{x}^*\right)=0$$

则 μ_k 的取值自由，只要满足大于等于 0 即可，此时极值在不等式围成的区域的边界点处取得。KKT 条件只是取得极值的必要条件而非充分条件。如果一个最优化问题是凸优化问题，则 KKT 条件是取得极小值的充分条件。

40.解释蒙特卡洛算法的原理。为什么蒙特卡洛算法能够收敛？

蒙特卡洛算法是一种随机算法，借助于随机数计算某些难以计算的数学期望和定积分。假设随机向量服从概率分布 $p(\mathbf{x})$。则其数学期望定义为

$$\mathrm{E}\left(f(\mathbf{x})\right)=\int_{\mathbb{R}^n} f(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

如果用蒙特卡洛算法计算，非常简单。首先从概率分布 $p(\mathbf{x})$ 抽取 N 个样本 $\mathbf{x}_1, \ldots, \mathbf{x}_N$。然后计算

$$\mathrm{E}\left(f(\mathbf{x})\right) \approx \frac{1}{N} \sum_{i=1}^{N} f\left(\mathbf{x}_i\right)$$

就是数学期望的估计值。在这里 $\mathbf{x}_i \sim p(\mathbf{x})$。随机抽取的样本频率蕴含了随机变量的概率值 $p(\mathbf{x})$。根据大数定律，如果样本 $\mathbf{x}_i$ 独立同分布，则它们的平均值收敛到数学期望 μ。

即下面的极限成立

$$\lim_{N\to+\infty}\frac{1}{N}\sum_{i=1}^{N}f\left(x_i\right)=\mathrm{E}\left(f\left(x\right)\right)$$

从而保证蒙特卡洛算法收敛。

41.解释熵的概念。

熵（entropy）是信息论中最基本的一个概念，定义于一个随机变量之上，用于对概率分布的随机性程度进行度量，反映了一组数据所包含的信息量的大小。

对于离散型随机变量，熵定义为

$$H\left(p\right)=\mathrm{E}_x\left[-\ln p\left(x\right)\right]=-\sum_{i=1}^{n}p_i\ln p_i$$

对于连续型随机变量，熵（也称为微分熵，differential entropy）定义为

$$H\left(p\right)=-\int_{-\infty}^{+\infty}p\left(x\right)\ln p\left(x\right)dx \quad （6.3）$$

将求和换成了定积分。

第 3 章 基本概念

1.名词解释：有监督学习，无监督学习。

有监督学习算法有训练过程，算法用训练集进行学习，用学习得到的模型进行预测。通常所见的机器学习应用，如图像识别、语音识别等都属于有监督学习问题。有监督学习的样本由输入值与标签值组成

$$\left(\mathbf{x},y\right)$$

其中 $\mathbf{x}$ 为样本的特征向量，是机器学习模型的输入值；y 为标签值，是模型的输出值。对于训练集，样本的标签值是由人工事先标注好的，例如为每张手写数字图像关联一个其对应的数字。

有监督学习的目标是给定训练样本集，根据它确定假设空间中的映射函数（也称为假设）

$$y=h\left(\mathbf{x}\right)$$

无监督学习对无标签的样本进行分析，发现样本集的结构或者分布规律。其典型代表是聚类，表示学习，以及数据降维。

聚类也是分类问题，但无训练过程。算法把一批没有标签的样本划分成多个类，使得在某种相似度指标下每一类中的样本之间尽量相似，不同类的样本之间尽量不同。且类别事先未定义。

无监督学习的另一类典型算法是表示学习，它从样本中自动学习出有用的特征，用于分类或聚类任务。

数据降维算法将 n 维空间中的向量 $\mathbf{x}$ 通过函数映射到更低维的 m 维空间中，在这里

$m \ll n$

$$\mathbf{y} = h(\mathbf{x})$$

2.什么是分类问题，什么是回归问题？

对于有监督学习，如果样本标签是整数则称为分类问题。此时的目标是确定样本的类别，以整数编号。预测函数是向量到整数的映射

$$\mathbb{R}^n \to \mathbb{Z}$$

此时的机器学习模型称为分类器。分类问题的样本标签通常从 0 或 1 开始，以整数编号。

如果标签值是连续实数则称为回归问题。此时预测函数是向量到实数的映射

$$\mathbb{R}^n \to \mathbb{R}$$

例如根据一个人的学历、工作年限等特征预测其收入，是典型的回归问题，收入是实数值而不是类别标签。

2.列举常见的有监督学习算法。

贝叶斯分类器，决策树，支持向量机，boosting 算法

3.列举常见的无监督学习算法。

k 均值算法，PCA，t-SNE，层次聚类

4.简述强化学习的原理。

强化学习模拟人的行为，源自于行为主义心理学。类似于有监督学习，通过对状态-动作-回报值序列进行学习，可以得到一个称为策略函数的模型，用于确定在每种状态下要执行的动作。训练时，对正确的动作做出奖励，对错误的动作进行惩罚，训练完成之后用得到的模型进行预测。

5.什么是生成模型？什么是判别模型？

假设 $\mathbf{x}$ 为特征向量，y 为该样本的标签值。如果机器学习算法对样本特征向量和标签的联合概率分布 $p(\mathbf{x}, y)$ 建模，则称为生成模型。如果对条件概率 $p(y|\mathbf{x})$ 进行建模，则称为判别模型。不使用概率模型的分类算法也属于判别模型，它直接预测样本的标签值而不关心样本的概率分布，这种情况的预测函数为

$$y = f(\mathbf{x})$$

这三种模型也分别被称为生成学习，条件学习，以及判别学习。

还有另外一种定义标准。生成模型对条件概率 $p(\mathbf{x}|y)$ 建模，判别模型对条件概率 $p(y|\mathbf{x})$ 建模。前者不仅可以通过贝叶斯公式用于分类问题，还可用于根据标签值 y（也称为隐变量）生成随机的样本数据 $\mathbf{x}$，而后者则只能用于根据样本特征向量 $\mathbf{x}$ 的值判断它的标签值 y 的分类任务。

6.概率模型一定是生成模型吗？

不一定，logistic 回归和 softmax 回归都是概率模型，但不是生成模型。

7.不定项选择。下面那些算法是生成模型？_____BGI_____哪些算法是判别模型？_____ACDEFH_____

A.决策树　　B.贝叶斯分类器　　C.全连接神经网络　D.支持向量机　　E. logistic 回归

F. AdaBoost 算法　　G.隐马尔可夫模型　　H.条件随机场　　I.受限玻尔兹曼机

8.如何判断是否发生过拟合？

模型在训练集上精度高，但在测试集上精度低。

9.发生过拟合的原因有哪些，应该怎么解决？

引起过拟合的可能原因有：

1. 模型本身过于复杂，拟合了训练样本集中的噪声。此时需要选用更简单的模型，或者对模型进行裁剪。
2. 训练样本太少或者缺乏代表性。此时需要增加样本数，或者增加样本的多样性。
3. 训练样本噪声的干扰，导致模型拟合了这些噪声，这时需要剔除噪声数据或者改用对噪声不敏感的模型。

10.列举常见的正则化方法。

L1 正则化

L2 正则化

决策树的剪枝算法

神经网络训练中的 dropout 技术，提前终止技术

11.解释 ROC 曲线的原理。

对于二分类问题可以通过调整分类器的灵敏度得到不同的分类结果，从而在二者之间折中。将各种灵敏度下的性能指标连成曲线可以得到 ROC 曲线，它能够更全面的反映算法的

性能。

真阳率（TPR）即召回率，是正样本被分类器判定为正样本的比例

$$TPR = TP / \left(TP + FN \right)$$

在目标检测任务中正样本是要检测的目标，真阳率即检测率，即目标能够被检测出来的比例。假阳率（FPR）是负样本被分类器判定为正样本的比例

$$FPR = FP / \left(FP + TN \right)$$

对于目标检测问题假阳率即误报率。

ROC 曲线的横轴为假阳率，纵轴为真阳率。当假阳率增加时真阳率会增加，它是一条增长的曲线。

12.解释精度，召回率，F1 值的定义。

测试样本中正样本被分类器判定为正样本的数量记为 TP，被判定为负样本的数量记为 FN；负样本中被分类器判定为负样本的数量记为 TN，被判定为正样本的数量记为 FP。

精度定义为被判定为正样本的测试样本中，真正的正样本所占的比例

$$P = \frac{TP}{TP+FP}$$

显然 $P \le 1$。召回率定义为被判定为正样本的正样本占所有正样本的比例

$$R = \frac{TP}{TP + FN}$$

同样的 $R \le 1$。精度值越接近 1，对正样本的分类越准确，即查的越准。召回率越接近于，正样本被正确分类的比例越大，即查的越全。

根据精度和召回率，F1 值定义为

$$F1 = \frac{2 \times P \times R}{P + R}$$

它是精度与召回率的调和平均值的倒数

$$\frac{1}{F1} = \frac{1}{2}\left(\frac{1}{P} + \frac{1}{R} \right)$$

F1 值综合考虑了精度与召回率，其值越大则模越准确，当精度与召回率均为 1 时 F1 有最大值 1。

13.解释交叉验证的原理。

交叉验证用于统计模型的精度值。k 折交叉验证将样本随机、均匀地分为 k 份，轮流用其中的 $k-1$ 份训练模型，1 份用于测试模型的准确率，用 k 个准确率的均值作为最终的准确率。

14.什么是过拟合，什么是欠拟合？

欠拟合也称为欠学习，指模型在训练集上的精度差。导致欠拟合的常见原因有模型简单，特征数太少无法正确的建立映射关系。

过拟合也称为过学习，指模型在训练集上表精度高，但在测试集上精度低，泛化性能差。过拟合产生的根本原因是训练数据包含抽样误差，算法训练时模型拟合了抽样误差。所谓抽样误差，是指抽样得到的样本集和整体数据集之间的偏差。

15.什么是没有免费午餐定理？

没有任何一个机器学习模型在所有样本集上表现是最优的。如果在一个数据集上算法 A 优于算法 B，则一定存在另外一个数据集，使得 B 优于 A。

16.简述奥卡姆剃刀原理。

简单的模型通常具有更好的泛化性能。

17.推导偏差-方差分解公式。

标签值由目标函数和随机噪声决定

$$y = f(\mathbf{x}) + \varepsilon$$

其中 ε 为随机噪声，其均值为 0，方差为 σ^2。模型的总误差为

$$\begin{aligned}
\mathrm{E}\left(\left(y-\hat{f}\right)^2\right) &= \mathrm{E}\left(y^2+\hat{f}^2-2y\hat{f}\right) \\
&= \mathrm{E}\left(y^2\right)+\mathrm{E}\left(\hat{f}^2\right)-\mathrm{E}\left(2y\hat{f}\right) \\
&= \mathrm{Var}\left(y\right)+\mathrm{E}^2\left(y\right)+\mathrm{Var}\left(\hat{f}\right)+\mathrm{E}^2\left(\hat{f}\right)-2f\mathrm{E}\left(\hat{f}\right) \\
&= \mathrm{Var}\left(y\right)+\mathrm{Var}\left(\hat{f}\right)+\left(f^2-2f\mathrm{E}\left(\hat{f}\right)+\mathrm{E}^2\left(\hat{f}\right)\right) \\
&= \mathrm{Var}\left(y\right)+\mathrm{Var}\left(\hat{f}\right)+\mathrm{E}\left(f-\mathrm{E}\left(\hat{f}\right)\right)^2 \\
&= \sigma^2+\mathrm{Var}\left(\hat{f}\right)+\mathrm{Bias}^2\left(\hat{f}\right)
\end{aligned}$$

18.证明如果采用均方误差函数，线性回归的优化问题是凸优化问题。

损失函数使用均方误差函数，定义为

$$L = \frac{1}{2l}\sum_{i=1}^{l}\left(h\left(\mathbf{x}_i\right)-y_i\right)^2$$

将回归函数代入损失函数，可以

$$L=\frac{1}{2l}\sum_{i=1}^{l}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i+b-y_i\right)^2$$

将权重向量和特征向量进行增广，即将 $\mathbf{w}$ 和 b 进行合并以简化表达，特征向量做相应的扩充，扩充后的向量为

$$\left[\mathbf{w},b\right]\rightarrow\mathbf{w}$$
$$\left[\mathbf{x},1\right]\rightarrow\mathbf{x}$$

目标函数简化为

$$L=\frac{1}{2l}\sum_{i=1}^{l}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i-y_i\right)^2$$

其二阶偏导数为

$$\frac{\partial^2 L}{\partial w_i\partial w_j}=\frac{1}{l}\sum_{k=1}^{l}x_{ki}x_{kj}$$

其中 x_{ki} 为第 k 个样本的特征向量的第 i 个分量。目标函数的 Hessian 矩阵为

$$\frac{1}{l}\sum_{k=1}^{l}\begin{bmatrix}x_{k1}x_{k1} & \cdots & x_{k1}x_{kn}\\ \cdots & \cdots & \cdots\\ x_{kn}x_{k1} & \cdots & x_{kn}x_{kn}\end{bmatrix}=\frac{1}{l}\begin{bmatrix}\sum_{k=1}^{l}x_{k1}x_{k1} & \cdots & \sum_{k=1}^{l}x_{k1}x_{kn}\\ \cdots & \cdots & \cdots\\ \sum_{k=1}^{l}x_{kn}x_{k1} & \cdots & \sum_{k=1}^{l}x_{kn}x_{kn}\end{bmatrix}$$

简写成矩阵形式为

$$\frac{1}{l}\begin{bmatrix}\mathbf{x}_1 & \cdots & \mathbf{x}_l\end{bmatrix}\begin{bmatrix}\mathbf{x}_1^{\mathrm{T}}\\ \cdots\\ \mathbf{x}_l^{\mathrm{T}}\end{bmatrix}=\frac{1}{l}\mathbf{X}^{\mathrm{T}}\mathbf{X}$$

其中 $\mathbf{X}$ 是所有样本的特征向量按照行构成的矩阵。对于任意非 $\mathbf{0}$ 向量 $\mathbf{x}$ 有

$$\mathbf{x}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{x}=\left(\mathbf{X}\mathbf{x}\right)^{\mathrm{T}}\left(\mathbf{X}\mathbf{x}\right)\geq 0$$

因此 Hessian 矩阵是半正定矩阵，目标函数是凸函数

19.推导线性回归的梯度下降迭代公式。

如果采用梯度下降法求解，损失函数对 w_j 的偏导数为

$$\frac{\partial L}{\partial w_j}=\frac{1}{l}\sum_{i=1}^{l}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i-y_i\right)x_{ij}$$

20.解释混淆矩阵的概念。

对于 k 分类问题，混淆矩阵为 $k \times k$ 的矩阵，它的元素 c_{ij} 表示第 i 类样本被分类器判定为第 j 类的数量

$$\begin{bmatrix} c_{11} & \cdots & c_{1k} \\ \cdots & \cdots & \cdots \\ c_{k1} & \cdots & c_{kk} \end{bmatrix}$$

如果所有样本都被正确分类，则该矩阵为对角阵。主对角线的元素之和 $\sum_{i=1}^{k} c_{ii}$ 为正确分类的样本数，其他元素之和为错误分类的样本数。因此对角线的值越大，分类器准确率越高。

21.解释岭回归的原理。

岭回归是带 L2 正则化项的线性回归，训练时优化的目标函数为

$$\min_{\mathbf{w}} \sum_{i=1}^{l} \left(\mathbf{w}^{\mathrm{T}} \mathbf{x}_i - y_i \right)^2 + \lambda \frac{1}{2} \mathbf{w}^{\mathrm{T}} \mathbf{w}$$

22.解释 LASSO 回归的原理。

LASSO 回归是使用 L1 正则化项的线性回归，训练时的目标函数为

$$\min_{\mathbf{w}} \sum_{i=1}^{l} \left(\mathbf{w}^{\mathrm{T}} \mathbf{x}_i - y_i \right)^2 + \lambda \left\| \mathbf{w} \right\|_1$$

第 4 章 贝叶斯分类器

1.什么是先验概率，什么是后验概率？

贝叶斯公式

$$p\left(A \middle| B\right) = \frac{p\left(A\right) p\left(B \middle| A\right)}{p\left(B\right)}$$

描述了先验概率和后验概率之间的关系。如果事件 A 是因，事件 B 是果，则称 $p\left(A\right)$ 为先验概率，意为事先已经知道其值。$p\left(A \middle| B\right)$ 称为后验概率，意为事后才知道其值。条件概率 $p\left(B \middle| A\right)$ 则称为似然函数。先验概率是根据以往经验和分析得到的概率，在随机事件发生之前即已经知道，是“原因”发生的概率。后验概率是根据“结果”信息所计算出的导致该

结果的原因所出现的概率。后验概率用于在事情已经发生的条件下，分析使得这件事情发生的原因概率。

2.推导朴素贝叶斯分类器的预测函数。

朴素贝叶斯分类器假设特征向量的分量之间相互独立。给定样本的特征向量 $\mathbf{x}$，该样本属于某一类 c_i 的概率为

$$p\left(y=c_i\left|\mathbf{x}\right.\right)=\frac{p\left(y=c_i\right)p\left(\mathbf{x}\left|y=c_i\right.\right)}{p\left(\mathbf{x}\right)}$$

由于假设特征向量各个分量相互独立，因此有

$$p\left(y=c_i\left|\mathbf{x}\right.\right)=\frac{p\left(y=c_i\right)\prod_{j=1}^{n}p\left(x_j\left|y=c_i\right.\right)}{Z}$$

其中 Z 为归一化因子。上式的分子可以分解为类概率 $p\left(c_i\right)$ 和该类每个特征分量的条件概率 $p\left(x_j\left|y=c_i\right.\right)$ 的乘积。

3.什么是拉普拉斯光滑？

在类条件概率的计算公式中，如果 $N_{x_i=v,y=c}$ 为 0，即特征分量的某个取值在某一类在训练样本中一次都不出现，则会导致如果预测样本的特征分量取到这个值时整个分类判别函数的值为 0。作为补救措施可以使用拉普拉斯平滑，具体做法是给分子和分母同时加上一个正数。如果特征分量的取值有 k 种情况，将分母加上 k，每一类的分子加上 1，这样可以保证所有类的条件概率加起来还是 1

$$p\left(x_i=v\left|y=c\right.\right)=\frac{N_{x_i=v,y=c}+1}{N_{y=c}+k}$$

对于每一个类，计算出待预测样本的各个特征分量的类条件概率，然后与类概率一起连乘，得到上面的预测值，该预测值最大的类为最后的分类结果。

4.推导正态贝叶斯分类器的预测函数。

假设特征向量服从 n 维正态分布，其中 $\boldsymbol{\mu}$ 为均值向量，$\boldsymbol{\Sigma}$ 为协方差矩阵。类条件概率密度函数为

$$p(\mathbf{x}|c)=\frac{1}{(2\pi)^{\frac{n}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

其中$|\mathbf{\Sigma}|$是协方差矩阵的行列式，$\mathbf{\Sigma}^{-1}$是协方差矩阵的逆矩阵。

在预测时需要寻找具有最大条件概率的那个类，即最大化后验概率，根据贝叶斯分类器的预测公式

$$\arg\max_c\left(p(c|\mathbf{x})\right)=\arg\max_c p(c)p(\mathbf{x}|c)$$

假设每个类的概率$p(c)$相等，则等价于求解该问题

$$\arg\max_c\left(p(\mathbf{x}|c)\right)$$

也就是计算每个类的$p(\mathbf{x}|c)$值然后取最大的那个。对$p(\mathbf{x}|c)$取对数，有

$$\ln\left(p(\mathbf{x}|c)\right)=\ln\left(\frac{1}{(2\pi)^{\frac{n}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}}\right)-\frac{1}{2}\left((\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

进一步简化为

$$\ln\left(p(\mathbf{x}|c)\right)=-\frac{n}{2}\ln(2\pi)-\frac{1}{2}\ln\left(|\mathbf{\Sigma}|\right)-\frac{1}{2}\left((\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

其中$-\frac{n}{2}\ln(2\pi)$是常数，对所有类都是相同。求上式的最大指等价于求下式的最小值

$$\ln\left(|\mathbf{\Sigma}|\right)+\left((\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

该值最小的那个类为最后的分类结果。其中$\ln\left(|\mathbf{\Sigma}|\right)$可以根据每一类的训练样本预先计算好，和$\mathbf{x}$无关，不用重复计算。预测时只需要根据样本$\mathbf{x}$计算$(\mathbf{x}-\boldsymbol{\mu})\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}$的值，而$\mathbf{\Sigma}^{-1}$也是在训练时计算好的，不用重复计算。

5.贝叶斯分类器是生成模型还是判别模型？

生成模型，它对$p(\mathbf{x}|y)$建模。

第 5 章 决策树

1.什么是预剪枝，什么是后剪枝？

决策树的剪枝算法可以分为两类，分别称为预剪枝和后剪枝。前者在树的训练过程中通过停止分裂对树的规模进行限制；后者先构造出一棵完整的树，然后通过某种规则消除掉部

分节点，用叶子节点替代。

2.什么是属性缺失问题？

在某些情况下样本特征向量中一些分量没有值，这称为属性缺失。

3.对于属性缺失问题，在训练时如何生成替代分裂规则？

替代分裂的目标是对训练样本的分裂结果要和主分裂尽可能接近，即被主分裂分到左边的样本要尽量被替代分裂分到左边；被主分裂分到右边的样本要尽量被替代分裂分到右边。主分裂和替代分裂对所有训练样本的分裂结果有 4 种情况，分别是

$$\mathrm{LL}, \mathrm{LR}, \mathrm{RL}, \mathrm{RR}$$

LL 表示被主分裂、替代分裂都分到了左子树的样本数。LR 表示被主分裂分到了左子树，被替代分裂分到了右子树的样本数。RL 表示被主分裂分到了右子树，被替代分裂分到了左子树的样本数。RR 表示被主分裂和替代分裂都分到了右子树的样本数。

因此 LL+RR 是主分裂和替代分裂的结果一致的样本数，LR+RL 是主分裂和替代分裂的结果不一致的样本数。由于可以将左右子树反过来，因此给定一个特征分量，在寻找替代分裂的分裂阈值时要让 LL+RR 或者 LR+RL 最大化，最后取它们的最大值

$$\max\left(\mathrm{LL}+\mathrm{RR}, \mathrm{LR}+\mathrm{RL}\right)$$

该值对应的分裂阈值为替代分裂的分裂阈值。对于除最佳分裂所用特征之外的其他所有特征，都找出该特征的最佳分裂和上面的值。最后取该值最大的那个特征和分裂阈值作为替代分裂规则。

4.列举分类问题的分裂评价指标。

熵不纯度。样本集 D 的熵不纯度定义为

$$E(D)=-\sum_i p_i \log_2 p_i$$

熵用来度量一组数据包含的信息量大小。当样本只属于某一类时熵最小，当样本均匀的分布于所有类中时熵最大。因此，如果能找到一个分裂让熵最小，这就是我们想要的最佳分裂。

Gini 不纯度。样本集的 Gini 不纯度定义为

$$G(D)=1-\sum_i p_i^2$$

当样本属于某一类时 Gini 不纯度的值最小，此时最小值为 0；当样本均匀的分布与每一类时 Gini 不纯度的值最大。

误分类不纯度。样本集的误分类不纯度定义为

$$E(D)=1-\max\left(p_i\right)$$

之所以这样定义是因为我们会把样本判定为频率最大的那一类，因此其他样本都会被错分，故错误分类率为上面的值。和上面的两个指标一样，当样本只属于某一类时误分类不纯

度有最小值 0，样本均匀的属于每一类时该值最大。

5.证明当各个类出现的概率相等时，Gini 不纯度有极大值；当样本全部属于某一类时，Gini 不纯度有极小值。

样本集的 Gini 不纯度定义为

$$G(D)=1-\sum_i p_i^2$$

概率分布必须满足下面的约束条件

$$\sum_i p_i=1$$

$$p_i\geq 0$$

构造拉格朗日乘子函数

$$L=1-\sum_{i=1}^{n}p_i^2+\lambda\left(\sum_{i=1}^{n}p_i-1\right)$$

对所有变量求偏导数，并令其为 0，得到下面的方程组

$$\frac{\partial L}{\partial p_i}=-2p_i+\lambda=0$$

$$\sum_{i=1}^{n}p_i-1=0$$

解得

$$p_i=1/n$$

6.ID3 用什么指标作为分裂的评价指标？

信息增益。假设用某种分裂规则将样本集 D 划分为 m 个不相交的子集 $D_1,...,D_m$，则该划分的信息增益定义为

$$G=H(D)-\sum_{i=1}^{m}\frac{|D_i|}{|D|}H(D_i)$$

其意义是划分之后熵的下降值。根据这种定义，ID3 训练时的优化目标为

$$L(T)=\sum_{t=1}^{|T|}N_tH_t(T)+\alpha|T|$$

其中 T 代表决策树，$|T|$ 为叶子节点数，N_t 为第 t 个叶子节点的训练样本数，$H_t(T)$ 为第 t 个叶子节点的训练样本集的熵。上式右端第一项是每个叶子节点的熵与其训练样本数的乘积，用于保证每个叶子节点的分类准确率尽可能高。第二项为正则化项，用于控制决策树的复杂度。

7.C4.5 用什么指标作为分裂的评价指标？

增益率。C4.5 是对 ID3 的改进，使用信息增益率来选择属性，克服了用信息增益选择属性时偏向选择取值多的属性的不足。另外还能够完成对连续属性的离散化处理。增益率定义为

$$GR = \frac{G}{IV}$$

其中 G 为信息增益，IV 为归一化因子，定义为

$$IV = -\sum_{i=1}^{m} \frac{|D_i|}{|D|} \ln \frac{|D_i|}{|D|}$$

其中 m 为类型数。

8.解释决策树训练时寻找最佳分裂的原理。

寻找最佳分裂时需要计算用每个阈值对样本集进行分裂后的指标值，寻找该值最大时对应的分裂，它就是最佳分裂。如果是数值型特征，对于每个特征将 l 个训练样本按照该特征的值从小到大排序，假设排序后的值为

$$x_1, ..., x_l$$

接下来从 x_1 开始，依次用每个 x_i 作为阈值，将样本分成左右两部分，计算分裂的指标值，该值最大的那个分裂阈值就是此特征的最佳分裂阈值。在计算出每个特征的最佳分裂阈值和上面的纯度值后，比较所有这些分裂的纯度值大小，该值最大的分裂为所有特征的最佳分裂。

9.对于分类问题，叶子节点的值如何设定？对于回归问题，决策树叶子节点的值如何设定？

对于分类树，将叶子节点的值设置成本节点的训练样本集中出现概率最大的那个类。即

$$y^* = \arg\max_y p_y$$

对于回归树，则设置为本节点训练样本标签值的均值

$$y^* = \frac{\sum_{i=1}^{N} y_i}{N}$$

10.决策树如何计算特征的重要性？

对每个特征分量在整个决策树中的分裂质量累加求和。对于分类树，分裂质量是 Gini

纯度值；对于回归树，分裂质量是每次分裂时回归误差的下降值。假设第i个特征分量的分裂质量之和为q_i，然后对所有特征分量求和后分裂质量进行归一化

$$\frac{q_i}{\sum_{i=1}^{n} q_i}$$

这个值即为该特征分量的重要性。其中n为特征向量的维数。显然该值越大变量越重要。统计所有节点的分裂质量需要对树进行遍历，可以使用任何一种遍历算法。这样做的依据是，如果一个变量被选来做分裂则说明它对分类或者回归很重要，如果它做分裂时的分裂质量很大，说明其对分类或者回归的贡献很大。

11.CART 对分类问题和回归问题分别使用什么作为分裂评价指标？

对于分类问题，训练时以基尼系数作为分裂的评价指标；对于回归问题，则用回归误差的下降值作为评价指标。样本集的 Gini 不纯度定义为

$$G(D) = 1 - \sum_i p_i^2$$

将类概率的计算公式代入 Gini 不纯度的定义，可以得到简化的计算公式

$$G(D) = 1 - \sum_i p_i^2 = 1 - \sum_i \left(N_i / N\right)^2 = 1 - \left(\sum_i N_i^2\right) / N^2$$

上面定义的是样本集的不纯度，我们需要评价的是分裂的好坏，因此需要根据样本集的不纯度构造出分裂的不纯度。分裂规则将节点的训练样本集分裂成左右两个子集，分裂的目标是把数据分成两部分之后这两个子集都尽可能的纯，因此我们计算左右子集的不纯度之和作为分裂的不纯度，显然求和需要加上权重，以反映左右两边的训练样本数。由此得到分裂的不纯度计算公式为

$$G = \frac{N_L}{N} G(D_L) + \frac{N_R}{N} G(D_R)$$

其中$G(D_L)$是左子集的不纯度，$G(D_R)$是右子集的不纯度，N是总样本数，N_L是左子集的样本数，N_R是右子集的样本数。

如果采用 Gini 不纯度指标，将 Gini 不纯度的计算公式代入上式可以得到

$$
\begin{aligned}
G &= \frac{N_L}{N}\left(1-\frac{\sum_i N_{L,i}^2}{N_L^2}\right)+\frac{N_R}{N}\left(1-\frac{\sum_i N_{R,i}^2}{N_R^2}\right) \\
&= \frac{1}{N}\left(N_L-\frac{\sum_i N_{L,i}^2}{N_L}+N_R-\frac{\sum_i N_{R,i}^2}{N_R}\right) \\
&= 1-\frac{1}{N}\left(\frac{\sum_i N_{L,i}^2}{N_L}+\frac{\sum_i N_{R,i}^2}{N_R}\right)
\end{aligned}
$$

其中 $N_{L,i}$ 是左子节点中第 i 类样本数，$N_{R,i}$ 是右子节点中第 i 类样本数。由于 N 是常数，要让 Gini 不纯度最小化等价于让下面的值最大化

$$
G = \frac{\sum_i N_{L,i}^2}{N_L}+\frac{\sum_i N_{R,i}^2}{N_R}
$$

这个值可以看做是 Gini 纯度，它的值越大，样本越纯。

对于回归树，衡量分裂的标准是回归误差即样本方差，每次分裂时选用使得方差最小化的那个分裂。假设节点的训练样本集有 l 个样本 $\left(\mathbf{x}_i, y_i\right)$，其中 $\mathbf{x}_i$ 为特征向量，y_i 为实数的标签值。节点的回归值为所有样本的均值，回归误差为所有样本的标签值与回归值的均方和误差，定义为

$$
E(D) = \frac{1}{l}\sum_{i=1}^{l}\left(y_i-\overline{y}\right)^2
$$

可以证明，回归值为均值的时候，上面的均方误差最小。把均值的定义带入上式，得到

$$
\begin{aligned}
E(D) &= \frac{1}{l}\sum_{i=1}^{l}\left(y_i-\frac{1}{l}\sum_{j=1}^{l}y_j\right)^2 \\
&= \frac{1}{l}\sum_{i=1}^{l}\left(y_i^2-2y_i\frac{1}{l}\sum_{j=1}^{l}y_j+\frac{1}{l^2}\left(\sum_{j=1}^{l}y_j\right)^2\right) \\
&= \frac{1}{l}\left(\sum_{i=1}^{l}y_i^2-\frac{2}{l}\left(\sum_{i=1}^{l}y_i\right)^2+\frac{1}{l}\left(\sum_{j=1}^{l}y_j\right)^2\right) \\
&= \frac{1}{l}\left(\sum_{i=1}^{l}y_i^2-\frac{1}{l}\left(\sum_{i=1}^{l}y_i\right)^2\right)
\end{aligned}
$$

根据样本集的回归误差，我们同样可以构造出分裂的回归误差。分裂的目标是最大程度的减小回归误差，因此把分裂的误差指标定义为分裂之前的回归误差减去分裂之后左右子树的回归误差

$$\Delta E = E(D) - \frac{N_L}{N} E(D_L) - \frac{N_R}{N} E(D_R)$$

将误差的计算公式代入上式，可以得到：

$$\Delta E = \frac{1}{N}\left(\sum_{i=1}^{N} y_i^2 - \frac{1}{N}\left(\sum_{i=1}^{N} y_i\right)^2\right) - \frac{N_L}{N}\left(\frac{1}{N_L}\left(\sum_{i=1}^{N_L} y_i^2 - \frac{1}{N_L}\left(\sum_{i=1}^{N_L} y_i\right)^2\right)\right) -$$

$$\frac{N_R}{N}\left(\frac{1}{N_R}\left(\sum_{i=1}^{N_R} y_i^2 - \frac{1}{N_R}\left(\sum_{i=1}^{N_R} y_i\right)^2\right)\right)$$

$$= -\frac{1}{N^2}\left(\sum_{i=1}^{N} y_i\right)^2 + \frac{1}{N}\left(\frac{1}{N_L}\left(\sum_{i=1}^{N_L} y_i\right)^2 + \frac{1}{N_R}\left(\sum_{i=1}^{N_R} y_i\right)^2\right)$$

由于 N 和 $-\frac{1}{N^2}\left(\sum_{i=1}^{N} y_i\right)^2$ 是常数，要让上式最大化等价于让下式最大化

$$\Delta E = \frac{1}{N_L}\left(\sum_{i=1}^{N_L} y_i\right)^2 + \frac{1}{N_R}\left(\sum_{i=1}^{N_R} y_i\right)^2$$

寻找最佳分裂时要计算上面的值，让该值最大化的分裂就是最佳分裂。回归树对类别型特征的处理和分类树类似，只是 E 值的计算公式不同，其他的过程相同。

第 6 章 k 近邻算法与距离度量学习

1.简述 k 近邻算法的预测算法的原理。

对于分类问题，给定 l 个训练样本 $(\mathbf{x}_i, y_i)$，其中 $\mathbf{x}_i$ 为维特征向量，y_i 为标签值，设定参数 k，假设类型数为 c，待分类样本的特征向量为 $\mathbf{x}$。预测算法的流程为

找出训练样本集 D 中找出离 $\mathbf{x}$ 最近的 k 个样本，假设这些样本的集合为 N

统计集合 N 中每个类的样本数 $C_i, i = 1,...,c$

返回 $\arg\max_i C_i$

在这里 $\arg\max_i C_i$ 表示最大的 C_i 值对应的那个类 i。如果 $k = 1$，k 近邻算法退化成最近邻算法。

k 近邻算法也可以用于回归问题。假设离测试样本最近的 k 个训练样本的标签值为 y_i，则对样本的回归预测输出值为

$$\hat{y} = \left(\sum_{i=1}^{k} y_i\right) / k$$

2.简述 k 的取值对 k 近邻算法的影响。

如果其值太小，则容易受到噪声的影响，导致泛函性能下降，出现过拟合。如果 k 值等于训练样数，则对于任意的预测样本，都会将其预测为训练样本集中数量最大的类。

3.距离函数需要满足哪些数学条件？

两个向量之间的距离为 $d\left(\mathbf{x}_i,\mathbf{x}_j\right)$，这是一个将两个维数相同的向量映射为一个实数的函数。距离函数必须满足以下 4 个条件。

第 1 个条件是非负性，即距离不能是一个负数

$$d\left(\mathbf{x}_i,\mathbf{x}_j\right)\geq 0$$

这意味着距离函数的值总是大于等于 0。

第 2 个条件是对称性，即 A 到 B 的距离和 B 到 A 的距离必须相等：

$$d\left(\mathbf{x}_i,\mathbf{x}_j\right)=d\left(\mathbf{x}_j,\mathbf{x}_i\right)$$

这意味着距离函数是对称函数。

第 3 个条件是区分性，如果两点间的距离为 0，则两个点必须相同

$$d\left(\mathbf{x}_i,\mathbf{x}_j\right)=0\Rightarrow\mathbf{x}_i=\mathbf{x}_j$$

第 4 个条件是三角不等式

$$d\left(\mathbf{x}_i,\mathbf{x}_k\right)+d\left(\mathbf{x}_k,\mathbf{x}_j\right)\geq d\left(\mathbf{x}_i,\mathbf{x}_j\right)$$

4.列举常见的距离函数。

欧氏距离。即 n 维欧氏空间中两点之间的距离。对于 $\mathbb{R}^n$ 空间中有两个点 $\mathbf{x}$ 和 $\mathbf{y}$，它们之间的距离定义为

$$d\left(\mathbf{x},\mathbf{y}\right)=\sqrt{\sum_{i=1}^{n}\left(x_i-y_i\right)^2}$$

Chebyshev 距离

$$d\left(\mathbf{x},\mathbf{y}\right)=\max_i\left(\left|x_i-y_i\right|\right)$$

Manhattan 距离

$$d\left(\mathbf{x},\mathbf{y}\right)=\sum_{i=1}^{n}\left|x_i-y_i\right|$$

Minkowski 距离

$$d\left(\mathbf{x},\mathbf{y}\right)=\left(\sum_{i=1}^{n}\left|x_i-y_i\right|^p\right)^{\frac{1}{p}}$$

Mahalanobis 距离是一种概率意义上的距离，给定两个向量 $\mathbf{x}$ 和 $\mathbf{y}$ 以及矩阵 $\mathbf{S}$，它定义为：

$$d(\mathbf{x},\mathbf{y})=\sqrt{(\mathbf{x}-\mathbf{y})^{\mathrm{T}}\mathbf{S}(\mathbf{x}-\mathbf{y})}$$

有时候也会使用该距离的平方

$$d(\mathbf{x},\mathbf{y})=(\mathbf{x}-\mathbf{y})^{\mathrm{T}}\mathbf{S}(\mathbf{x}-\mathbf{y})$$

要保证根号内的值非负，且满足区分性条件，则矩阵 $\mathbf{S}$ 必须是正定的。

Bhattacharyya 距离（也称为巴氏距离）定义了两个离散型或连续型概率分布的相似性。对于离散型随机变量的分布，它的定义为：

$$d(\mathbf{x},\mathbf{y})=-\ln\left(\sum_{i=1}^{n}\sqrt{x_i\cdot y_i}\right)$$

其中 x_i, y_i 为可以看作两个随机变量取某一值的概率，它们是向量 $\mathbf{x}$ 和 $\mathbf{y}$ 的分量，它们的值必须非负。两个向量越相似，这个距离值越小。

5.解释距离度量学习的原理。

通过机器学习的方式得到距离函数。

6.解释 LMNN 算法的原理。

LMNN 寻找一个变换矩阵，使得变换后每个样本的个最近邻居都和它是同一个类，而不同类型的样本通过一个大的间隔被分开。

假设原始的样本点为 $\mathbf{x}$，变换之后的点为 $\mathbf{y}$，在这里要寻找的是如下线性变换

$$\mathbf{y}=\mathbf{L}\mathbf{x}$$

其中 $\mathbf{L}$ 为线性变换矩阵。首先定义目标邻居的概念。一个样本的目标邻居是和该样本同类型的样本。我们希望通过学习得到的线性变换让样本最接近的邻居就是它的目标邻居

$$j\rightsquigarrow i$$

表示训练样本 $\mathbf{x}_j$ 是样本 $\mathbf{x}_i$ 的目标邻居。这个概念不是对称的，$\mathbf{x}_j$ 是 $\mathbf{x}_i$ 的目标邻居不等于 $\mathbf{x}_i$ 是 $\mathbf{x}_j$ 的目标邻居。

为了保证 kNN 算法能准确的分类，任意一个样本的目标邻居样本要比其他类别的样本更接近于该样本。对每个样本，我们可以将目标邻居想象成为这个样本建立起了一个边界，使得和本样本标签值不同的样本无法入侵进来。训练样本集中，侵入这个边界并且和该样本不同标签值的样本称为冒充者，这里的目标是最小化冒充者的数量。

为了增强 kNN 分类的泛化性能，要让冒充者离由目标邻居估计出的边界的距离尽可能的远。通过在 kNN 决策边界周围加上一个大的安全间隔，可以有效的提高算法的鲁棒性。

接下来定义冒充者的概念。对于训练样本 $\mathbf{x}_i$，其标签值为 y_i，目标邻居为 $\mathbf{x}_j$，冒充者是指那些和 $\mathbf{x}_i$ 有不同的标签值并且满足如下不等式的样本 $\mathbf{x}_l$：

$$\left\|\mathrm{L}\left(\mathbf{x}_i-\mathbf{x}_l\right)\right\|^2 \leq \left\|\mathrm{L}\left(\mathbf{x}_i-\mathbf{x}_j\right)\right\|^2+1$$

其中 $\mathbf{L}$ 为线性变换矩阵，左乘这个矩阵相当于对向量进行线性变换。根据上面的定义，冒充者就是闯入了一个样本的分类间隔区域并且和该样本标签值不同的样本。这个线性变换实际上确定了一种距离定义：

$$\left\|\mathbf{L}\left(\mathbf{x}_i-\mathbf{x}_j\right)\right\|=\sqrt{\left(\mathbf{L}\left(\mathbf{x}_i-\mathbf{x}_j\right)\right)^{\mathrm{T}}\left(\mathbf{L}\left(\mathbf{x}_i-\mathbf{x}_j\right)\right)}=\sqrt{\left(\mathbf{x}_i-\mathbf{x}_j\right)^{\mathrm{T}}\mathbf{L}^{\mathrm{T}}\mathbf{L}\left(\mathbf{x}_i-\mathbf{x}_j\right)}$$

其中，$\mathbf{L}^{\mathrm{T}}\mathbf{L}$ 就是 Mahalanobis 距离中的矩阵。训练时优化的损失函数由推损失函数和拉损失函数两部分构成。拉损失函数的作用是让和样本标签相同的样本尽可能与它接近

$$\varepsilon_{pull}\left(\mathbf{L}\right)=\sum_{j\rightsquigarrow i}\left\|\mathbf{L}\left(\mathbf{x}_i-\mathbf{x}_j\right)\right\|^2$$

推损失函数的作用是把不同类型的样本推开

$$\varepsilon_{push}\left(\mathbf{L}\right)=\sum_{i,j\rightsquigarrow i}\sum_{l}\left(1-y_{il}\right)\left[1+\left\|\mathbf{L}\left(\mathbf{x}_i-\mathbf{x}_j\right)\right\|^2-\left\|\mathbf{L}\left(\mathbf{x}_i-\mathbf{x}_l\right)\right\|^2\right]_+$$

如果 $y_i=y_j$，则 $y_{ij}=1$，否则 $y_{ij}=0$。函数 $\left[z\right]_+$ 定义为

$$\left[z\right]_+=\max\left(z,0\right)$$

如果两个样本类型相同，则有

$$1-y_{il}=0$$

因此推损失函数只对不同类型的样本起作用。总损失函数由这两部分的加权和构成

$$\varepsilon\left(\mathbf{L}\right)=\left(1-\mu\right)\varepsilon_{pull}\left(\mathbf{L}\right)+\mu\varepsilon_{push}\left(\mathbf{L}\right)$$

在这里 μ 是人工设定的参数。求解该最小化问题即可得到线性变换矩阵。

7.解释 ITML 算法的原理。

ITML 的优化目标是在保证同类样本距离相近，不同类样本之间距离远的约束条件下，迫使度量矩阵所代表的正态分布接近于某一先验概率分布。算法使用了信息论中的 KL 散度，因此得名。这是一种有监督的局部度量学习算法。

假设有 n 个 $\mathbb{R}^d$ 中的样本点 $\mathbf{x}_1,...,\mathbf{x}_n$。度量矩阵为 $\mathbf{A}$，这里的距离采用马氏距离的平方和。如果两个样本点之间相似，则有如下的不等式约束

$$d_{\mathbf{A}}\left(\mathbf{x}_i,\mathbf{x}_j\right)\leq u$$

即它们之间的距离小于某一较小的阈值 u。这一约束通常用于同类的样本点之间。反之

如果两个样本点之间不相似，则有如下不等式约束

$$d_{\mathbf{A}}\left(\mathbf{x}_i,\mathbf{x}_j\right)\geq l$$

其中 l 为一个较大的阈值。这一约束通常用于不同类的样本点之间。

矩阵 $\mathbf{A}$ 通常要符合某些先验知识。例如，如果数据服从正态分布，则该矩阵为正态分布协方差矩阵的逆矩阵；而对有些场景，欧氏距离平方作为距离函数有很好的效果，此时该矩阵为单位矩阵。因此可以对矩阵 $\mathbf{A}$ 正则化，迫使其尽可能接近于某一已知的马氏距离矩阵 $\mathbf{A}_0$。

因此需要衡量 $\mathbf{A}$ 与 $\mathbf{A}_0$ 之间的接近程度。如果以度量矩阵作为协方差矩阵的逆矩阵，则此多维正态分布为

$$p\left(\mathbf{x};\mathbf{A}\right)=\frac{1}{Z}\exp\left(-\frac{1}{2}d_{\mathbf{A}}\left(\mathbf{x},\boldsymbol{\mu}\right)\right)$$

其中 Z 为归一化常数，$\boldsymbol{\mu}$ 为均值向量，$\mathbf{A}^{-1}$ 为协方差矩阵。如果将马氏距离所作用的样本集看作服从正态分布，则可以用 KL 距离衡量二者的差异。根据 KL 散度的定义，这两个度量矩阵所代表的正态分布之间的 KL 散度为

$$D_{\mathrm{KL}}\left(p\left(\mathbf{x};\mathbf{A}_0\right)\middle\|p\left(\mathbf{x};\mathbf{A}\right)\right)=\int_{\mathbb{R}^n}p\left(\mathbf{x};\mathbf{A}_0\right)\ln\frac{p\left(\mathbf{x};\mathbf{A}_0\right)}{p\left(\mathbf{x};\mathbf{A}\right)}d\mathbf{x}$$

因此得到如下优化问题

$$\begin{aligned}&\min_{\mathbf{A}}D_{\mathrm{KL}}\left(p\left(\mathbf{x};\mathbf{A}_0\right)\middle\|p\left(\mathbf{x};\mathbf{A}\right)\right)\\&d_{\mathbf{A}}\left(\mathbf{x}_i,\mathbf{x}_j\right)\leq u\qquad\left(i,j\right)\in S\\&d_{\mathbf{A}}\left(\mathbf{x}_i,\mathbf{x}_j\right)\geq l\qquad\left(i,j\right)\in D\end{aligned}$$

其中 S 为相似的样本对的集合，D 为不相似的样本对的集合。目标函数为两个矩阵之间的 KL 散度，实现先验知识。由于 $\mathbf{x}_i$ 和 $\mathbf{x}_j$ 都是常数，因此不等式约束为线性约束。

如果定义

$$D_{ld}\left(\mathbf{A},\mathbf{A}_0\right)=\mathrm{tr}\left(\mathbf{A}\mathbf{A}_0^{-1}\right)-\ln\det\left(\mathbf{A}\mathbf{A}_0^{-1}\right)-n$$

如果假设两个正态分布的矩阵相等，则它们之间的 KL 散度为

$$\begin{aligned}D_{\mathrm{KL}}\left(p\left(\mathbf{x};\mathbf{A}_0\right)\middle\|p\left(\mathbf{x};\mathbf{A}\right)\right)&=\frac{1}{2}D_{ld}\left(\mathbf{A}_0^{-1},\mathbf{A}^{-1}\right)\\&=\frac{1}{2}D_{ld}\left(\mathbf{A},\mathbf{A}_0\right)\end{aligned}$$

从而得到如下的优化问题

$$\min_{\mathbf{A}} D_{ld}\left(\mathbf{A},\mathbf{A}_0\right)$$
$$\mathrm{tr}\left(\mathbf{A}\left(\mathbf{x}_i-\mathbf{x}_j\right)\left(\mathbf{x}_i-\mathbf{x}_j\right)^{\mathrm{T}}\right)\le u \qquad (i,j)\in S$$
$$\mathrm{tr}\left(\mathbf{A}\left(\mathbf{x}_i-\mathbf{x}_j\right)\left(\mathbf{x}_i-\mathbf{x}_j\right)^{\mathrm{T}}\right)\ge l \qquad (i,j)\in D$$

8.解释 NCA 算法的原理。

NCA 同样与 k 近邻算法有关。在保证其优化目标是使得每个样本的同类样本被 k 近邻算法正确分类的概率最大化，以此构造目标函数。这是一种有监督的局部度量学习算法。

首先定义每个样本点的邻居的概率分布，是其他样本所有样本是此样本邻居的概率。样本 $\mathbf{x}_j$ 是 $\mathbf{x}_i$ 的邻居的概率定义通过 softmax 归一化进行计算

$$p_{ij}=\frac{\exp\left(-\left\|\mathbf{A}\mathbf{x}_i-\mathbf{A}\mathbf{x}_j\right\|^2\right)}{\sum_{k,k\neq i}\exp\left(-\left\|\mathbf{A}\mathbf{x}_i-\mathbf{A}\mathbf{x}_k\right\|^2\right)}$$

这两个样本点经过变换之后相距越远则此概率值越小；反之则越大。样本成为其自身的邻居的概率定义为 0，即 $p_{ii}=0$。在对样本点 i 进行分类时，如果采用这些邻接作为其标签值，则可以计算出样本点被正确分类的概率。定义 p_i 为样本点 i 被正确的分类的概率，是它所有同类样本成为其邻居的概率之和

$$p_i=\sum_{j\in C_i}p_{ij}$$

其中 C_i 为 i 的同类样本集合，即 $C_i=\left\{j\middle|y_i=y_j\right\}$。$y_i$ 为样本的类别标签值。NCA 的优化目标是所有样本的 p_i 之和

$$L\left(\mathbf{A}\right)=\sum_i p_i=\sum_i\sum_{j\in C_i}p_{ij}$$

如果定义向量 $\mathbf{x}_{ij}=\mathbf{x}_i-\mathbf{x}_j$，则优化变量对 $\mathbf{A}$ 的梯度为

$$\nabla_{\mathbf{A}}L=-2\mathbf{A}\sum_i\sum_{j\in C_i}p_{ij}\left(\mathbf{x}_{ij}\mathbf{x}_{ij}^{\mathrm{T}}-\sum_k p_{ik}\mathbf{x}_{ik}\mathbf{x}_{ik}^{\mathrm{T}}\right)$$

整理后可得

$$\nabla_{\mathbf{A}}L=2\mathbf{A}\sum_i\left(p_i\sum_k p_{ik}\mathbf{x}_{ik}\mathbf{x}_{ik}^{\mathrm{T}}-\sum_{j\in C_i}p_{ij}\mathbf{x}_{ij}\mathbf{x}_{ij}^{\mathrm{T}}\right)$$

计算出梯度之后可用梯度下降法或其他优化算法进行迭代。

第 7 章 数据降维

1.使用数据降维算法的目的是什么？

数据降维算法的目标是将向量变换到低维空间中，并保持原始空间中的某些结构信息，以达到某种目的，如避免维数灾难，数据可视化。

2.列举常见的数据降维算法。

PCA，KPCA，LDA，KLDA，LLE，拉普拉斯特征映射，局部保持投影，等距映射，SEN，t-SNE，MDS。

3.常见的降维算法中，哪些是监督降维，哪些是无监督降维？

LDA 和 KLDA 是有监督的，其他的都是无监督的。

4.什么是流形？

流形是几何中的一个概念，它是高维空间中的几何结构，即空间中的点构成的集合，可以简单的将流形理解成二维空间的曲线，三维空间的曲面在更高维空间的推广。

5.根据最小化重构误差准则推导 PCA 投影矩阵的计算公式。

最小化如下误差函数

$$L(a,\mathbf{e})=\sum_{i=1}^{n}\left\|\mathbf{m}+a_i\mathbf{e}-\mathbf{x}_i\right\|^2$$

为了求这个函数的极小值，对 a_i 求偏导数并令其为 0 可以得到：

$$2\mathbf{e}^{\mathrm{T}}\left(\mathbf{m}+a_i\mathbf{e}-\mathbf{x}_i\right)=0$$

变形后得到：

$$a_i\mathbf{e}^{\mathrm{T}}\mathbf{e}=\mathbf{e}^{\mathrm{T}}\left(\mathbf{x}_i-\mathbf{m}\right)$$

由于 $\mathbf{e}$ 是单位向量，因此 $\mathbf{e}^{\mathrm{T}}\mathbf{e}=1$，最后得到：

$$a_i=\mathbf{e}^{\mathrm{T}}\left(\mathbf{x}_i-\mathbf{m}\right)$$

这就是样本和均值的差对向量 $\mathbf{e}$ 做投影。现在的问题是 $\mathbf{e}$ 的值如何选确定。定义如下的散布矩阵

$$\mathbf{S}=\sum_{i=1}^{n}\left(\mathbf{x}_i-\mathbf{m}\right)\left(\mathbf{x}_i-\mathbf{m}\right)^{\mathrm{T}}$$

这个矩阵是协方差矩阵的 n 倍，协方差矩阵的计算公式为

$$\mathbf{\Sigma}=\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i-\mathbf{m}\right)\left(\mathbf{x}_i-\mathbf{m}\right)^{\mathrm{T}}$$

将上面求得的a_i代入目标函数中，得到只有变量$\mathbf{e}$的函数

$$\begin{aligned}L\left(\mathbf{e}\right)&=\sum_{i=1}^{n}\left(\alpha_i\mathbf{e}+\mathbf{m}-\mathbf{x}_i\right)^{\mathrm{T}}\left(\alpha_i\mathbf{e}+\mathbf{m}-\mathbf{x}_i\right)\\&=\sum_{i=1}^{n}\left(\left(\alpha_i\mathbf{e}\right)^{\mathrm{T}}\alpha_i\mathbf{e}+2\left(\alpha_i\mathbf{e}\right)^{\mathrm{T}}\left(\mathbf{m}-\mathbf{x}_i\right)+\left(\mathbf{m}-\mathbf{x}_i\right)^{\mathrm{T}}\left(\mathbf{m}-\mathbf{x}_i\right)\right)\\&=\sum_{i=1}^{n}a_i^2-2\sum_{i=1}^{n}a_i^2+\sum_{i=1}^{n}\left(\mathbf{m}-\mathbf{x}_i\right)^{\mathrm{T}}\left(\mathbf{m}-\mathbf{x}_i\right)\\&=-\sum_{i=1}^{n}\left(\mathbf{e}^{\mathrm{T}}\left(\mathbf{x}_i-\mathbf{m}\right)\right)^2+\sum_{i=1}^{n}\left(\mathbf{m}-\mathbf{x}_i\right)^{\mathrm{T}}\left(\mathbf{m}-\mathbf{x}_i\right)\\&=-\sum_{i=1}^{n}\left(\mathbf{e}^{\mathrm{T}}\left(\mathbf{x}_i-\mathbf{m}\right)\left(\mathbf{x}_i-\mathbf{m}\right)^{\mathrm{T}}\mathbf{e}\right)+\sum_{i=1}^{n}\left(\mathbf{m}-\mathbf{x}_i\right)^{\mathrm{T}}\left(\mathbf{m}-\mathbf{x}_i\right)\\&=-\mathbf{e}^{\mathrm{T}}\mathbf{S}\mathbf{e}+\sum_{i=1}^{n}\left(\mathbf{m}-\mathbf{x}_i\right)^{\mathrm{T}}\left(\mathbf{m}-\mathbf{x}_i\right)\end{aligned}$$

上式的后半部分和$\mathbf{e}$无关，由于$\mathbf{e}$是单位向量，因此有$\left\|\mathbf{e}\right\|=1$的约束，这可以写成$\mathbf{e}^{\mathrm{T}}\mathbf{e}=1$。要求解的是一个带等式约束的极值问题，可以使用拉格朗日乘数法。构拉格朗日函数：

$$L\left(\mathbf{e},\lambda\right)=-\mathbf{e}^{\mathrm{T}}\mathbf{S}\mathbf{e}+\lambda\left(\mathbf{e}^{\mathrm{T}}\mathbf{e}-1\right)$$

对$\mathbf{e}$求梯度并令其为$\mathbf{0}$可以得到：

$$-2\mathbf{S}\mathbf{e}+2\lambda\mathbf{e}=\mathbf{0}$$

即：

$$\mathbf{S}\mathbf{e}=\lambda\mathbf{e}$$

λ就是散度矩阵的特征值，$\mathbf{e}$为它对应的特征向量，因此上面的最优化问题可以归结为矩阵的特征值和特征向量问题。矩阵$\mathbf{S}$的所有特征向量给出了上面极值问题的所有极值点。矩阵$\mathbf{S}$是实对称半正定矩阵。这里需要最大化$\mathbf{e}^{\mathrm{T}}\mathbf{S}\mathbf{e}$的值，由于

$$\mathbf{e}^{\mathrm{T}}\mathbf{S}\mathbf{e}=\lambda\mathbf{e}^{\mathrm{T}}\mathbf{e}=\lambda$$

因此λ为散度矩阵最大的特征值时，$\mathbf{e}^{\mathrm{T}}\mathbf{S}\mathbf{e}$有极大值，目标函数取得极小值。将上述结论从一维推广到d'维，每个向量可以表示成：

$$\mathbf{x}=\mathbf{m}+\sum_{i=1}^{d'}a_i\mathbf{e}_i$$

在这里$\mathbf{e}_i$都是单位向量，并且相互正交，即寻找低维空间中的标准正交基。误差函数变

成：

$$\sum_{i=1}^{n}\left\|\mathbf{m}+\sum_{j=1}^{d'}a_{ij}\mathbf{e}_j-\mathbf{x}_i\right\|^2$$

和一维情况类似，可以证明，使得该函数取最小值的$\mathbf{e}_j$为散度矩阵最大的d'个特征值对应的单位长度特征向量。即求解下面的优化问题：

$$\min_{\mathbf{W}}-\mathrm{tr}\left(\mathbf{W}^{\mathrm{T}}\mathbf{S}\mathbf{W}\right)$$

$$\mathbf{W}^{\mathrm{T}}\mathbf{W}=\mathbf{I}$$

其中tr为矩阵的迹，$\mathbf{I}$为单位矩阵，该等式约束保证投影基向量是标准正交基。与一维的情况相同，其解为如下矩阵的特征值

$$\left(\sum_{i=1}^{n}\left(\mathbf{x}_i-\mathbf{m}\right)\left(\mathbf{x}_i-\mathbf{m}\right)^{\mathrm{T}}\right)\mathbf{w}=\lambda\mathbf{w}$$

矩阵$\mathbf{W}$的列$\mathbf{e}_j$是要求解的基向量。

6.解释 PCA 降维算法的流程。

计算投影矩阵的流程：

1.计算样本集的均值向量。将所有向量减去均值，这称为白化。

2.计算样本集的协方差矩阵。

3.对方差矩阵进行特征值分解，得到所有特征值与特征向量。

4.将特征值从大到小排序，保留最大的一部分特征值对应的特征向量，以它们为行，形成投影矩阵。

投影算法的流程：

1.将样本减掉均值向量。

2.左乘投影矩阵，得到降维后的向量。

7.解释 PCA 重构算法的流程。

向量重构的流程为：

1.输入向量左乘投影矩阵的转置矩阵。

2.加上均值向量，得到重构后的结果。

8.解释 LLE 的原理。

局部线性嵌入将高维数据投影到低维空间中，并保持数据点之间的局部线性关系。其核心思想是每个点都可以由与它相邻的多个点的线性组合来近似重构，投影到低维空间之后要保持这种线性重构关系，即有相同的重构系数。

假设数据集由l个D维向量$\mathbf{x}_i$组成，它们分布在D维空间中的一个流形附近。每个数

据点和它的邻居位于或者接近于流形的一个局部线性片段上，即可以用邻居点的线性组合来重构，组合系数刻画了局部面片的几何特性：

$$\mathrm{x}_i \approx \sum_j w_{ij}\mathrm{x}_j$$

权重 w_{ij} 为第 j 个数据点对第 i 个点的组合权重，这些点的线性组合被用来近似重构数据点 i。权重系数通过最小化下面的重构误差确定：

$$\min_{w_{ij}} \sum_{i=1}^{l} \left\| \mathrm{x}_i - \sum_{j=1}^{l} w_{ij}\mathrm{x}_j \right\|^2$$

在这里还加上了两个约束条件：每个点只由它的邻居来重构，如果 x_j 不在 x_i 的邻居集合里则权重值为 0。另外限定权重矩阵的每一行元素之和为 1，即：

$$\sum_j w_{ij} = 1$$

这是一个带约束的优化问题，求解该问题可以得到权重系数。假设算法将向量从 D 维空间的 x 映射为 d 维空间的 y。每个点在 d 维空间中的坐标由下面的最优化问题确定：

$$\min_{\mathrm{y}_i} \sum_{i=1}^{l} \left\| \mathrm{y}_i - \sum_{j=1}^{l} w_{ij}\mathrm{y}_j \right\|^2$$

这里的权重和上一个优化问题的值相同，在前面已经得到。优化的目标是 y_i，这个优化问题等价于求解稀疏矩阵的特征值问题。得到 y 之后，即完成了从 D 维空间到 d 维空间的非线性降维。

9.名词解释：图的拉普拉斯矩阵。

假设图 G 的邻接矩阵为 $\mathbf{W}$，加权度矩阵为 $\mathbf{D}$。图拉普拉斯矩阵定义为加权度矩阵与邻接矩阵之差

$$\mathbf{L} = \mathbf{D} - \mathbf{W}$$

10.解释 t-SNE 的原理。

t-SNE 是对 SNE 的改进。t-SNE 采用了对称的概率计算公式，另外在低维空间中计算样本点之间的概率时使用 t 分布代替了正态分布。

在 SNE 中 $p_{i|j}$ 和 $p_{j|i}$ 是不相等的，因此概率值不对称。可以用两个样本点的联合概率替代它们之间的条件概率解决此问题。在高维空间中两个样本点的联合概率定义为

$$p_{ij}=\frac{\exp\left(-\left\|\mathbf{x}_i-\mathbf{x}_j\right\|^2/2\sigma^2\right)}{\sum_{k\neq l}\exp\left(-\left\|\mathbf{x}_k-\mathbf{x}_l\right\|^2/2\sigma^2\right)}$$

显然这个定义是对称的，即 $p_{ij}=p_{ji}$。同样的，低维空间中两个点的联合概率为

$$q_{ij}=\frac{\exp\left(-\left\|\mathbf{y}_i-\mathbf{y}_j\right\|^2\right)}{\sum_{k\neq i}\exp\left(-\left\|\mathbf{y}_k-\mathbf{y}_l\right\|^2\right)}$$

目标函数采用 KL 散度，定义为

$$L\left(\mathbf{y}_i\right)=D_{\mathrm{KL}}\left(P\left|Q\right.\right)=\sum_{i=1}^{l}\sum_{j=1}^{l}p_{ij}\log\frac{p_{ij}}{q_{ij}}$$

但这样定义联合概率会存在异常值问题。如果某一个样本 $\mathbf{x}_i$ 是异常点即离其他点很远，则所有的 $\left\|\mathbf{x}_i-\mathbf{x}_j\right\|^2$ 都很大，因此与 $\mathbf{x}_i$ 有关的 p_{ij} 很小，从而导致低维空间中的 $\mathbf{y}_i$ 对目标函数影响很小。解决方法是重新定义高维空间中的联合概率，具体为

$$p_{ij}=\frac{p_{j|i}+p_{i|j}}{2n}$$

这样能确保对所有的 $\mathbf{x}_i$ 均有

$$\sum_j p_{ij}>\frac{1}{2n}$$

这种方法称为对称 SNE。

对称 SNE 虽然对 SNE 做了改进，但还存在拥挤问题，各类样本降维后聚集在一起而缺乏区分度。解决方法是用 t 分布替代高斯分布，计算低维空间中的概率值。相比于正态分布，t 分布更长尾。如果在低维空间中使用 t 分布，则在高维空间中距离近的点，在低维空间中距离也要近；但在高维空间中距离远的点，在低维空间中距离要更远。因此可以有效的拉大各个类之间的距离。使用 t 分布之后，低维空间中的概率计算公式为

$$q_{ij}=\frac{\left(1+\left\|\mathbf{y}_i-\mathbf{y}_j\right\|^2\right)^{-1}}{\sum_{k\neq l}\left(1+\left\|\mathbf{y}_k-\mathbf{y}_l\right\|^2\right)^{-1}}$$

目标函数同样采用 KL 散度。目标函数对的 $\mathbf{y}_i$ 梯度为

$$\nabla_{\mathbf{y}_i}L=4\sum_j\left(p_{ij}-q_{ij}\right)\left(\mathbf{y}_i-\mathbf{y}_j\right)\left(1+\left\|\mathbf{y}_i-\mathbf{y}_j\right\|^2\right)^{-1}$$

11.解释 KPCA 的原理。

核主成分分析是核技术与主成分分析相结合的产物，它使用核技术将向量映射到更高维的空间，然后用 PCA 进行处理。核技术用一个映射（称为核映射）将原始向量变换到另外一个空间，然后进行处理。

12.证明图的拉普拉斯矩阵半正定。

根据拉普拉斯矩阵的定义 $\mathbf{L}=\mathbf{D}-\mathbf{W}$，并且 $\mathbf{W}$ 是对称矩阵，有

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(y_i-y_j\right)^2 w_{ij} &= \sum_{i=1}^{n}\sum_{j=1}^{n}\left(y_i^2+y_j^2-2y_iy_j\right)w_{ij}\\ &=\sum_{i=1}^{n}\sum_{j=1}^{n}y_i^2w_{ij}+\sum_{i=1}^{n}\sum_{j=1}^{n}y_j^2w_{ij}-\sum_{i=1}^{n}\sum_{j=1}^{n}2y_iy_jw_{ij}\\ &=\sum_{i=1}^{n}y_i^2\left(\sum_{j=1}^{n}w_{ij}\right)+\sum_{j=1}^{n}y_j^2\left(\sum_{i=1}^{n}w_{ij}\right)-\sum_{i=1}^{n}\sum_{j=1}^{n}2y_iy_jw_{ij}\\ &=\sum_{i=1}^{n}y_i^2d_{ii}+\sum_{j=1}^{n}y_j^2d_{jj}-\sum_{i=1}^{n}\sum_{j=1}^{n}2y_iy_jw_{ij}\\ &=2\left(\sum_{i=1}^{n}y_i^2d_{ii}-\sum_{i=1}^{n}\sum_{j=1}^{n}2y_iy_jw_{ij}\right)\\ &=2\left(\mathbf{y}^{\mathrm{T}}\mathbf{D}\mathbf{y}-\mathbf{y}^{\mathrm{T}}\mathbf{W}\mathbf{y}\right)\\ &=2\mathbf{y}^{\mathrm{T}}\mathbf{L}\mathbf{y}\end{aligned}$$

因此拉普拉斯矩阵是半正定的。

13.解释拉普拉斯特征映射的原理。

拉普拉斯特征映射为样本集构造邻接图，计算图的拉普拉斯矩阵，对其进行特征值分解，从而完成数据降维。

算法的第一步是构造样本集的邻接图，保证距离相近的样本之间的边权重更大。

第二步是计算图的拉普拉斯矩阵。

第三步是特征映射。求解如下广义特征值和特征向量问题

$$\mathbf{L}\mathbf{f}=\lambda\mathbf{D}\mathbf{f}$$

假设 $\mathbf{f}_0,...,\mathbf{f}_{k-1}$ 是这个广义特征值问题的解，它们按照特征值的大小升序排列，即：

$$0=\lambda_0\le...\le\lambda_{k-1}$$

去掉值为 0 的特征值 λ_0，用剩下的前 m 个特征向量来构造投影矩阵，将向量投影到以它们为基的空间中。

拉普拉斯特征映射在特定的意义下最好的保留了数据的局部信息。根据一组数据点 $\mathbf{x}_1,...,\mathbf{x}_n$，我们构造了带权重的图，假设这个图是联通的。首先考虑最简单的情况，将这个图映射到一条直线上，保证相连的点在映射之后靠的越近越好。假设这些点映射之后的坐标为 $\mathbf{y}=\left(y_1,y_2,...,y_n\right)^{\mathrm{T}}$。则目标函数可以采用下面的定义

$$\min \sum_{i=1}^{n}\sum_{j=1}^{n}\left(y_i - y_j\right)^2 w_{ij}$$

这个目标函数意味着，如果 $\mathbf{x}_i$ 和 $\mathbf{x}_j$ 距离很近，则 y_i 和 y_j 也必须距离很近，否则会出现大的损失函数值，因为 w_{ij} 的值很大。反之，如果两个点 $\mathbf{x}_i$ 和 $\mathbf{x}_j$ 距离很远，则 w_{ij} 的值很小，如果 y_i 和 y_j 距离很远，也不会导致大的损失值。可以证明，对任意的 $\mathbf{y}$，下式成立

$$\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(y_i - y_j\right)^2 w_{ij} = \mathbf{y}^{\mathrm{T}}\mathbf{L}\mathbf{y}$$

这个最优化问题可以表述为

$$\min_{\mathbf{y}} \mathbf{y}^{\mathrm{T}}\mathbf{L}\mathbf{y}$$
$$\mathbf{y}^{\mathrm{T}}\mathbf{D}\mathbf{y} = 1$$

这里的等式约束条件 $\mathbf{y}^{\mathrm{T}}\mathbf{D}\mathbf{y} = 1$ 消除了投影向量 $\mathbf{y}$ 的缩放，因为 $\mathbf{y}$ 与 $k\mathbf{y}$ 本质上是一个投影结果。矩阵 $\mathbf{D}$ 提供了对图的顶点的一种度量，如果 d_{ii} 越大，则其对应的第 i 个顶点提供的信息越大，这也符合我们直观的认识，如果一个顶点连接的边的总权重越大，则其在图里起的作用也越大。上面的问题可以采用拉格朗日乘数法求解，构造拉格朗日乘子函数

$$L\left(\mathbf{y},\lambda\right) = \mathbf{y}^{\mathrm{T}}\mathbf{L}\mathbf{y} + \lambda\left(\mathbf{y}^{\mathrm{T}}\mathbf{D}\mathbf{y} - 1\right)$$

对 $\mathbf{y}$ 求梯度并令梯度为 0，可以得到

$$\nabla L\left(\mathbf{y},\lambda\right) = 2\mathbf{L}\mathbf{y} + 2\lambda\mathbf{D}\mathbf{y} = \mathbf{0}$$

由此可得

$$\mathbf{L}\mathbf{y} = \lambda\mathbf{D}\mathbf{y}$$

下面把这个结果推广到高维，假设将向量投影到 m 维的空间，则投影结果是一个 $n\times m$ 的矩阵，记为 $\mathbf{Y} = \left[\mathbf{y}_1\mathbf{y}_2...\mathbf{y}_m\right]$，其第 i 个行为第 i 个顶点投影后的坐标。仿照一维的情况构造目标函数

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\left\|\mathbf{y}_i - \mathbf{y}_j\right\|^2 w_{ij} = \mathrm{tr}\left(\mathbf{Y}^{\mathrm{T}}\mathbf{L}\mathbf{Y}\right)$$

这等价于求解如下问题

$$\min \mathrm{tr}\left(\mathbf{Y}^{\mathrm{T}}\mathbf{L}\mathbf{Y}\right)$$
$$\mathbf{Y}^{\mathrm{T}}\mathbf{D}\mathbf{Y} = \mathbf{I}$$

这个问题的最优解还是上面广义特征值问题的解，是最小的 m 个广义特征值对应的特征向量。

14.解释等距映射的原理。

等距映射（Isomap 使用了微分几何中测地线的思想，它希望数据在向低维空间映射之后能够保持流形上的测地线距离。

测地线源自于大地测量学，是地球上任意两点之间在球面上的最短路径。算法计算任意两个样本之间的测地距离，然后根据这个距离构造距离矩阵。最后通过距离矩阵求解优化问题完成数据的降维，降维之后的数据保留了原始数据点之间的距离信息。

在这里测地线距离通过图构造，是图的两个节点之间的最短距离。算法的第一步构造样本集的邻居图，这和前面介绍的两种方法相同。如果两个数据点之间的距离小于指定阈值或者其中一个节点在另外一个节点的邻居集合中，则两个节点是联通的。假设有 N 个样本，则邻居图有 N 个节点。邻居图的节点 i 和 j 之间边的权重为它们之间的距离 w_{ij}，距离的计算公式可以有多种选择。

第二步计算图中任意两点之间的最短路径长度，可以通过经典的 Dijkstra 算法实现。假设最短路径长度为 $d_G(i,j)$，由它构造如下矩阵：

$$\mathrm{D}_G = \left\{ d_G(i,j) \right\}$$

其元素是所有节点对之间的最短路径长度。算法的第三步根据矩阵 D_G 构造 d 维嵌入，这通过求解如下最优化问题实现：

$$\min_{\mathrm{y}} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(d_G(i,j) - \left\| \mathrm{y}_i - \mathrm{y}_j \right\| \right)^2$$

这个问题的解 y_i 即为降维之后的向量。这个目标函数的意义是向量降维之后任意两点之间的距离要尽量的接近在原始空间中这两点之间的最短路径长度，因此可以认为降维尽量保留了数据点之间的测地距离信息。

15.PCA 是有监督学习还是无监督学习？

无监督学习，因为样本集没有标签值。

绝大部分习题的答案在《机器学习-原理、算法与应用》一书中都有详细的讲解。购买链接为：

https://item.jd.com/12685964.html?dist=jd