· 名词解释（5' * 4）

1. 机器学习/数据挖掘

    数据挖掘：是通过对(大规模)观测数据集的分析,寻找确信的关系,并将数据以一种可理解的且利于使用的新颖方式概括数据的方法.

    机器学习：如果说计算机程序可以从经验E中学习有关某类任务T和绩效指标P的信息，则该计算机程序是否可以通过经验E来提高在任务T中的绩效（由P衡量）

2. 主动学习/无监督学习/有监督学习/强化学习/半监督学习/在线学习/（课本P13）

    主动学习通过一定的算法查询最有用的未标记样本，并交由专家进行标记，然后用查询到的样本训练分类模型来提高模型的精确度。

    深度学习（英语：deep learning）是机器学习的分支，是一种以人工神经网络为架构，对资料进行表征学习的算法。

3. ID3（决策树算法。）（C4.5/CART算法）

    ID3算法（Iterative Dichotomiser 3 迭代二叉树3代）是一个由Ross Quinlan发明的用于决策树的算法。以信息增益为标准来选择划分属性。

    C4.5算法是由Ross Quinlan开发的用于产生决策树的算法。该算法是对Ross Quinlan之前开发的ID3算法的一个扩展。C4.5算法以增益率为标准来选择最有划分属性。C4.5算法产生的决策树可以被用作分类目的，因此该算法也可以用于统计分类。

4. 神经网络/支持向量机（VC维）/集成学习/K-means

    神经网络：（人工）神经网络是模仿大脑学习过程的计算模型，它们具有神经元的基本特征及其在大脑中的相互连接，通常情况下，计算机会编程来模拟这些特征。

    VC维：VC维被定义为算法可以破碎（shatter）的最大点集的基数，在这里破碎（shatter）意为若对于一个假设空间H，如果存在m个数据样本能够被假设空间H中的函数按所有可能的2^h种形式分开，则称假设空间H能够把m个数据样本破碎（shatter）

    集成学习：

    K-means：k均值聚类算法（k-means clustering algorithm）是一种迭代求解的聚类分析算法

神经网络的特点：大规模并行处理、结实、自适应和组织、足以模拟非线性关系、硬件

激活函数：

批量学习：在批处理学习中，在呈现所有N个训练样本之后，对多层感知器的突触权重进行调整。一次代表所有N个样本的训练过程称为训练的一个时期。因此，批处理学习的成本函数由平均误差能量Eav定义。

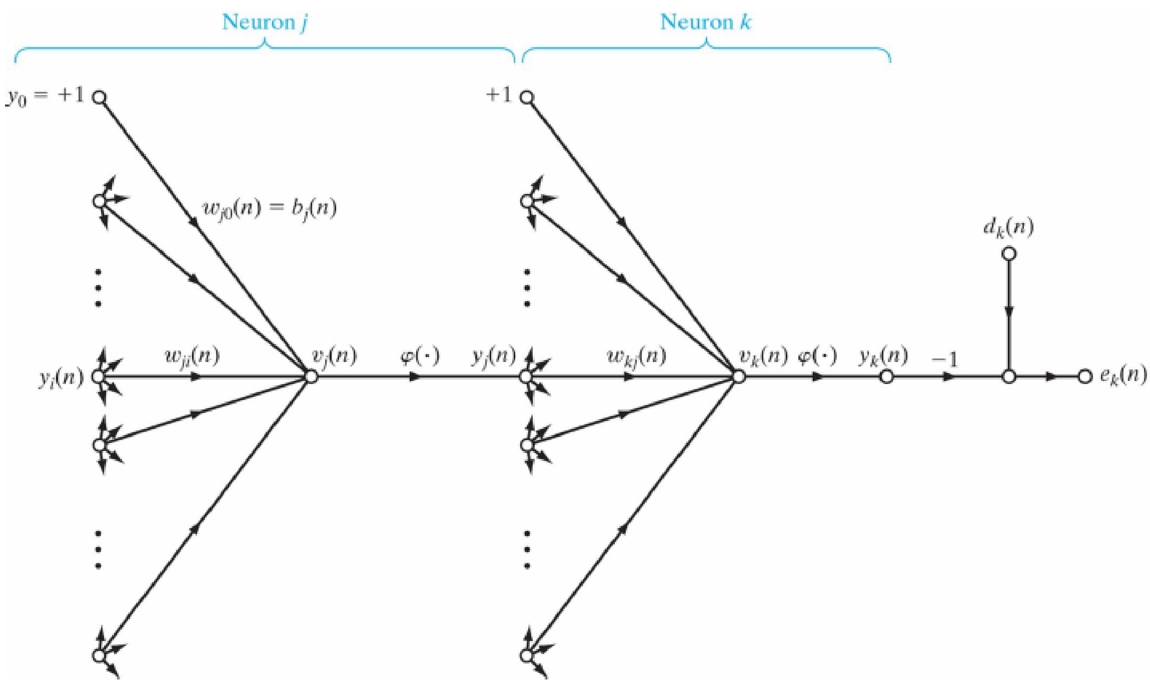
优点

准确估计梯度向量

并行化，速度快

坏处

更多存储要求



简答题（10' * 3）

1. parzen窗简述。为什么可以选用高斯密度函数作为窗函数？

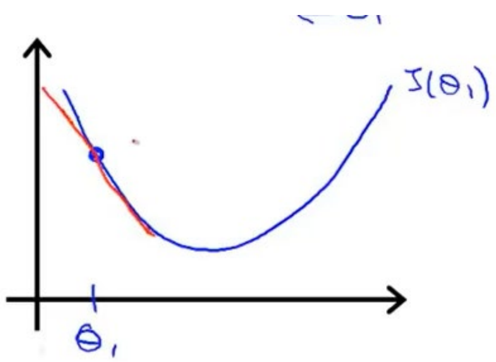
   P135

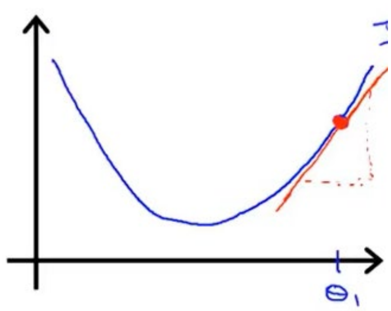
   窗函数需要满足的条件：

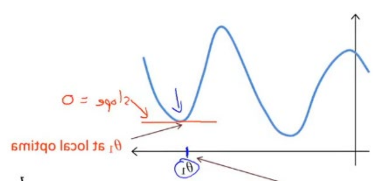
$$\varphi(x) \geq 0$$
$$\int \varphi(u)du = 1$$

2. 梯度下降算法与牛顿法的基本思想和区别。证明为什么梯度下降算法可以保证目标函数下降

P185







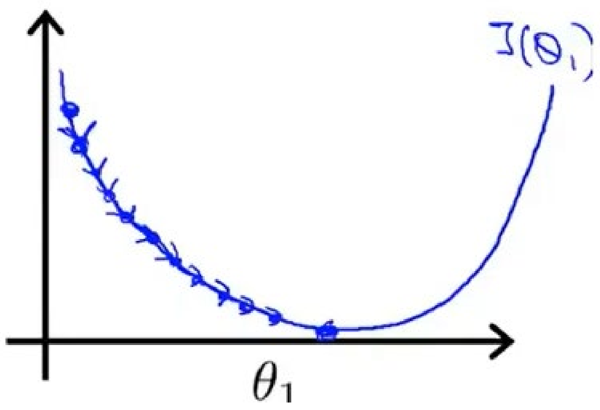
当梯度>0时，

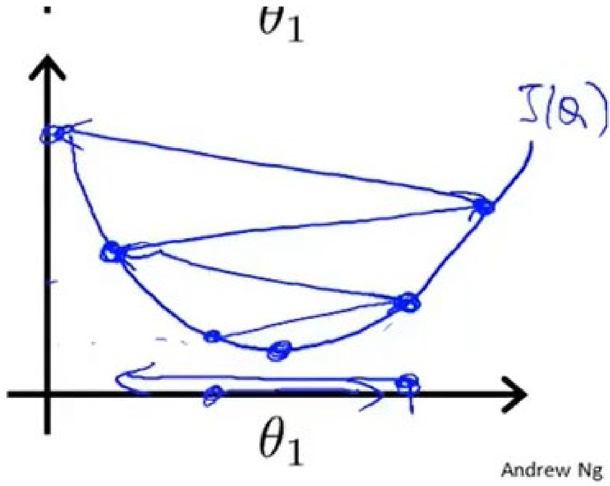
$a_1 = a_0 - \propto$(a变小了接近最小值点)

当梯度<0

$a_1 = a_0 - \propto$(a变大了接近最小值点)

当达到局部最优解时：梯度为零，收敛。





学习率太小时：收敛的过于缓慢。

学习率太大时：可能无法收敛甚至发散。

3. 什么是过拟合？模型为什么会出现过拟合？如何避免过拟合？

过拟合是训练误差在统计上小于测试误差。模型在训练集上表现很好，但在测试集上却表现很差。模型对训练集"死记硬背"，没有理解数据背后的规律，泛化能力差。

1、训练数据集样本单一，样本不足。

2、训练数据中噪声干扰过大。

3、模型过于复杂。

4.对模型进行了过度训练

1. 获取和使用更多的数据（数据集增强）——解决过拟合的根本性方法

2. 采用合适的模型（控制模型的复杂度）

3. 降低特征的数量：对于一些特征工程而言，可以降低特征的数量——删除冗余特征，人工选择保留哪些特征。

4. L1 / L2 正则化

5. Dropout：Dropout 指的是在训练过程中每次按一定的概率（比如50%）随机地“删除”一部分隐藏单元（神经元）。

6. Early stopping（提前终止）：在模型对训练数据集迭代收敛之前停止迭代来防止过拟合。

7.决策树中可以使用剪枝技术防止过拟合。

. 综合分析题

1. 从期望损失角度解释adaboost，如分布和分类器权重更新的依据。（20'）

   见课本二P173（从式8.4写到式8.19）

2. SVM。（1）从VC维和结构风险角度分析为什么margin要最大化。（2）推导优化函数的对偶形式。（3）简述SVM线性不可分的情况下如何求解(30')

（1）
VC维是指被定义为算法可以破碎（shatter）的最大点集的基数，在这里破碎（shatter）意为若对于一个假设空间H，如果存在m个数据样本能够被假设空间H中的函数按所有可能的2^m种形式分开，则称假设空间H能够把m个数据样本破碎（shatter）.

因为风险=经验风险+结构风险（学习模型结构带来的风险）
期望风险≤经验风险+

Expected Risk $R(\boldsymbol{\alpha})=\int \frac{1}{2}\left|y-f(\mathbf{x},\boldsymbol{\alpha})\right|dP(\mathbf{x},y)$

Empirical Risk $R_{emp}(\boldsymbol{\alpha})=\frac{1}{2l}\sum_{i=1}^{l}\left|y_i-f(\mathbf{x}_i,\boldsymbol{\alpha})\right|$

$$P\left(R(\alpha)\le R_{emp}(\alpha)+\underbrace{\sqrt{\frac{h(\log(2l/h)+1)-\log(\eta/4)}{l}}}_{\text{VC Confidence}}\right)=1-\eta$$

h is the VC dimension; l is the number of samples

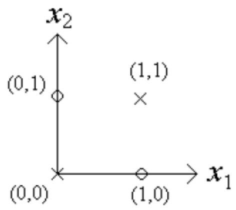
h是VC维，l是样本数，最后一项是学习率。
因此我们的目标是让h尽可能的小
令x属于半径为R的球面，分隔超平面的margin集的VC维大小h存在如下关系：

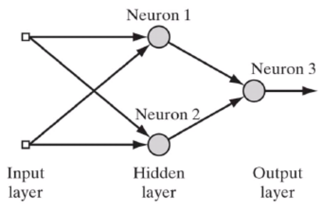
$$h\le min\left(\left(\frac{R}{\gamma}\right)^2,d\right)+1$$

其中γ是margin的大小，d是x的维数，为了让h尽可能的小，就要让$(\frac{R}{\gamma})$尽可能的小，就是让γ尽可能大，所以margin最大时泛化能力最强。

（2）课本二P121
公式6.1~公式6.11
（3）课本二P126
公式6.19~6.24

异或神经网络计算过程





(a)

- The weights are initialized as:

  $\underline{w}_1(0)=(-1.2,1,1)^T$, $\underline{w}_2(0)=(0.3,1,1)^T$, $\underline{w}_3(0)=(0.5,0.4,0.8)^T$

- η=0. 5
- When the sample (1,1) is given to the network:

$$\begin{cases} y_1=\dfrac{1}{1+\exp[-((-1.2)\times(-1)+1\times1+1\times1)]}=0.96 \\ y_2=\dfrac{1}{1+\exp[-(0.3\times(-1)+1\times1+1\times1)]}=0.84 \\ z=\dfrac{1}{1+\exp[-(0.5\times(-1)+0.4\times0.96+0.8\times0.84)]}=0.63 \end{cases}$$

We have:

$$\delta_j(n)=e_j(n)\varphi'(v_j(n))$$
$$=(d_j(n)-O_j(n))O_j(n)(1-O_j(n))$$

$$\delta_3=(0-0.63)\times0.63\times(1-0.63)=-0.147$$

$$\delta_j(n)=\varphi'(v_j(n))\sum_k\delta_k(n)w_{kj}(n)$$
$$=y_j(n)(1-y_j(n))\sum_k\delta_k(n)w_{kj}(n)$$

$$\delta_1=0.96\times(1-0.96)\times(-0.147)\times0.4=-0.002$$

$$\delta_2=0.84\times(1-0.84)\times(-0.147)\times0.8=-0.0158$$

- Then the weights are updated as:

$$\underline{w}_1(1)=(-1.2,1,1)^T+0.5\times(-0.0002)(-1,1,1)^T=(-1.199,0.999,0.999)^T$$
$$\underline{w}_2(1)=(0.3,1,1)^T+0.5\times(-0.0158)(-1,1,1)^T=(0.3079,0.992,0.992)^T$$
$$\underline{w}_3(1)=(0.5,0.4,0.8)^T+0.5\times(-0.147)(-1,0.96,0.84)^T=(0.5735,0.329,0.738)^T$$

- Finally, we can have:

$$\underline{w}_1(1)=(-1.198,0.912,1.179)^T$$
$$\underline{w}_2(1)=(0.294,0.826,0.98-)^T$$
$$\underline{w}_3(1)=(0.216,0.384,-0.189)^T$$