机器学习（双语）习题参考答案

一

1.$\frac{\partial f}{\partial w_i}=(\sum_{i=1}^{d}w_ix_i-y)x_i$
2.$\frac{\partial f}{\partial w_i}=(\sum_{i=1}^{d}w_ix_i-y)x_i+2\lambda w_i$
3.$\frac{\partial f}{\partial w_i}=\frac{e^{-w_i}}{(1+e^{-w_i})^2}$
4.$\frac{\partial f}{\partial w_i}=\frac{4}{(e^{w_i}+e^{-w_i})^2}$
5.$\nabla_{\boldsymbol{x}}f(\boldsymbol{x})=\boldsymbol{x}$
6.$\nabla_{\boldsymbol{x}}f(\boldsymbol{x})=\frac{(A+A^T)}{2}\boldsymbol{x}+\boldsymbol{b}$

注：
$$\frac{\partial a^{\top}x}{\partial x}=\frac{\partial x^{\top}a}{\partial x}=a$$
$$\frac{\partial x^{\top}Ax}{\partial x}=\left(A+A^{\top}\right)x$$

二

1.将x_j扩充为$(1,x_j)$，$w=(w_0,w_1,...,w_d)^T$，则线性回归数学表达式可表示为：

$$h(x,w)=wx,w\in\mathbb{R}^{m\times(d+1)}$$

注：形式不唯一。

带有二范数正则项的均方误差损失函数为：

$$L(w)=\frac{1}{2m}\sum_{i=1}^{m}(y_i-h(x_i,w))^2+\lambda||w||_2^2$$

其关于w的梯度：

$$\frac{\partial L(w)}{\partial w_i}=\frac{1}{m}\sum_{i=1}^{m}(y_i-h(x_i,w))x_i+2\lambda w$$

2.记$\phi(x)=(\phi_0(x),\phi_1(x),...,\phi_k(x))^T$，其中$\phi_0(x)=1$。同理，广义线性回归模型的数学表达式为：

$$h(x,w)=w\phi(x),w\in\mathbb{R}^{m\times(\mathbf{K}+1)}$$

均方误差损失函数为：

$$L(w)=\frac{1}{2m}\sum_{i=1}^{m}(y_i-h(x_i,w))^2$$

其关于w的梯度：

$$\frac{\partial L(w)}{\partial w_i}=\frac{1}{m}\sum_{i=1}^{m}(y_i-h(x_i,w))\phi_i(x)$$

3.逻辑回归模型的数学表达式为：

$$h(x,w)=\frac{1}{1+e^{-wx}},w\in\mathbb{R}^{m\times(d+1)}$$

交叉熵误差损失函数为：

$$L(w)=\frac{1}{2m}\sum_{i=1}^{m}[-y_ilogh(x_i,w)-(1-y_i)log(1-h(x_i,w))]$$

其关于w的梯度：

$$\frac{\partial L(w)}{\partial w_i}=\frac{1}{m}\sum_{i=1}^{m}(y_i-h(x_i,w))x_i$$

4.最大间隔分类器的优化目标函数：

$$\begin{gathered}\underset{\omega,b}{min}\frac{1}{2}||\omega||^2\\s.t.y_i(\omega^Tx_i+b)\geq1\end{gathered}$$

构造拉格朗日函数：

$$L(x,w)=\frac{1}{2}||w||^2+\sum_{i=1}^{m}\lambda_i(1-y_i(\omega^Tx_i+b))$$

其梯度为：

$$\frac{\partial L(w)}{\partial w} = w + \sum_{i=1}^{m} \lambda_i(-y_i)x_i$$

其几何意义：

三

略

四

$$\frac{|b|}{||w||}$$

$$\frac{|w^T x + b|}{||w||}$$

五

1.$w^{(1)} \in \mathbb{R}^{m\times(d+1)}$,$w^{(2)} \in \mathbb{R}^{\mathbf{k}\times(m+1)}$

2.将偏置项转化为$x_0 = 1$，则前向传递函数为：$z = \sigma(w^{(1)}x))$；将偏置项转化为$z_0 = 1$，则前向传递函数为：$y = \sigma(w^{(2)}z))$

3.第一层网络参数的导数：

4.前向传播的计算复杂度为：$O(w)$；反向传播的计算复杂度为：$O(w)$。数值差分由于要在每个参数上执行两次前向传播，故计算复杂度为：$O(w^2)$。其中w为网络中所有参数的数目。

5.随机梯度下降过程：

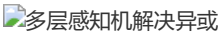
- 随机初始化网络参数w。
- 对于更新到第τ步,即将更新到$\tau + 1$步的情况,从所有的训练样本中随机挑选一个样本x_n,计算损失函数的值$E_n(w^{(\tau)})$及其梯度$\nabla E_n(w^{(\tau)})$
- 采用此式更新网络参数：$w^{(\tau+1)} = w^{(\tau)} - \alpha\nabla E_n(w^{(\tau)})$,其中$\alpha$为学习率。
- 令$\tau = \tau + 1$,判断此时是否已经满足更新次数，为满足转第二步，满足则结束。
  学习率对网络训练的影响：

- 学习率设置太大会造成网络不能收敛。
- 如果学习率设置太小，网络收敛非常缓慢，会增大找到最优值的时间，但是这很可能会进入局部极值点就收敛，没有真正找到的最优解。

6.不能，线性函数无论叠加多少层，都是线性的，只是斜率和截距不同，叠加网络对解决实际问题没有多大帮助；因为需要神经网络解决的实际问题基本都是非线性的。使用线性函数的话,加深神经网络的层数就没有意义了。

7.使用神经网络的动机：
在一些分类问题中，需要的参数可能会非常多，而且这些参数可能会以高次形式组合，变成更多的参数项。比如图片中每一个像素点（pixel）的表示，需要像素点坐标、灰度/RGB来表示。所以可以通过神经网络（Neural network）来解决复杂分类问题。同时对于输入特征较多的情况下，普通模型不能有效地处理这些特征。

8.能解决异或问题的多层感知机：


多层感知机解决异或
9.感知机的目标函数：
给定数据集 $T=\{(x_1,y_1),(x_2,y_2),...,(x_n,y_n)\}$
其中$x_i \in \mathcal{X}=\mathbf{R}^n, y_i \in \mathcal{Y}=\{-1,1\}, i=1,2,...,N$,求参数$w,b$，使得以下目标函数最小的解：

$$min_{w,b}L(w,b)=-\sum_{x_i \in M} y_i(w \cdot x_i+b)$$

其中M为错误分类点的集合。

注：通常来说，感知机的目标函数是分错样本的个数，但由于该函数不为连续可导函数，故用上式替代。

假设误分类点集合M是固定的，那么损失函数$L(w,b)$的梯度由下式给出：

$$\nabla_w L(w,b)=-\sum_{x_i \in M} y_i x_i$$
$$\nabla_b L(w,b)=-\sum_{x_i \in M} y_i$$

若使用随机梯度下降的方法优化的话，那么随机选取一个误分类点(x_i,y_i)，对w,b进行更新：

$$w \rightarrow w+\alpha y_i x_i$$
$$b \rightarrow b+\alpha y_i$$

其中α是学习率。

综上写出如下的算法过程：
输入：训练数据集$T=\{(x_1,y_1),(x_2,y_2),...,(x_n,y_n)\}$,其中$x_i\in\mathcal{X}=\mathbf{R}^n,y_i\in\mathcal{Y}=\{-1,1\},i=1,2,...,N$,$\alpha$为学习率$(0<\alpha\leq1)$
输出：w,b；感知机模型$f(x)=sign(w\cdot x+b)$
(1) 选取初值w_0,b_0
(2) 在训练集中选取数据(x_i,y_i)
(3) 如果$y_i(w\cdot x_i+b)\leq0$,

$$w\rightarrow w+\alpha y_ix_i$$
$$b\rightarrow b+\alpha y_i$$

(4) 转至(2)，直至训练集中没有误分类点。

六

1. 将多分类问题“拆解”为二分类问题的方案及优缺点：
   三种比较经典的方案：

- one-versus-the-rest
- one-versus-one
- many-versus-many
  首先是第一种方法：这种方法构建$|C|$个独立的分类器，其中第k个分类器以C_k类别作为正例，其余类别的样本作为负例。取置信度最大的类别作为预测结果。这种方法优点在于：存储开销和测试开销都相对较小。缺点在于：不同的分类器是在不同的任务上进行训练的，无法保证不同分类器产生的实数值$y_k(x)$具有恰当的标度。另一个问题是训练集合不平衡。
  其次是第二种方法：给定数据集$D=\{(x_1,y_1),(x_2,y_2),...,(x_m,y_m)\},y_i\in\{C_1,C_2,...,C_N\}$,给这N个类别两两配对，产生$\frac{N(N-1)}{2}$个二分类任务，进而得到与之数目相当的分类器，最终结果由投票产生，即预测类别最多的取为最终结果。优点在于在类别很多时，训练时间相对较短，缺点是存储开销和测试开销都很大。预测结果可能导致歧义性。
  最后是第三种方法：每次将若干类作为正例，若干类作为反例。此时正反类需要特殊的构造，不能够随意选取。优点在于这种方法对于错误以及各个分类器的输出的歧义性具有鲁棒性，缺点就是需要比较复杂的编码解码设计，比较繁琐，而且实际上作为上述两种方法的特例，如果设计不好的话会导致同时出现上述两种方法的缺点。

七

1. 贝叶斯定理的内容就是如下的公式：

$$P(\omega|\mathcal{D}) = \frac{P(\mathcal{D}|\omega)P(\omega)}{P(\mathcal{D})}$$

其中ω为我们选择的模型的参数，$\mathcal{D}$为我们所观测到的数据集，$P(\omega)$称为先验概率，$P(\omega|D)$称为后验概率。$P(\mathcal{D}|\omega)$称为似然。我们可以用后验概率分布和似然函数来表达贝叶斯定理的分母$p(\mathcal{D}) = \int p(\mathcal{D}|\omega)p(\omega)d\omega$。

其意义在于：
在观察到数据之前，我们有一些关于参数w的假设，这以先验概率$p(w)$的形式给出。观测数据$\mathcal{D} = \{t_1, ..., t_N\}$的效果可以通过条件概率$p(\mathcal{D}|w)$表达。
它让我们能够计算得到后验概率$p(\omega|\mathcal{D})$，即在观测到$\mathcal{D}$之后估计ω的不确定性。

2. 1 维与多维高斯分布的概率密度函数：
   一维：$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^2}\exp\{\frac{1}{2\sigma^2}(x-\mu)^2\}$
   μ被称为均值，σ^2称为方差。
   多维：$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{D}{2}}}\frac{1}{|\Sigma|^{\frac{1}{2}}}\exp\{\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\}$
   x为输入数据，为D维，μ为均值向量，Σ为协方差矩阵。
3.

$$P\left(x_0 \mid \omega_1\right) = P\left(x_0 \mid \omega_2\right) \Rightarrow \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\frac{(x_0-\mu_1)^2}{\sigma_2}} = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\frac{(x_0-\mu_2)^2}{\sigma^2}} \Rightarrow x_0 = \frac{\mu_1+\mu_2}{2}$$

则最小错误率：

$$\begin{aligned}
p(e) &= p\left(x \in R_2, c_1\right) + p\left(x \in R_1, c_2\right) && (1)\\
&= p\left(x \in R_2 \mid c_1\right) p\left(c_1\right) + p\left(x \in R_1 \mid c_2\right) p\left(c_2\right) && (2)\\
&= \int_{R_2} p\left(x \mid c_1\right) p\left(c_1\right) dx + \int_{R_1} p\left(x \mid c_2\right) p\left(c_2\right) dx && (3)\\
&= \frac{1}{2}\int_{\frac{\mu_1+\mu+\mu_2}{2}}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma}\right)^2} p\left(c_1\right) dx + \frac{1}{2}\int_{-\infty}^{\frac{\mu_1+\mu_2}{2}} \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\left(\frac{x-\mu_2}{\sigma}\right)^2} p\left(c_2\right) dx && (4)
\end{aligned}$$

若$p(c_1) = \lambda, p(w_2) = 1-\lambda$，则：

$$P(e) = \lambda\int_{\frac{\mu+\mu}{2}}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma}\right)^2} dx + (1-\lambda)\int_{-\infty}^{\frac{\mu_1+\mu_2}{2}} \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\left(\frac{x-\mu_2}{\sigma}\right)^2} dx$$

八

1.该结果的似然函数及其对数：

$$p(D|\mu)=\prod_{i=1}^{N}p(x_i|\mu)=\prod_{i=1}^{N}\mu^{x_i}(1-\mu_i)^{1-x_i}=\mu^{\sum_{i=1}^{N}x_i}(1-\mu)^{N-\sum_{i=1}^{N}x_i}$$

$$log\ p(D|\mu)=(\sum_{i=1}^{N}x_i)log\ \mu+(N-\sum_{i=1}^{N}x_i)log\ (1-\mu)$$

令$\frac{d\ log\ p(D|\mu)}{d\mu}=0$：

$$\frac{(\sum_{i=1}^{N}x_i)}{\mu}+\frac{(N-\sum_{i=1}^{N}x_i)}{1-\mu}=0$$

得μ的极大似然估计为：

$$\mu_{ML}=\frac{\sum_{i=1}^{N}x_i}{N}$$

2.在给定数据集大小为N的前提下，正面朝上的观测总数为m。则：

$$p(y|N,\mu)=C_N^m\mu^m(1-\mu)^{N-m}$$

其中y表示给定数据集大小为N的前提下，正面朝上的观测总数为m的整个事件。

$$p(\mu|a,b)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$$

故有联合概率密度：

$$\begin{gathered}p(y,\mu|N,a,b)=p(y|N,\mu)p(\mu|a,b)\\=C_N^m\mu^m(1-\mu)^{N-m}\ \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}\\=C_N^m\ \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\ \mu^{m+a-1}(1-\mu)^{N-m+b-1}\\=C_N^m\frac{1}{B(a,b)}\mu^{m+a-1}(1-\mu)^{N-m+b-1}\\=\frac{B(a+m,b+N-m)}{B(a,b)}C_N^m\frac{1}{B(a+m,b+N-m)}\mu^{m+a-1}(1-\mu)^{N-m+b-1}\\=h(y)g(\mu,y)\end{gathered}$$

其中：

$$h(y)=\frac{B(a+m,b+N-m)}{B(a,b)}C_N^m$$
$$g(\mu,y)=\frac{1}{B(a+m,b+N-m)}\mu^{m+a-1}(1-\mu)^{N-m+b-1}$$

可以看出$g(\mu,y)$是形状参数为$a+m,b+N-m$的Beta分布。

实际上，μ的后验概率分布即为$g(\mu,y)$。（证明略，见【统计学进阶知识（一）】深入理解Beta分布：从定义到公式推导 - 知乎 (zhihu.com)）

即：

$$\begin{aligned}p(\mu\mid y,a,b)&=g(\mu,y)=\mathrm{Beta}(a+m,b+N-m)\\&=\frac{1}{B(a+m,b+N-m)}\mu^{m+a-1}(1-\mu)^{N-m+b-1}\\&=p(\mu|D)\end{aligned}$$

有：

$$\begin{aligned}p(x=1|D)&=\int_0^1 p(x=1|\mu)p(\mu|D)d\mu\\&=\int_0^1 \mu p(\mu|D)d\mu\\&=E(\mu|D)\end{aligned}$$

由于Beta分布的期望：

$$
\begin{aligned}
E[X] &= \int_0^1 x f(x;\alpha,\beta) \\
&= \int_0^1 x \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \mathbf{d}x \\
&= \frac{1}{B(\alpha,\beta)} \int_0^1 x^{\alpha}(1-x)^{\beta-1} \mathbf{d}x \\
&= \frac{B(\alpha+1,\beta)}{B(\alpha,\beta)} \\
&= \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \\
&= \frac{\alpha}{\alpha+\beta}
\end{aligned}
$$

故:

$$
\begin{aligned}
p(x=1|D) &= \int_0^1 p(x=1|\mu)p(\mu|D)d\mu \\
&= \int_0^1 \mu p(\mu|D)d\mu \\
&= E(\mu|D) \\
&= \frac{a+m}{a+b+N}
\end{aligned}
$$

九

1.极大似然估计:

$$
\begin{aligned}
p(D \mid \mu) = \prod_{i=1}^{N} p\left(x_i|\mu,\sigma\right) &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(x_i-\mu\right)^2}{2\sigma^2}\right) \\
&= (\frac{1}{\sqrt{2\pi}\sigma})^N \exp\left(\sum_{i=1}^{N} -\frac{\left(x_i-\mu\right)^2}{2\sigma^2}\right) \\
\log p(D|\mu) &= -\frac{N}{2}\log 2\pi - N\log\sigma - \sum_{i=1}^{N}\frac{\left(x_i-\mu\right)^2}{2\sigma^2} \\
\frac{d\log P(D \mid \mu)}{d\mu} &= \sum_{i=1}^{N}\frac{\mu-x_i}{\sigma^2}
\end{aligned}
$$

$$\frac{d\log p(D|\mu)}{d\mu}=0 \quad \Rightarrow \mu_{mL}=\frac{1}{N}\sum_{i=1}^{N}X_i$$

2.若考虑一个新的观测点X_{N+1}，则：

$$\mu_{ML}^{(N+1)}=\frac{1}{N+1}\sum_{i=1}^{N+1}X_i=\frac{1}{N+1}\left(N\mu_{ML}^{(N)}+X_{N+1}\right)$$
$$=\frac{N}{N+1}\mu_{ML}^{(N)}+\frac{1}{N+1}X_{N+1}=\mu_{ML}^{(N)}+\frac{1}{N+1}\left(X_{N+1}-\mu_{ML}^{(N)}\right)$$

3.

注：如何理解「共轭分布」？ - 知乎 (zhihu.com)

4.可预测分布(predictive distribution)：

以下内容来自PRML：

在实际应用中，我们通常感兴趣的不是w本身的值，而是对于新的x值预测出t的值。这需要我们计算出预测分布(predictive distribution)，定义为：

$$p(t|\mathbf{t},a,B)=\int p(t|\mathbf{w},B)p(\mathbf{w}|\mathbf{t},a,B)dw$$

其中$\mathbf{t}$是训练数据的目标变量的值组成的向量。并且，为了简化记号，我们在右侧省略了条件概率中出现的输入向量。目标变量的条件概率分布$p(t|\mathbf{t},a,B)$由上式给出。预测分布的含义就是在机器学习的过程中，我们的最终目的还是需要根据训练集估计出来的参数w来预测测试集中的样本标签。在上式中给出了从纯粹的贝叶斯观点的角度出发来求取预测分布的过程（即需要对参数w的空间进行积分)。

我们看到上式涉及到两个高斯分布的卷积，因此我们看到预测分布的形式为：

$$p(t|\mathbf{x},\mathbf{t},\alpha,\beta)=N(t|\mathbf{m}_N^T\phi(x),\sigma_N^2(x))$$

其中预测分布的方差$\sigma_N^2(x)$为

$$\sigma_N^2(x)=\frac{1}{\beta}+\phi(x)^T\mathbf{S}_N\phi(x)$$

上式的第一项表示数据中的噪声，而第二项反映了与参数w关联的不确定性。由于噪声和w的分布是相互独立的高斯分布，因此它们的值是可以相加的。在极限$N\to\infty$的情况下，上式的第二项趋于零，从

而预测分布的方差只与参数β控制的具有可加性的噪声有关。

+

1. 支持向量机：边缘（间隔）（margin）被定义为决策边界与最近的数据点之间的垂直距离。最大化边缘会生成对决策边界的一个特定的选择。这个决策边界的位置由数据点的一个子集即最大化边缘后距离决策边界最近的若干数据点确定，这个数据点的子集被称为支持向量。
2. 分为两种，硬（hard）间隔与软（soft）间隔：
   边缘（间隔）（margin）被定义为决策边界与最近的数据点之间的垂直距离。
   使得所有样本都分类正确时最大化间隔称为硬间隔。
   而允许部分样本不满足约束条件时，此时的间隔称为软间隔。
   硬间隔SVM表达式：

$$\begin{aligned} &\underset{\omega,b}{min}\frac{1}{2}||\omega||^2 \\ &s.t. y_i(\omega^T x_i+b)\geq 1 \end{aligned}$$

软间隔SVM表达式：

$$\begin{aligned} &\underset{\omega,b}{min}\frac{1}{2}||\omega||^2+C\sum_{i=1}^{m}\xi_i \\ &s.t. y_i(\omega^T x_i+b)\geq 1-\xi_i \\ &\xi_i\geq 0 \quad i=1,2,...,m \end{aligned}$$

3. 支持向量机模型的对偶优化模型：

$$\begin{aligned} &\max_{\alpha}\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^T x_j \\ &s.t.\sum_{i=1}^{m}\alpha_i y_i=0 \\ &0\leq\alpha_i\leq C \quad i=1,2,...,m \end{aligned}$$

4. 以下摘抄自李航老师的统计学习方法一书中133页的核函数定义。
   核函数：设$\mathcal{X}$为输入空间（欧几里得空间$\mathbb{R}^n$的子集或离散集合），又设$\mathcal{H}$为特征空间（希尔伯特空间），如果存在一个$\mathcal{X}$到$\mathcal{H}$的映射：

$$\phi(x):\mathcal{X}\rightarrow\mathcal{H}$$

使得对所有$x,z\in\mathcal{X}$,函数$K(x,z)$满足条件：

$$K(x,z) =< \phi(x), \phi(z) >$$

则称K(x,z)为核函数 **(kernel function)**，$\phi(x)$为映射函数，式中$< \phi(x), \phi(z) >$为$\phi(x)$和$\phi(z)$的内积。

常见的核函数有：

线性核函数：$K(x_i, x_j) = x_i^T x_j$

多项式函数：$K(x_i, x_j) = (x_i^T x_j)^d, d \geq 1$为多项式的系数

高斯核：$K(x_i, x_j) = exp(-\frac{||x_i - x_j||^2}{2\sigma^2}), \sigma > 0$

拉普拉斯核：$K(x_i, x_j) = exp(-\frac{||x_i - x_j||}{\sigma}), \sigma > 0$

Sigmoid核：$K(x_i, x_j) = tanh(\beta x_i^T x_j + \theta)$,$tanh$为双曲正切函数，$\beta > 0, \theta < 0$

5. 首先证明其对称性：

$$K_{ij} = K\left(x^{(i)}, x^{(j)}\right) = \phi\left(x^{(i)}\right)^T \phi\left(x^{(j)}\right) = \phi\left(x^{(j)}\right)^T \phi\left(x^{(i)}\right) = K\left(x^{(j)}, x^{(i)}\right) = K_{ji}$$

在此再证明其半正定性：

$$\begin{aligned}
z^T K z &= \sum_i \sum_j z_i K_{ij} z_j \\
&= \sum_i \sum_j z_i \phi\left(x^{(i)}\right)^T \phi\left(x^{(j)}\right) z_j \\
&= \sum_i \sum_j z_i \sum_k \phi_k\left(x^{(i)}\right) \phi_k\left(x^{(j)}\right) z_j \\
&= \sum_k \sum_i \sum_j z_i \phi_k\left(x^{(i)}\right) \phi_k\left(x^{(j)}\right) z_j \\
&= \sum_k \left(\sum_i z_i \phi_k\left(x^{(i)}\right)\right)^2 \\
&\geq 0
\end{aligned}$$

6. 可以通过函数组合得到新的核函数：

- 若K_1和K_2为核函数，则对于任意正数γ_1, γ_2，其线性组合为$\gamma_1 K_1 + \gamma_2 K_2$也是核函数
- 若K_1和K_2为核函数，则核函数的直积$K_1 \otimes K_2(x,z) = K_1(x,z)K_2(x,z)$也是核函数
- 若K_1为核函数，则对于任意函数$g(x)$,$K_1(x,z) = g(x)K_1(x,z)g(z)$也是核函数

7. 利用核技巧推广逻辑回归模型：

需要说明的是此处不同于之前的logsitc回归采取的label标注的0和1，而是运用的label标注：-1和1.

（和SVM是相同的）。那么可以写出新的损失函数（依据最大似然估计）或者说优化形式如下：
以下是推广下的逻辑斯蒂回归形式$h(x) = \frac{1}{1+exp(-\omega^T \phi(x))}$

$$\min_{\omega} \frac{\lambda}{2N}\omega^T\omega + \frac{1}{N}\sum_{i=1}^{N} log(1 + exp(-y_i\omega^T\phi(x_i)))$$

下面可以运用核技巧推广以上模型：

Representer Theorem：对于任何的L2-正则化线性模型：

$$\min_{\omega} \frac{\lambda}{2N}\omega^T\omega + \frac{1}{N}\sum_{i=1}^{N} err(y_i, \omega^T z_i)$$

最佳的$\omega = \sum_{i=1}^{N} \beta_i z_i$ 其中$z_i = \phi(x_i)$。
（证明可参见机器学习技法）
下面以此作为基本事实进行核技巧在logistc模型上的运用。

由于ω已经满足上述形式，于是我们带入函数形式即可得到下列结果：

$$\min_{\beta} \frac{\lambda}{N}\sum_{n=1}^{N}\sum_{m=1}^{N}\beta_n\beta_m K(\mathbf{x}_n, \mathbf{x}_m) + \frac{1}{N}\sum_{n=1}^{N}\log\left(1 + \exp\left(-y_n\sum_{m=1}^{N}\beta_m K(\mathbf{x}_m, \mathbf{x}_n)\right)\right)$$

？？？？

8. 我唯一想到的联系就是拥有相似的数学表达。。。
   区别当然很大，一种是模型，一种是可以应用在任何L2正则化线性模型中的被人熟知的技巧。

十一

1.使用数据降维算法的目的：

- Curse of dimensionality
- To leverage the difficulty of a learning task

2.主成分分析与线性鉴别分析的区别：
1）LDA是有监督的降维方法，而PCA是无监督的降维方法

2）LDA降维最多降到类别数k-1的维数，而PCA没有这个限制。

3）LDA除了可以用于降维，还可以用于分类。

4）LDA选择分类性能最好的投影方向，而PCA选择样本点投影具有最大方差的方向。

3.最近重构性：样本点到这个超平面的距离都足够近。

x_i在欧几里得空间中的标准正交基为$\{w_1, ..., w_d\}$,w_i标准正交向量，$||w_2||_2 = 1$，$w_i^T w_j = 0(i \neq j)$。若丢弃若干坐标，降低到d'维，$d' < d$。x_i在低维空间投影为$z_i = (z_{i1}; ...; z_{id'})$，$z_{ij} = w_j^T x_i$，新的投影重构样本$\hat{x_i} = \sum_{j=1}^{d'} z_{ij} w_j$。考虑整个训练集，原样本点$x_i$与基于投影重构的样本点$\hat{x_i}$之间的距离为：

$$\sum_{i=1}^{m} ||\hat{x_i} - x_i||_2^2 = \sum_{i=1}^{m} || \sum_{j=1}^{d'} z_{ij} w_j - x_i||_2^2 = \sum_{i=1}^{m} z_i^T z_i - 2 \sum_{i=1}^{m} z_i^T W^T x_i + const$$
$$\propto -tr(W^T (\sum_{i=1}^{m} x_i x_i^T) W)$$

其中$w = \{w_1, ..., w_d\}$。

故可得：

$$\min_{\omega} -tr(W^T X X^T W)$$
$$s.t. W^T W = I$$

最大可分性：样本点在这个超平面上的投影都尽可能分开。

样本点x在新空间中超平面投影是$W^T x_i$，要使投影后的样本点尽快呢分开，就必须使投影后的样本点方差最大，写出协方差矩阵$\sum_i W^T x_i x_i^T W$，由此可得优化目标：

$$\min_{\omega} -tr(W^T X X^T W)$$
$$s.t. W^T W = I$$

求解步骤如下：

step1:$x_i \leftarrow x_i - \frac{1}{m} \sum x_i$，中心化样本点

step2:对协方差矩阵$X X^T$做特征值分解

step3:取最大的d'个特征值对应的特征向量$w_i, ..., w_{d'}$

output:投影矩阵$W^* = (w_1, ..., w_{d'})$

4.二分类LDA：对于$D = \{(x1, y1), (x2, y2), ..., ((xm, ym))\}$，其中任意样本$x_i$为n维向量，$y_i \in \{C_1, C_2, ..., C_k\}$

我们定义$N_j(j=1,2...k)$为第j类样本的个数，$X_j(j=1,2...k)$为第j类样本的集合，而$\mu_j(j=1,2...k)$为第j类样本的均值向量，定义$\Sigma_j(j=1,2...k)$为第j类样本的协方差矩阵。对于二分类问题：

$$\min_w -w^T S_b w$$
$$s.t. w^T S_w w = 1$$

由于$S_b w = \lambda w = \lambda(\mu_o - \mu_1)$：

得：

$$w = S^{-1}(\mu_0 - \mu_1)$$

多分类LDA：优化目标函数如下：

$$\max_W \frac{|W^T S_b W|}{|W^T S_w W|} = \frac{tr(W^T S_b W)}{tr(W^T S_w W)}$$

其中$S_b = \sum_{j=1}^k N_j(\mu_j - \mu)(\mu_j - \mu)^T$,$\mu$为所有样本均值向量。$S_w = \sum_{j=1}^k S_{w_j} = \sum_{j=1}^k \sum_{x \in X_j}(x - \mu_j)(x - \mu_j)^T$
求解方法是写出$S_w^{-1} S_b$的最大的d'个特征值向量，作为投影矩阵W。

5.常用的特征选择算法：

Exhaustive Search（穷举法）：评估所有的C_d^m个子集

Branch-n-Bound Search（分支定界法）：使用分支限界搜索方法；只需要枚举所有可能的特征子集中的一小部分来找到最优子集。

Best Individual Feature：将所有的特征排序，选择最好的m个特征。

Sequencial Forward Selection (SFS)：选择最佳的单个特征，然后一次添加一个特征，并与所选特征相结合，使criterion function最大化。

Sequencial Backward Selection (SBS)：从所有的d特性开始，并一次连续删除一个特性。同样使得criterion function最大化

“Plus l-take away r” Selection：首先使用前向选择通过l特征扩大特征子集，然后使用后向选择删除r特征，同样使得criterion function最大化

Sequencial Forward Floating Search (FSFS) and Sequencial Backward Floating Search (SBFS)：上一种方法的推广，l和r的值自动确定并且动态更新

6.Filter方法：根据独立于分类器的指标J来评价所选择的特征子集S，在所有可能的特征子集中搜索出使得可分性判据J最大的特征子集作为最优特征子集。不考虑所使用的学习算法。Wrapper方法：将特征选择和分类器结合在一起，在分类过程中表现优异的的特征子集会被选中。

7.流形学习：
设$Y \subset R^d$是一个低维流形，$f:Y \to R^D$是一个光滑嵌入，其中$D > d$。数据集{ y_i }为随机生成的，且经过f映射为观察空间的数据$\{x_i = f(y_i)\}$.流形学习就是在给定观察样本集$\{x_i = f(y_i)\}$。流形学习就是在给定观察样本集$\{x_i\}$的条件下重构f和$\{y_i\}$

LLE的主要思想是：对于一组具有嵌套流形的数据集，在嵌套空间与内在低维空间局部领域间的点的关系应该不变。即在嵌套空间每个采样点可以用它的近邻点线性表示，在低维空间中保持每个邻域中的权值不变，重构原数据点，是重构误差最小。

算法推导如下：
首先为每个样本x_i找到近邻下标集合Q_i,然后计算出基于Q_i中的样本点对x_i进行线性重构的系数w_i：

$$\min_{w_1,w_2,...,w_m} \sum_{i=1}^{m} ||x_i - \sum_{j \in Q_i} w_{ij} x_j||_2^2$$
$$s.t. \sum_{j \in Q_i} w_{ij} = 1$$

以上式子令$C_{jk} = (x_i - x_j)^T (x_i - x_k)$,$w_{ij}$有闭式解。

$$w_{ij} = \frac{\sum_{k \in Q_i} C_{jk}^{-1}}{\sum_{l,s \in Q_i} C_{ls}^{-1}}$$

之后LLE在低维空间中保持w_{ij}不变，于是x_i对应的低维空间坐标z_i可通过以下式子求解：

$$\min_{z_1,z_2,...,z_m} \sum_{i=1}^{m} ||z_i - \sum_{j \in Q_i} w_{ij} z_j||_2^2$$

令$Z = (z_1, z_2, ..., z_m) \in R^{d' \times m}$,$(W)_{ij} = w_{ij}$,$M = (I - W)^T (I - W)$
优化式可写为下述形式：

$$\min_Z tr(ZMZ^T)$$
$$s.t. ZZ^T = I$$

上述式子可以通过特征值分解进行求解：M的最小的前d'个特征值对应的特征向量矩阵即为Z^T

LLE算法流程：

输入：样本集$D=x_1,x_2,...x_m$;近邻参数k;低维空间维数d'
过程:
1.for $i=1,2,...,m$ do:

2.确定x_i的k近邻;

3.从上述推导中求得w_{ij},$j\in Q_i$

4.对于$j\notin Q_i$,令$w_{ij}=0$
5.end for
6.从上式求得M
7.对M进行特征值分解
8.return M的最小d'个特征值对应的特征向量。

十二

1. 聚类与分类有何区别和联系？什么情况下需要使用聚类算法?
   “分类”与“聚类”的区别在于：学习方法和对数据要求不同，“分类”是已知分类规则或分类标签，利用分类规则或分类标签构造分类器；而“聚类”不知道分类规则或分类标签，完全通过学习集数据样本找出分类规则，构造分类器，确定分类。
   联系在于两种方法都是为了发掘数据的类别信息而进行的学习算法。
   当给出的训练数据不存在标签信息，而任务要求我们发掘数据内部的类别信息时，需要使用聚类算法。
2. 给定数据集$D=\{x_1,x_2,...,x_m\}$,需要将数据划分为$\mathcal{C}=\{C_1,C_2,...,C_k\}$最小化下列平方误差，即该算法的目标函数：

$$E=\sum_{i=1}^{k}\sum_{x\in C_i}||x-\mu_i||_2^2$$

$\mu_i=\frac{1}{|C_i|}\sum_{x\in C_i}x$是簇的均值向量。
求解由于上述式子的最优解需要考察所有D中可能的簇划分，故是一个NP-hard问题。k-均值算法实际上采取了贪心的策略，故搜索到的划分不一定是最优的。
k-均值算法的结果受初始值的影响，该算法实际上是启发式的算法，不同的初始化往往求解出来不同的局部最优解。

3. 基于 DBSCAN 的概念定义，若 x 为核心对象，由 x 密度可达的所有样本构成的集合为 X。试证明 X 满足连接性与最大性。
   $X=\{x'\in D|x'由x密度可达\}$

首先依据定义：

连接性(connectivity):$x_i \in C, x_j \in C \Rightarrow x_i$和$x_j$密度相连

最大性(maximality):$x_i \in C, x_j$由x_i密度可达 $\Rightarrow x_j \in C$

先证明连接性

任取X中的两个元素:x_i, x_j按照定义都与x密度可达，故x_i和x_j密度相连。即满足连接性

再证明最大性：

按照定义，这近乎显然，采取反证法，假设X不满足最大性：那么，存在$x_i, x_j, x_i \in X$时，x_j由x_i密度可达,且$x_j \notin X$,这样的话，由X的定义可知，$x_j \in X$,产生了矛盾，故X满足最大性。

十三

1. 对于一个学习器，我们根据学习器预测的结果对所有样本进行排序，按照最有可能为正例的顺序到最不可能为正例的顺序排序，然后按照顺序依次取前1个，前2个，前3个，一直到前N个为预测正例（N为数据集总个数），依次统计confusion matrix，计算查全率（recall），查准率（precision），以precision为纵轴，recall为横轴作图，得到P-R曲线（题中的F1曲线应该为此意）。再通过confusion matrix计算真正例率（True Positive Rate，简称为TPR）和假正例率（False Positive Rate，简称FPR），按如上的方法计算，以真正例率为纵轴，假正例率为横轴，作图得到的是ROC曲线。
2. 将数据集D划分为k个大小相似的互斥子集，即$D = D_1 \cup D_2 ... \cup D_k, D_i \cap D_j = \emptyset (i \neq j)$,每个$D_i$都尽可能保持与原数据集分布的一致性，即从D中分层抽样得到，然后用其中依次用k-1个子集的并做训练集，剩余的那个做验证集，最后可以得到k组训练验证集，从而进行k次训练和测试，最终返回这k次测试结果的均值，这种方法称为k折交叉验证法。

   k最大为$|D|$，即数据集中数据的个数。
3. 留一法即上述k折交叉验证的基础上，取k=m时的方法称为留一法（m为D数据集中数据的个数）
4. 推导偏差-方差分解公式：

   令$\bar{f}(x) = E_D[f(x;D)], var = E_D[(f(x;D) - \bar{f}(x))^2], bias^2 = (\bar{f}(x) - y)^2, \varepsilon^2 = E_D[(y_D - y)^2]$并假定噪声期望为0，即$E_D[y_D - y] = 0$，有：

$$
\begin{aligned}
E(f;D) =& \mathbb{E}_D\left[\left(f(\boldsymbol{x};D)-y_D\right)^2\right] \\
=& \mathbb{E}_D\left[\left(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x})+\bar{f}(\boldsymbol{x})-y_D\right)^2\right] \\
=& \mathbb{E}_D\left[\left(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x})\right)^2\right]+\mathbb{E}_D\left[\left(\bar{f}(\boldsymbol{x})-y_D\right)^2\right] \\
&+\mathbb{E}_D\left[2\left(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x})\right)\left(\bar{f}(\boldsymbol{x})-y_D\right)\right] \\
=& \mathbb{E}_D\left[\left(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x})\right)^2\right]+\mathbb{E}_D\left[\left(\bar{f}(\boldsymbol{x})-y_D\right)^2\right] \\
=& \mathbb{E}_D\left[\left(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x})\right)^2\right]+\mathbb{E}_D\left[\left(\bar{f}(\boldsymbol{x})-y+y-y_D\right)^2\right] \\
=& \mathbb{E}_D\left[\left(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x})\right)^2\right]+\mathbb{E}_D\left[\left(\bar{f}(\boldsymbol{x})-y\right)^2\right]+\mathbb{E}_D\left[\left(y-y_D\right)^2\right] \\
&+2\mathbb{E}_D\left[\left(\bar{f}(\boldsymbol{x})-y\right)\left(y-y_D\right)\right] \\
=& \mathbb{E}_D\left[\left(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x})\right)^2\right]+\left(\bar{f}(\boldsymbol{x})-y\right)^2+\mathbb{E}_D\left[\left(y_D-y\right)^2\right]
\end{aligned}
$$

其中$E[X+Y]=E[X]+E[Y], E[XY]=E[X]E[Y]$。
注：

$$
\begin{aligned}
& E_D\left[2\left(f(x;D)-\bar{f}(x)\right)\left(\bar{f}(x)-y_D\right)\right] \\
=& E_D\left[2(f(x;D)-\bar{f}(x))\bar{f}(x)-2(f(x;D)-\bar{f}(x))y_D\right] \\
=& E_D\left[2\left(f(x;D)\bar{f}(x)-\bar{f}^2(x)\right)\right]-\left[2E_D\left(f\left(x;D\right)y_D\right)-2E_D\left(\bar{f}(x)y_D\right)\right] \\
=& 2\bar{f}(x)\cdot E_D[f(x;D)]-2\bar{f}^2(x)-\left[2\bar{f}(x)\cdot E_D\left(y_D\right)-2\bar{f}(x)\cdot E_D\left(y_D\right)\right] \\
=& 0-0=0
\end{aligned}
$$

5. 请分别针对偏差与方差的大小分布的不同情况，分析对应机器学习模型的改进策略。
偏差度量的是学习算法的期望预测和真实结果的偏离程度，即刻画了学习算法本身的拟合能力；方差度量了同样大小的训练集的变动所导致的学习性能的变化，即刻画了数据扰动所造成的影响；噪声是当前任务上任何学习算法能达到的期望泛化误差的下界，即刻画了学习问题本身的难度。
    - 如果模型的偏差较大时，此时偏差主导了泛化错误率，说明学习器的拟合能力不够，即训练不足，需要通过调整超参数增加学习器的训练程度。
    - 如果模型的方差较大时，此时方差主导了泛化错误率，学习器发生了过拟合的问题，训练数据发生的任何轻微扰动都有可能让学习器发生显著变化，需要通过调整超参数、降低模型复杂度、增加数据量或通过早停减少学习器的训练程度。