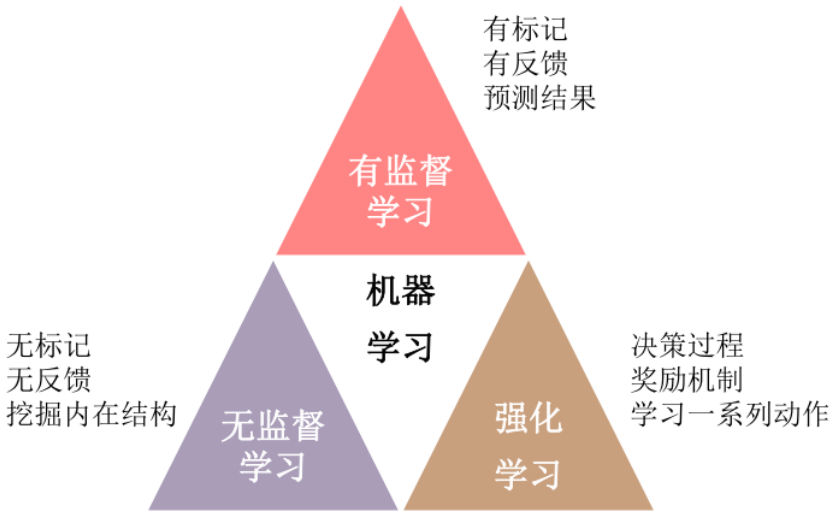
第一章 机器学习的基本概念

重点一 基本术语

- **数据**

	编号	色泽	根蒂	敲声	好瓜
		特征			标记
训练集	1	青绿	蜷缩	浊响	是
	2	乌黑	蜷缩	沉闷	是
	3	青绿	硬挺	清脆	否
	4	乌黑	稍蜷	沉闷	否
测试集	1	青绿	蜷缩	沉闷	?

- 学习



- **泛化能力**：模型适用于新样本的能 力为泛化(generalization)能力
- 模型偏好：学习算法自身在一个可能很庞大的假设空间中对假设进行选择的启发式或"价值观"

  如 奥卡姆剃刀——若有多个假设与观察一致，选最简单的那个
- 误差与损失函数：机器学习模型的输出结果与其对应的真实值之间的差异被称为模型的输出误差，损失函数用于度量模型对于单个样本的输出误差
- 泛化误差：机器学习模型在样本集合D上的整体误差称为该模型关于该学习任务的泛化误差
- **过拟合与欠拟合**

  过拟合：同时拟合训练样本的共性特征和个性特征（噪声）

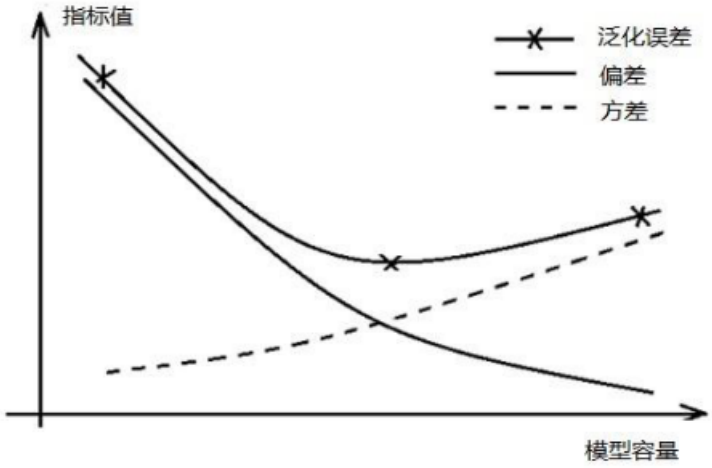
  欠拟合：未能充分拟合训练样本共性特征造成模型泛化误差较大而导致模型泛化能力较弱
- **偏差与方差（重点）**

$$R_exp(f) = E[L(Y, f(X))] = E[(f(x) - \hat{y})^2] = var[f(x)] + [bias[f(x)]^2] + \varepsilon^2$$

`var[f(x)]`：表达了同样大小训练集的变动导致的学习性能变化，刻画了数据扰动带来的影响

`bias[f(x)]^2`：表达了期望预测与真实值的偏离程度，刻画了学习算法对训练集的拟合能力

`ε^2`：表达了期望泛化误差下界， 刻画了学习问题的难度



【笔记】：

第二章 感知机模型

基本概念

- 处理二分类问题，输入为实例的特征向量，输出为实例的类别，取+1和-1。
- 利用梯度下降法对损失函数进行极小化
- 具有简单而易于实现的优点，但是也有如下缺点：
  - 感知机算法存在许多解，既依赖于初值，也依赖迭代过程中误分类点的选择顺序
  - 线性不可分数据集，迭代震荡

学习算法

- 梯度下降法

  影响因素：特征缩放（数据归一化），学习率

  特征缩放保证特征具有相近的尺度（无量纲化），可以使梯度下降法更快的收敛
- 优化目标

$$\min_{\mathbf{w},b} L(\mathbf{w},b) = \min_{\mathbf{w},b} -\sum_{\mathbf{x}_i \in M} y_i(\mathbf{w}^T\mathbf{x}_i + b)$$

- SGD
  - 首先任意选择一个超平面，即初始化w, b，然后不断极小化目标函数,损失函数L的梯度

$$\nabla_{\mathbf{w}} L(\mathbf{w},b) = -\sum_{\mathbf{x}_i \in M} y_i \mathbf{x}_i \qquad \nabla_b L(\mathbf{w},b) = -\sum_{\mathbf{x}_i \in M} y_i$$

  - 逐个选取误分类点更新

$$\mathbf{w} \leftarrow \mathbf{w} + \eta y_i \mathbf{x}_i \qquad b \leftarrow b + \eta y_i$$

练习

给定正例`x1=(3,3)`，`x2=(4,3)`，负例`x3=(1,1)`

请计算初值`w=0`、`b=0`、学习率为1情况下的分类超平面

第三章 支持向量机

基本概念

- 二分类模型
- 基本模型是定义在特征空间上的间隔最大的线性分类器， 间隔最大使它有别于感知机
- 支持向量机还包括核技巧，这使它成为实质上的非线性分类器
- 支持向量机的学习策略就是间隔最大化，可形式化为一 个求解凸二次规划(convex quadratic programming)的问题， 也等价于正则化的合页损失函数的最小化问题
- 支持向量机的学习算法是求解凸二次规划的最优化算法
- 引入软间隔

  解决问题：允许支持向量机在一些样本上不满足约束

  基本思想：最大化间隔的同时, 让不满足约束的样本应尽可能少
- 引入核技巧

  解决问题：非线性可分问题

  优点：本质虽然是将特征进行从低维到高维的转换，但核函数的本质是在它事先在低维上进行计算， 而将实质上的分类效果表现在了高维上，也就避免了直接在高维空间中的复杂计算

几何间隔

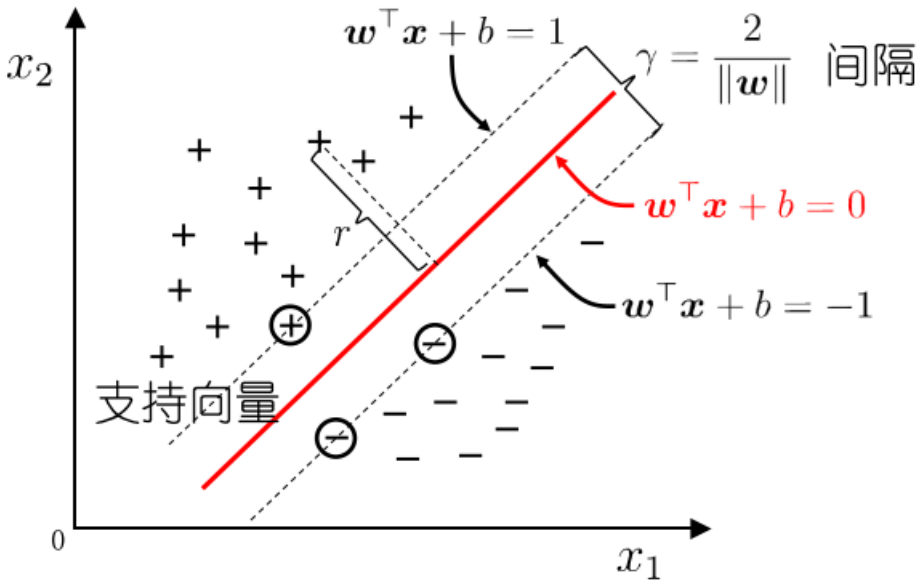
- 定义

$$\hat{d}_i = \frac{y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b)}{\|\boldsymbol{w}\|} = \frac{d_i}{\|\boldsymbol{w}\|}$$

- 优化问题

$$\max_{\boldsymbol{w},b} \hat{d} = \frac{d}{\|\boldsymbol{w}\|}, s.t.\, \hat{d}_i \geq \hat{d}$$

- 超平面方程



SMO算法

- 基本思路： 不断执行如下两个步骤直至收敛
  1. 选取一对需更新的变量 `αi` 和 `αj`
  2. 固定 `αi` 和 `αj` 以外的参数, 求解对偶问题更新 `αi` 和 `αj`
- 仅考虑 `αi` 和 `αj` 是，对偶问题约束变为

$$\alpha_i y_i + \alpha_j y_j = -\sum_{k \neq i,j} \alpha_k y_k, \quad \alpha_i \geq 0, \quad \alpha_j \geq 0.$$

练习

给定正例 `x1=(3,3)`， `x2=(4,3)`，负例 `x3=(1,1)`

计算 支持向量、分类决策函数 和 分类超平面方程

第四章 朴素贝叶斯分类

基本概念

- 贝叶斯决策就是在不完全情报下，对部分未知的状态用主观概率估计，然后用贝叶斯公式对发生概率进行修正，最后再利用期望值和修正概率做出最优决策，基本思想为：
  - 已知类条件概率密度参数表达式和先验概率
  - 利用贝叶斯公式转换成后验概率
  - 根据后验概率大小进行决策分类
- 判别式模型`(discriminative models)`：给定`x`，通过直接建模来预测，如决策树、`DNN`、`SVM`
- 生成式模型`(generative models)`：先对联合概率分布`P(x, c)`建模，再由此获得`P(c|x)`
- 朴素贝叶斯分类器：每个属性独立地对分类结果发生影响

  基于属性条件独立性假设

$$P(c \mid \mathbf{x}) = \frac{P(c)P(\mathbf{x} \mid c)}{P(\mathbf{x})} = \frac{P(c)}{P(\mathbf{x})} \prod_{i=1}^{d} P(x_i \mid c)$$

  其中`d`为属性数目，`xi`为`x`在第`i`个属性上的取值

  由于对所有类别来说`P(x)`相同，因此贝叶斯判定准则有

$$h_{nb}(\mathbf{x}) = \underset{c \in y}{\operatorname{argmax}} P(c) \prod_{i=1}^{d} P(x_i \mid c)$$

练习

1. 某抽奖游戏使用三个外观一致碗和三张抽奖券，其中两张1元券 和一张1000元券。游戏主持人分别用每个碗盖住一张券且不让抽奖者知道每个碗盖的是几元券，在抽奖者选定一个碗之后翻开剩下两个碗中的一个，使得翻开的碗盖的是1元券。抽奖者如何选择才能以较高的 概率获得1000元券?

2. 例子：用西瓜数据集3.0训练一个朴素贝叶斯分类器，对测试例“测1”进行分类

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
测 1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	?

第五章 决策树（重点）

基本概念

- 基本思想：模拟人类进行级联选择或决策的过程，按照属性的某个优先级依次对数据的全部属性进行判别，从而得到输入数据所对应的预测输出。
- 模型结构：一个根结点、若干内部结点和叶结点（叶结点表示决策的结果，内部结点表示对样本某一属性判别）
- 解决问题的步骤：
  - 模型构建（归纳）：通过对训练集合的归纳，建立分类模型
  - 预测应用（推论）：根据分类模型，对测试集合进行测试

学习算法

- 如何选择最优划分属性？

  如果结点对应数据子集中的样本基本属于同一个类别，则无需 对结点的数据子集做进一步划分，否则就要对该结点的数据子集做进一步划分，生成新的判别标准

- 如何在最优划分属性下划分样本集？

- 信息熵：度量样本集合纯度最常用的一种指标

  定当前样本集合D中第k类样本所占的比为$p_k(K=1,2,...,|\mathcal{Y}|)$
  则D的信息熵定义为

  $$\mathrm{Ent}(D)=-\sum_{k=1}^{|\mathcal{Y}|}p_k\log_2 p_k$$

  `Ent(D)`的值越小，则`D`的纯度越高

- ID3

  离散属性a有V个可能的取值$\{a^1,a^2,...,a^V\}$，用a来进行划分，则会产生V个分支结点，其中第v个分支结点包含了D中所有在属性a上取值为a^v的样本，记为D^v。则可计算出用属性a对样本集D进行划分所获得的“信息增益”：

  $$\mathrm{Gain}(D,a)=\mathrm{Ent}(D)-\sum_{v=1}^{V}\frac{|D^v|}{|D|}\mathrm{Ent}(D^v)$$

  为分支结点权重，样本数越多的分支结点的影响越大

  - 一般而言，信息增益越大，则意味着使用属性a来进行划分所获得的“纯度提升”越大
  - ID3决策树学习算法[Quinlan, 1986]以信息增益为准则来选择划分属性

  存在问题：信息增益对可取值数目较多的属性有所偏好

- C4.5

  增益率

$$\text{Gain\_ratio}(D,a) = \frac{\text{Gain}(D,a)}{\text{IV}(a)}$$

其中 $\text{IV}(a) = -\sum_{v=1}^{V} \frac{|D^v|}{|D|} \log_2 \frac{|D^v|}{|D|}$

存在问题：增益率准则对可取值数目较少的属性有所偏好

- CARD
  - 数据集 的纯度可用“基尼值”来度量

$$\text{Gini}(D) = \sum_{k=1}^{|\mathcal{Y}|} \sum_{k' \neq k} p_k p_{k'} = 1 - \sum_{k=1}^{|\mathcal{Y}|} p_k^2$$

D的基尼值越小，数据集D的纯度越高

  - 属性a的基尼指数定义为

$$\text{Gini\_index}(D,a) = \sum_{v=1}^{V} \frac{|D^v|}{|D|} \text{Gini}(D^v)$$

  - 应选择那个使划分后基尼指数最小的属性作为最优划分属性，即

$$a_* = \underset{a \in A}{\text{argmin}}\ \text{Gini\_index}(D,a)$$

  - CART [Breiman et al., 1984]采用“基尼指数”来选择划分属性

练习

1. 计算下图西瓜数据集中根结点的信息熵、色泽属性的信息的信息熵、以及色泽属性带来的信息增益。（ID3算法）

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否
10	青绿	硬挺	清脆	清晰	平坦	软粘	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	否

2. 使用基尼指数作为划分属性来构建上方西瓜数据集的决策树。（CARD算法）

第六章 集成学习

基本概念

- 集成学习将多个性能一般的普通模型进行有效集成，形成一个性能优良的集成模型
- 自助采样法

  假设D中包含有n个样本数据，自助采样对D进行n次有放回的随机采样并将采样样本纳入训练集

  对样本数据集D进行多次自助采样就可以分别生成多个具有一定差异的训练样本子集$D1$, $D2$,⋯, DK, 可分别通过对这些子集的训练构造出所需的弱学习器

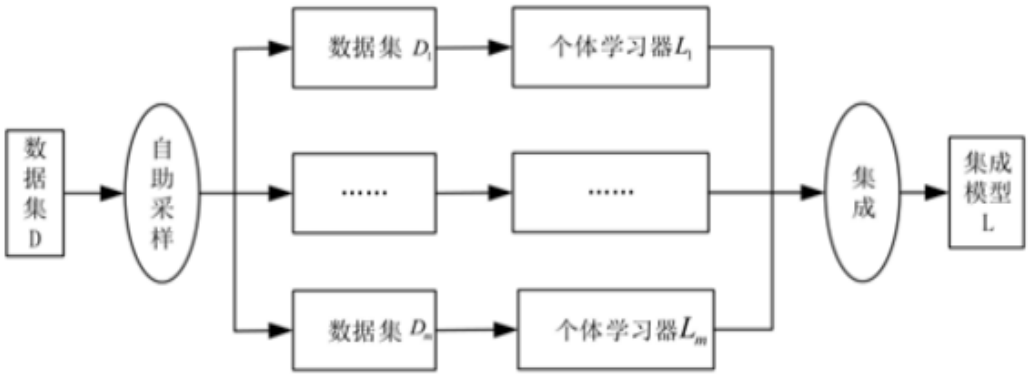


图　　Bagging 集成学习流程图

- 随机森林
  - 简单有效的常用监督学习模型
  - Bagging集成学习方法将多个决策树模型作为弱学习器集成起来，构建一个较强泛化性能的森林模型作为强学习器

- Boosting基本思想

Adaboost 学习算法（重点）

- 算法过程

（1）令$i=1$并设定弱学习器的数目m。使用均匀分布初始化训练样本集的权重分布，令n维向量w^i表示第i次需更新的样本权重，则有：$w^1=(w_{i1},w_{i2},\cdots,w_{in})^T=\left(\frac{1}{n},\frac{1}{n},\cdots,\frac{1}{n}\right)^T$

（2）使用权重分布为w^i的训练样本集D_i学习得到第i个弱学习器f_i

（3）计算f_i在训练样本集D_i上的分类错误率e_i ：

$$e_i=\sum_{k=1}^{n}w_{ik}I(f_i(X_k)\neq y_k)$$

（4）确定弱学习器f_i的组合权重α_i。由于弱学习器f_i的权重取值应与其分类性能相关，对于分类错误率e_i越小的f_i，则其权重α_i应该越大，故有$\alpha_i=\frac{1}{2}\ln\frac{1-e_i}{e_i}$

（5）依据弱学习器f_i对训练样本集D_i的分类错误率e_i更新样本权重，更新公式为$w_{i+1,j}=\frac{w_{ij}\exp(-\alpha_i y_k L_i(X_k))}{Z_i}$，其中$Z_i=\sum_{k=1}^{n}w_{ij}\exp(-\alpha_i y_k f_i(X_k))$为归一化因子，保证更新后权重向量为概率分布

（6）若$i<m$，则令$i=i+1$并返回步骤（2），否则执行步骤（7）

（7）对于m个弱分类器$f_1,f_2,\cdots,f_m$，分别将每个f_i按权重α_i进行组合：$G=sign(\sum_{i=1}^{m}\alpha_i f_i(X))$，得到并输出所求集成模型$G$，算法结束

【笔记】

练习

1. 试以所示数据集为训练样本，使用 `AdaBoost` 集成学习算法构建集成模型

编号	1	2	3	4	5	6	7	8	9	10
X	0	1	2	3	4	5	6	7	8	9
y	1	1	1	−1	−1	−1	1	1	1	−1

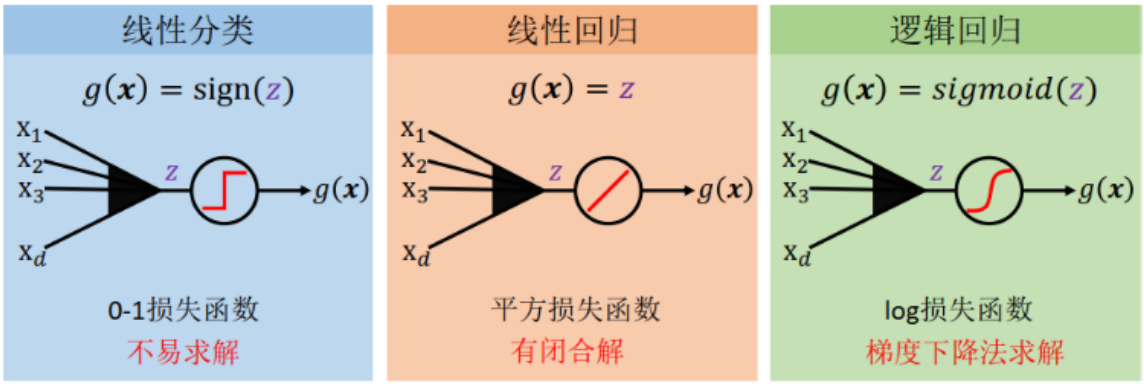
2. 某公司招聘职员考查身体、业务能力、发展潜力这3项。身 体分为合格1、不合格0两级，业务能力和发展潜力分为上1、 中2、下3三级。分类为合格1、不合格0两类。已知数据如 下表所示，以决策树桩为基学习器，试用Adaboost算法学习一个强分类器。

	1	2	3	4	5	6	7	8	9	10
身体	0	0	1	1	1	0	1	1	1	0
业务能力	1	3	2	1	2	1	1	1	3	2
发展潜力	3	1	2	3	3	2	2	1	1	1
分类	-1	-1	-1	-1	-1	-1	-1	1	1	-1

第七章 线性回归、逻辑回归与softmax回归

基本概念

- 均分误差最小化、最小二乘法
  - 给定训练样本$\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m$，使用最小二乘法，即基于均方误差最小化进行模型求解：
    $$\mathbf{w}^* = \operatorname{argmin} \sum_{i=1}^{m} (f(\boldsymbol{x}_i) - yi)^2 = \operatorname{argmin} \sum_{i=1}^{m} (\mathbf{w}^T \boldsymbol{x}_i - yi)^2$$
  - 则：
    $$J(\mathbf{w}) = \sum_{i=1}^{n} (\mathbf{w}^T \boldsymbol{x}_i - yi)^2 = ||\mathbf{X}\mathbf{w} - \boldsymbol{y}||^2$$
    其中$\mathbf{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m)^T = (x_{ij})_{m \times n}$， $\boldsymbol{y} = (y_1, y_2, \ldots, y_m)^T$
  - 令$\boldsymbol{J}(\mathbf{w})$对参数向量$\mathbf{w}$各分量的偏导数为0，即：
    $$\frac{\partial \boldsymbol{J}}{\partial \boldsymbol{w}} = \mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{w}) = 0$$
  - 则由$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{w}) = 0$解得：
    $$\mathbf{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$



z是关于特征$\boldsymbol{x}$的线性函数

正则化（深刻理解）

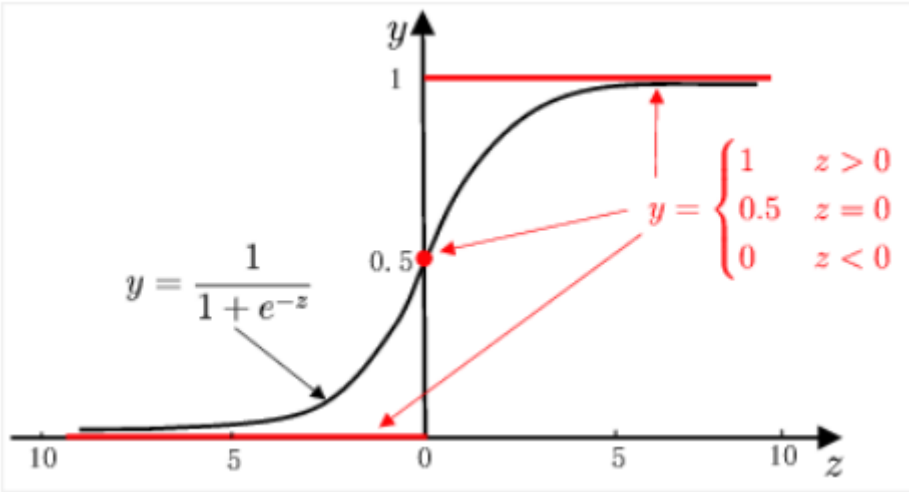
- 缓解过拟合的方法
- 正则化的**本质是对权重w的约束**，使模型简单，解决了特征过多，模型过于复杂的问题。

【笔记】

逻辑回归

- 替代函数——逻辑函数（logistic function）

$$\sigma = \frac{1}{1+e^{-z}}$$



- 对数损失（交叉熵损失）

$$P(y|\boldsymbol{x}) = \begin{cases} g(\boldsymbol{x}) & y=1 \\ 1-g(\boldsymbol{x}) & y=0 \end{cases} \Rightarrow$$

$$P(y|\boldsymbol{x}) = [g(\boldsymbol{x})]^{y}[1-g(\boldsymbol{x})]^{1-y}$$

$$l(y|h(x)) = -logP(y|x) = -ylogg(x) - (1-y)log(1-g(x))$$

- 经验风险

$$J(\mathbf{w},b)=R_{emp}(\mathbf{w},b)=-\frac{1}{m}\sum_{i=1}^{m}\left\{y^{(i)}\cdot\log g\left(\boldsymbol{x}^{(i)}\right)+\left(1-y^{(i)}\right)\cdot\log\left(1-g\left(\boldsymbol{x}^{(i)}\right)\right)\right\}$$

- 学习算法 - 梯度下降法

$$J(\mathbf{w},b)=-\frac{1}{m}\sum_{i=1}^{m}\left\{y^{(i)}\cdot\log g\left(\boldsymbol{x}^{(i)}\right)+\left(1-y^{(i)}\right)\cdot\log\left(1-g\left(\boldsymbol{x}^{(i)}\right)\right)\right\}$$

$$g(\boldsymbol{x})=\sigma\left(\mathbf{w}^{T}\boldsymbol{x}+b\right)=\frac{1}{1+\exp\left(-\mathbf{w}^{T}\boldsymbol{x}-b\right)}$$

$$\frac{\partial g}{\partial\mathbf{w}}=g(1-g)\boldsymbol{x}\quad\frac{\partial g}{\partial b}=g(1-g)$$

$$\frac{\partial\log g}{\partial\mathbf{w}}=(1-g)\boldsymbol{x}\quad\frac{\partial\log g}{\partial b}=1-g\qquad\frac{\partial J}{\partial\mathbf{w}}=\frac{1}{m}\sum_{i=1}^{m}\left(g\left(\boldsymbol{x}^{(i)}\right)-y^{(i)}\right)\boldsymbol{x}^{(i)}$$

$$\frac{\partial\log(1-g)}{\partial\mathbf{w}}=-g\boldsymbol{x}\quad\frac{\partial\log(1-g)}{\partial b}=-g\qquad\frac{\partial J}{\partial b}=\frac{1}{m}\sum_{i=1}^{m}\left(g\left(\boldsymbol{x}^{(i)}\right)-y^{(i)}\right)$$

1. 初始化w ,b
2. 更新w ,b（使用Batch GD，如果使用 SGD呢？）

$$\mathbf{w}:=\mathbf{w}-\eta\sum_{i=1}^{m}\left(g\left(\boldsymbol{x}^{(i)}\right)-y^{(i)}\right)\boldsymbol{x}^{(i)}$$

$$b:=b-\eta\sum_{i=1}^{m}\left(g\left(\boldsymbol{x}^{(i)}\right)-y^{(i)}\right)$$

Batch GD

3. 检查是否收敛，如果不收敛转2

【笔记】

Softmax回归

- 逻辑回归模型

$$\begin{cases}P(y=1\mid\boldsymbol{x})=\sigma\left(\mathbf{w}^{T}\boldsymbol{x}+b\right)=\dfrac{1}{1+e^{-\left(\mathbf{w}^{T}\boldsymbol{x}+b\right)}}\\[2ex]P(y=0\mid\boldsymbol{x})=1-\sigma\left(\mathbf{w}^{T}\boldsymbol{x}+b\right)=\dfrac{e^{-\left(\mathbf{w}^{T}\boldsymbol{x}+b\right)}}{1+e^{-\left(\mathbf{w}^{T}\boldsymbol{x}+b\right)}}\end{cases}$$

$$\frac{P(y=1\mid\boldsymbol{x})}{P(y=0\mid\boldsymbol{x})}=e^{\mathbf{w}^{T}\boldsymbol{x}+b}\Rightarrow logit=\log\frac{P(y=1\mid\boldsymbol{x})}{P(y=0\mid\boldsymbol{x})}=\mathbf{w}^{T}\boldsymbol{x}+b$$

- 扩展到k个类别

$$\log\frac{P(y=j\mid\boldsymbol{x})}{P(y=1\mid\boldsymbol{x})}=\mathbf{w}_j{}^T\boldsymbol{x}+b_j\Rightarrow P(y=j\mid\boldsymbol{x})=e^{\mathbf{w}_j{}^T\boldsymbol{x}+b_j}P(y=1\mid\boldsymbol{x}),\quad j=1..k$$

$$\Downarrow$$

$$P(y=1\mid\boldsymbol{x})=\frac{1}{1+\sum_{j=2}^{k}e^{\mathbf{w}_j{}^T\boldsymbol{x}+b_j}}\qquad P(y=j\mid\boldsymbol{x})=\frac{e^{\mathbf{w}_j{}^T\boldsymbol{x}+b_j}}{1+\sum_{c=2}^{k}e^{\mathbf{w}_c{}^T\boldsymbol{x}+b_c}}$$

$$\Downarrow$$

$$P(y=j\mid\boldsymbol{x})=\frac{e^{\mathbf{w}_j{}^T\boldsymbol{x}+b_j}}{\sum_{c=1}^{k}e^{\mathbf{w}_c{}^T\boldsymbol{x}+b_c}}$$

样本x属于第k类(y=k)的概率

$$P(y=k\mid\mathbf{x})=\frac{\exp(\boldsymbol{\theta}_k^T\boldsymbol{x})}{\sum_{j=1}^{C}\exp(\boldsymbol{\theta}_j^T\boldsymbol{x})}\in(0\quad 1),\quad \boldsymbol{\theta}=\{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\cdots\boldsymbol{\theta}_c\}$$

- 损失函数与经验风险
  - 对数损失

$$l(y,h(\boldsymbol{x}))=-\log P(y\mid\boldsymbol{x};\boldsymbol{\theta})=-\log\left\{\prod_{c=1}^{k}\left[P(y=k\mid\boldsymbol{x};\boldsymbol{\theta})\right]^{I(y=c)}\right\}$$

$$=\log\sum_{c=1}^{k}e^{(\mathbf{w}_c^T\boldsymbol{x}+b_c)}-\sum_{c=1}^{k}I(y=c)(\mathbf{w}_c^T\boldsymbol{x}+b_c)$$

  - 经验风险

$$J(\boldsymbol{\theta})=R_{emp}(\boldsymbol{\theta})=\frac{1}{m}\left\{\sum_{i=1}^{m}l(y_i,h(\boldsymbol{x}_i))\right\}$$

例题

1. 使用逻辑回归模型实现手写体图像数字0和1的识别 （图像大小均是28*28）

2. 使用Softmax回归模型实现手写体图像数字识别

3. 对于多分类问题，Softmax回归模型一定比逻辑回归模型 （分解成多个二分类任务）更合适吗?

第八章 神经网络与深度学习（重点）

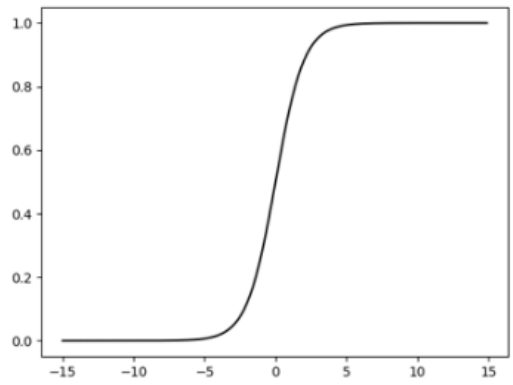
基本概念

- 深度学习与浅层学习的关系
  - 神经元数目足够多的神经网络模型可以逼近任意函数
  - 神经网络模型增加神经网络的隐含层层数比直接增加某一隐含层的结点数目更能提高模型的拟合能力
  - 浅层学习：数据处理层数较少的神经网络模型，其容量较低
  - 虽然从理论上讲浅层学习模型可以逼近任意函数，但其模型容量或灵活性远不及具有较深层次的网络模型，难以满足对复杂任务求解的需求
- 激活函数
  - Sigmoid

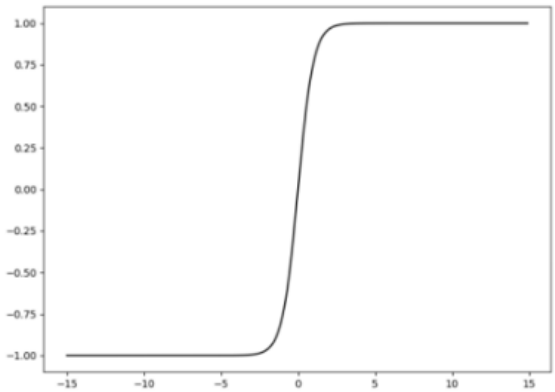
$$f(t)=\frac{1}{1+e^{-t}}$$
$$f(t)^{'}=f(t)*(1-f(t))$$

  - tanh

$$f(t)=tanh(t)=\frac{e^{t}-e^{-t}}{e^{t}+e^{-t}}$$
$$f(t)^{'}=1-f(t)^{2}$$



Sigmoid激活函数



tanh激活函数

多层感知机

BP算法

- 第一步：初始化W,b
- 第二步：
- Do{ 1.前向传播：

$$for\ l=1:L-1\quad z_j^{(l+1)}\left(x^{(n)}\right)=\sum_{i=1}^{N_l}w_{ij}^{(l)}a_i^{(l)}\left(x^{(n)}\right)+b_j^{(l)};\quad a_j^{(l+1)}\left(x^{(n)}\right)=f^{(l+1)}\left(z_j^{(l+1)}\left(x^{(n)}\right)\right)$$

2.误差后向传播：

$$l=L:\quad \delta_j^{(L)}\left(x^{(n)}\right)=\frac{\partial J\left(W,b;x^{(n)},y^{(n)}\right)}{\partial a_j^{(L)}}f^{(L)\prime}\left(z_i^{(L)}\left(x^{(n)}\right)\right)$$

$$for\ l=L-1:2\quad \left\{\begin{array}{c}\delta_i^{(l)}\left(x^{(n)}\right)=f^{(l)\prime}\left(z_i^{(L)}\left(x^{(n)}\right)\right)\left[\sum_{j=1}^{N_{l+1}}w_{ij}^{(l)}\delta_j^{(l+1)}\left(x^{(n)}\right)\right]\\ \frac{\partial J\left(W,b;x^{(n)},y^{(n)}\right)}{\partial b_j^{(l)}}=\delta_j^{(l+1)}\left(x^{(n)}\right);\frac{\partial J\left(W,b;x^{(n)},y^{(n)}\right)}{\partial w_{ij}^{(l)}}=\delta_j^{(l+1)}a_i^{(l)}\left(x^{(n)}\right)\end{array}\right\}$$

$$l=1:\quad \frac{\partial J\left(W,b;x^{(n)},y^{(n)}\right)}{\partial b_j^{(1)}}=\delta_j^{(2)}\left(x^{(n)}\right);\quad \frac{\partial J\left(W,b;x^{(n)},y^{(n)}\right)}{\partial w_{ij}^{(1)}}=\delta_j^{(2)}\left(x^{(n)}\right)x_i^{(n)}$$

3.更新权值：

$$w_{ij}^{(l)}:=w_{ij}^{(l)}-\alpha\frac{1}{m}\sum_{n=1}^{m}\frac{\partial J\left(W,b;x^{(n)},y^{(n)}\right)}{\partial w_{ij}^{(l)}};\quad b_j^{(l)}:=b_j^{(l)}-\alpha\frac{1}{m}\sum_{n=1}^{m}\frac{\partial J\left(W,b;x^{(n)},y^{(n)}\right)}{\partial b_j^{(l)}}$$

- }Until convergence

- 梯度消失
  - 原因：传统神经网络模型 多采用 `Sigmoid` 激活函数或 `tanh` 激活函数。由于这些激活函数的梯度任何情况下的取值均小于1，故梯度值将会在深度网络模型的训练过程中逐层衰减并最终接近于 0
  - 应对策略：
    - 改进激活函数，如ReLU、Leaky ReLU
    - 中间层引入损失函数，如GoogLeNet
    - 残差连接，如ResNet
- 局部最优
  - 原因：深度网络模型目标函数过于复杂
  - 应对策略：
    - 改进初始化方式，如DAE、DBN
    - 随机梯度下降，跳出局部极值
    - 设置冲量，加速优化

【笔记】

练习

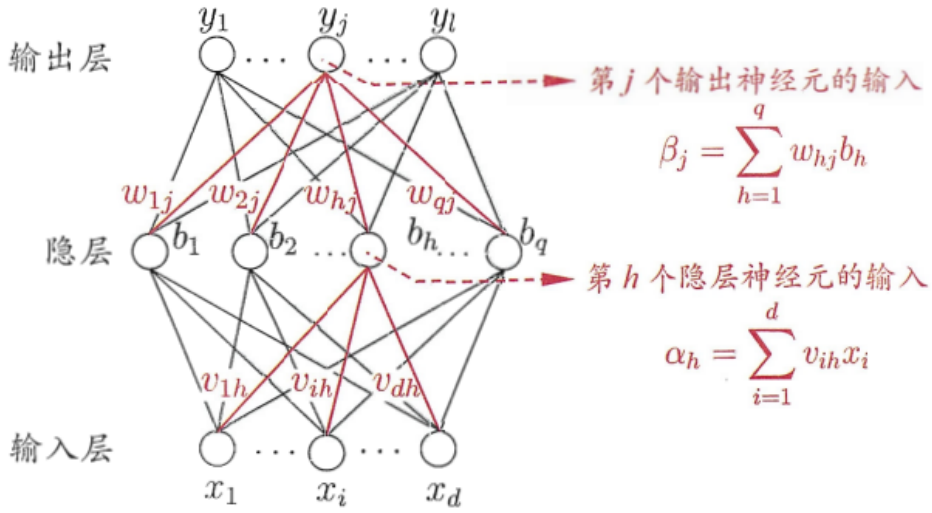
1. 现有如图所示的神经网络模型，其中各神经元均使用Sigmoid激活函数，假设当前网络的权重向量$\boldsymbol{w}$取值为：

$$\boldsymbol{W} = \left(w_{11}^{(1)}, w_{12}^{(1)}, w_{21}^{(1)}, w_{22}^{(1)}, b_1^{(2)}, b_2^{(2)}, w_{11}^{(2)}, w_{12}^{(2)}, w_{21}^{(2)}, w_{22}^{(2)}, b_1^{(3)}, b_2^{(3)}\right)^T$$
$$= (0.2,03,0.4,0.5,0.3,0.3,0.7,0.5,0.8,0.3,0.5,0.5)^T$$

试用样本 (X, $\boldsymbol{y}$) = [0.4,0.6 , (0.6,0.6)]完成对连接权重的一次更新

2. 手写体数字识别
    - 识别数字2
    - 识别10个数字

3. 如下的多层神经网络，激活函数使用Sigmoid，对于输入`( xk， yk)`

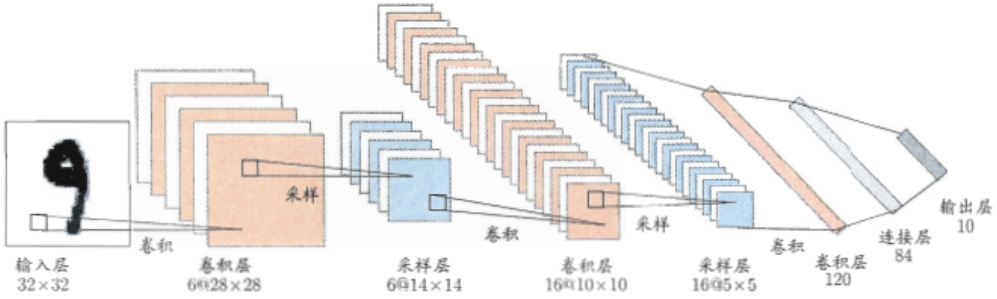


- 写出均方误差 `Ek` 的表达式
- 写出均方误差对第 `j` 层输出神经元的输入的偏导
- 写出均方误差对第 `h` 个隐层神经元的输入的偏导
- 给出

  $$\triangle w_{hj}、\triangle v_{ih}$$

  的更新方程

4. 给出下面用于手写体识别的卷积神经网络每层参数的个数



第九章 聚类

基本概念

- 无监督学习通过比较样本之间的某种联系实现对样本的数据分析。最大特点是学习算法的输入是无标记样本。
-

k-均值划分聚类

- 基本思想：同类样本在特征空间中应该相距不远
- 主要方法：将集中在特征空间某一区域内的样本划分为 同一个簇
- 区域位置的界定主要通过样本特征值的均值确定

1：令$s=0$，并从D中随机生成k个作为初始聚类中心的数据点$u_1^0, u_2^0, \ldots, u_k^0$；

2：计算D中各样本与各簇中心之间的距离w，并根据w值将其分别划分到簇中心点与其最近的簇中；

3：分别计算各簇中所有示例样本数据的均值，并分别将每个簇所得到的均值作为该簇新的聚类中心 $u_1^{s+1}, u_2^{s+1}, \ldots, u_k^{s+1}$；

4：若$u_j^{s+1}=u_j^s$，则终止算法并输出最终簇，否则令$s=s+1$，并返回步骤2

DBSCAN密度聚类

– 任意选取一个点p

– 得到所有从p关于 ε 和 $MinPts$ 密度可达的点

– 如果p是一个核心点, 则找到一个聚类

– 如果p是一个边界点, 没有从p密度可达的点, DBSCAN 将访问数据库中的下一个点

– 继续这一过程, 直到数据库中的所有点都被处理

- 算法基本思想（划分聚类、密度聚类），解决什么问题，优缺点

例题

1. 表为某机构15支足球队在2017-2018年间的积分，各队在各赛事中的水平发挥有所不同。若将球队的水平分为三个不同的层次水平，试用k-均值聚类方法分析哪些队伍的整体水平比较相近

队伍	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
赛事1	50	28	17	25	28	50	50	50
赛事2	50	9	15	40	40	50	40	40
赛事3	9	4	3	5	2	1	9	9
队伍	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	
赛事1	40	50	50	50	40	40	50	
赛事2	40	50	50	50	40	32	50	
赛事3	5	9	5	9	9	17	9	

2. 试分析划分聚类算法和密度聚类算法所能得到的簇边界有何区别

3. 划分聚类

   – K均值聚类

   – 模糊k-均值聚类

   密度聚类

   – DBSCAN密度聚类

   – OPTICS密度聚类

   – DENCLUE密度聚类

   说明以上5个算法的基本思想、优缺点以及解决什么问题。

第十章 主成分分析

基本概念

- 主成分分析（principal component analysis, PCA）是一种常用的无监督学习方法
- 该方法利用正交变换把由线性相关变量表示的观测数 据转换为少数几个由线性无关变量表示的数据，线性无关的变量称为主成分
- 主成分的个数通常小于原始变量的个数，所以主成分分析属于降维方法
- 主成分分析主要用于发现数据中的基本结构，即数据中 变量之间的关系

- 方差贡献率

- 若$k=m$，则转换后数据保留了原数据的全部信息；若$k=0$，则相当于完全不展示原数据的信息
- 在确定k的具体取值时，通常会考虑不同k值可保留方差的百分比并称这种分量方差占总方差的百分比为该分量对总方差的贡献率，简称为方差贡献率
- 令$\lambda_1,\lambda_2,\ldots,\lambda_n$表示协方差矩阵$\boldsymbol{C}$的全部特征值且按由大到小顺序排列，$\boldsymbol{w}_i$为特征值$\lambda_i$所对应的特征向量，若保留变换后样本数据前$k$个分量，则得到相应累计方差贡献率$\Omega$为

$$\Omega=\sum_{i=1}^{k}\lambda_i\Big/\sum_{i=1}^{n}\lambda_i$$

- 通常选择k以保留99%或97%的累计方差贡献率，即选取满足$\Omega\geq 0.99$或$\Omega\geq 0.97$的最小k值

基本思想

- 主成分分析中，首先对给定数据进行规范化，使得数据 每一变量的平均值为0，方差为1
- 之后对数据进行正交变换，原来由线性相关变量表示的 数据，通过正交变换变成由若干个线性无关的新变量表 示的数据
- 新变量是可能的正交变换中变量的方差的和（信息保 存）最大的，方差表示在新变量上信息的大小
- 可以用主分成近似地表示原始数据，发现数据的基本结 构
- 也可以把数据由少数主成分表示，对数据降维

例题

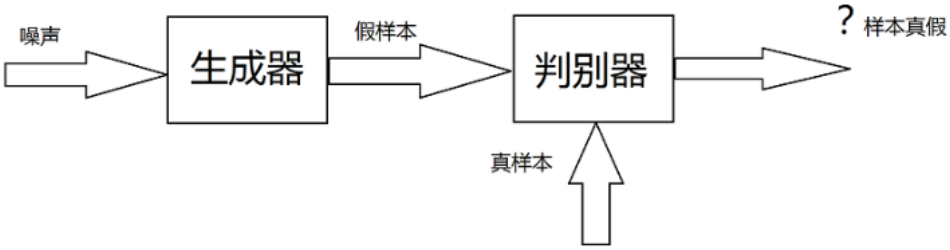
1. 现有我国大陆30个省、直辖市、自治区的经济发展 状况数据集如表4-12所示，包括8项经济指标：国民生产总 值（$a1$）；居民消费水平（$a2$）；固定资产投资（$a3$）； 职工平均工资（$a4$）；货物周转量（$a5$）；居民消费指数 （$a6$）；商品零售价格指数（$a7$）；工业总产值（$a8$）， 试用PCA方法将这8项经济指标融合成3项综合指标

	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8
北京	1394.89	2505	519.01	8144	373.9	117.3	112.6	843.43
天津	920.11	2720	345.46	6501	342.8	115.2	110.6	582.51
河北	2849.52	1258	704.87	4839	2033.3	115.2	115.8	1234.85
山西	1092.48	1250	290.9	4721	717.3	116.9	115.6	697.25
内蒙古	832.88	1387	250.23	4134	781.7	117.5	116.8	419.39
辽宁	2793.37	2397	387.99	4911	1371.1	116.1	114	1840.55
吉林	1129.2	1872	320.45	4430	497.4	115.2	114.2	762.47
黑龙江	2014.53	2334	435.73	4145	824.8	116.1	114.3	1240.37
上海	2462.57	5343	996.48	9279	207.4	118.7	113	1642.95
江苏	5155.25	1926	1434.95	5934	1025.5	115.8	114.3	2026.64
浙江	3524.79	2249	1006.39	6619	754.4	116.6	113.5	916.59
安徽	2003.58	1254	474	4609	908.3	114.8	112.7	824.14
福建	2160.52	2320	553.97	5857	609.3	115.2	114.4	433.67
江西	1205.1	1182	282.84	4211	411.7	116.9	115.9	571.84
山东	5002.34	1527	1229.55	5145	1196.6	117.6	114.2	2207.69
河南	3002.74	1034	670.35	4344	1574.4	116.5	114.9	1367.92
湖北	2391.42	1527	571.68	4685	849	120	116.6	1220.72
湖南	2195.7	1408	422.61	4797	1011.8	119	115.5	843.83
广东	5381.72	2699	1639.83	8250	656.5	114	111.6	1396.35
广西	1606.15	1314	382.59	5150	556	118.4	116.4	554.97
海南	364.17	1814	198.35	5340	232.1	113.5	111.3	64.33
四川	3534	1261	822.54	4645	902.3	118.5	117	1431.81
贵州	630.07	942	150.84	4475	301.1	121.4	117.2	324.72
云南	1206.68	1261	334	5149	310.4	121.3	118.1	716.65
西藏	55.98	1110	17.87	7382	4.2	117.3	114.9	5.57
陕西	1000.03	1208	300.27	4396	500.9	119	117	600.98
甘肃	553.35	1007	114.81	5493	507	119.8	116.5	468.79
青海	165.31	1445	47.76	5753	61.6	118	116.3	105.8
宁夏	169.75	1355	61.98	5079	121.8	117.1	115.3	114.4
新疆	834.57	1469	376.95	5348	339	119.7	116.7	428.76

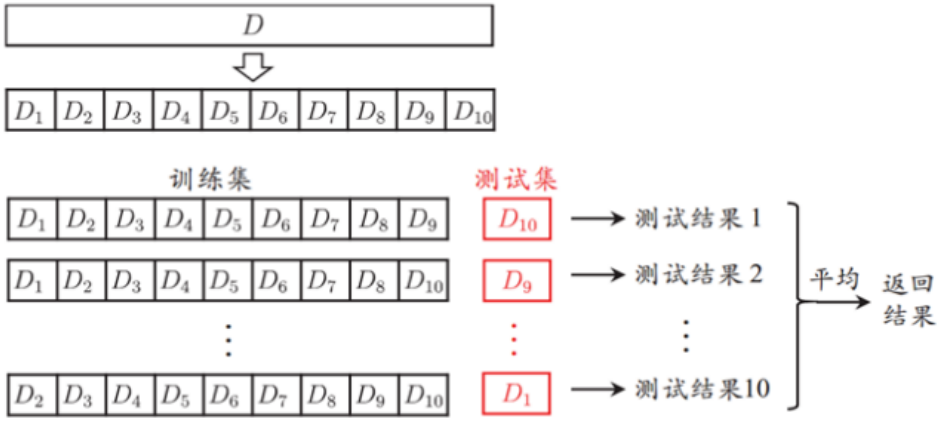
第十一章 问题与策略

基本概念

- 数据问题
  - 生成对抗网络基本思想：生成器生成新的样本，将生成的虚拟样本和训练集中实际图像样本随机输入判别器中，通过GAN模型的判别器判别输入样本是否为虚拟样本



- 评测问题
  - 留出法
    - 直接将数据集划分为两个互斥集合
    - 训练/测试集划分要尽可能保持数据分布的一致性
    - 一般若干次随机划分、重复实验取平均值
    - 训练/测试样本比例通常为`2:1~4:1`
  - 交叉验证法
    - 将数据集分层采样划分为k个大小相似的互斥子集
    - 每次用k-1个子集的并集作为训练集，余下的子集作为测试集， 最终返回k个测试结果的均值
    - k最常用的取值是10



- 自助法
  - 以自助采样法为基础，对数据集D有放回采样m次得到训练集， 其余作为测试集
  - 实际模型与预期模型都使用m个训练样本
  - 约有1/3的样本没在训练集中出现（Why？）
- 比较

  自助法在数据集较小、难以有效划分训练/测试集时很有用；

  由于改变了数据集分布可能引入估计偏差，在数据量足够时，留出法和交叉验证法更常用

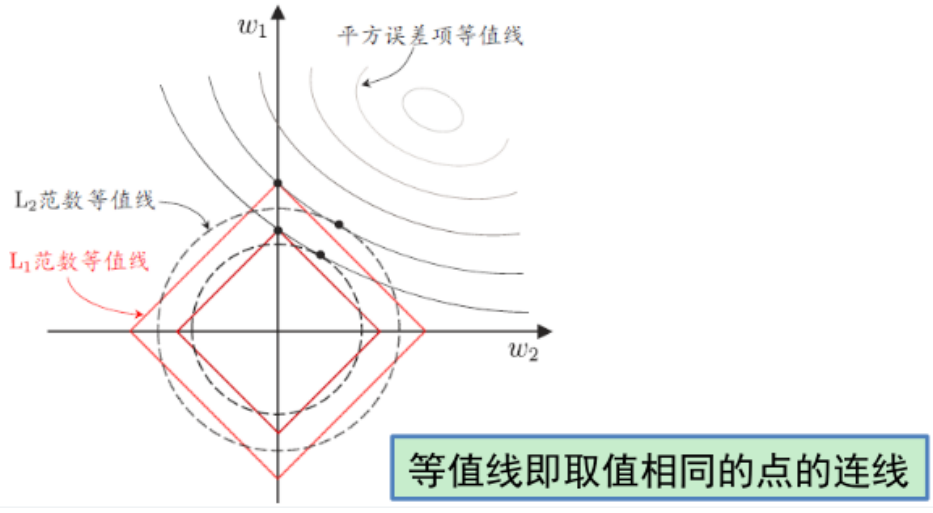
- 正则化

结构风险最小化

$$\arg\min_{w,b} \sum_{i=1}^{n} l\left(g\left(\boldsymbol{x}^i;\boldsymbol{w}\right),y^i\right)+\lambda L(\boldsymbol{w})$$

- 简单化： L_2范数-岭回归 $L(\boldsymbol{w})=\|\boldsymbol{w}\|_2^2$
- 稀疏化： L_1范数 $L(\boldsymbol{w})=\|\boldsymbol{w}\|_1$

- L1和L2范数正则化
  - 假设$\boldsymbol{x}$仅有两个属性，那么$\boldsymbol{w}$有两个分量$w1$和$w2$. 那么目标优化的解要在平方误差项与正则化项之间折中,即出现在图中平方误差项等值线与正则化等值线相交处
  - 从图中看出,采用L1范数时交点常出现在坐标轴上,即产生 $w1$或者$w2$为0的稀疏解



练习

1. 试分析L1和L2范数正则化的优缺点

第十二章 计算学习理论

参考例题

1. 下表中每一行表示一个样本，每个样本有四个0-1属 性，每个样本的类别y见最后一列。请使用前8个样本作为训练集D构造一棵决策 树。
    - 计算D的熵。
    - 使用信息增益作为优化目标，该决策树的树根所使用 的属性是?
    - 使用你所选择的属性，树根的信息增益是多少？写出计算过程。
    - 画出你所构造的决策树。
    - 使用你的决策树，对 x(9) 的分类结果是什么?

	x_1	x_2	x_3	x_4	y
$x^{(1)}$	0	0	0	0	0
$x^{(2)}$	0	0	1	0	0
$x^{(3)}$	1	1	0	1	0
$x^{(4)}$	1	0	0	1	1
$x^{(5)}$	0	1	1	0	1
$x^{(6)}$	0	0	1	1	1
$x^{(7)}$	0	0	0	1	1
$x^{(8)}$	1	1	0	0	1
$x^{(9)}$	1	1	1	1	?

2. .二维空间的五个样本点 A(1, 5)、B(5, 5)、C(3, 3)、D(1, 1)、E(5, 1)，其中 A,B,D,E 是正样本，C 是负样本。利用这 5 个样本，使用 AdaBoost 算法训练一个强分类器，弱分类器为 Decision Stump。请回答一下问题：

1）下面哪一个可能是一个弱分类器 $h_1(x)$？（A）

A.sign(x_1)　　B. sign(x_1-2)　　C. sign(x_1-4)　　D. sign(x_1-6)

2）上面选择的弱分类器的错误率是多少？（A）

A.0.2　　B. 0.4　　C.0.6　　D. 0.8

3）$h_1(x)$的权值为（B）

A.$\frac{1}{2}ln2$　　B. $ln2$　　C.$\frac{1}{2}ln\frac{3}{2}$　　D. $\frac{1}{2}ln\frac{2}{3}$

4）选择 $h_2(x)$是样本的权重是：A（　），B（　），C（　），D（　），E（　）

5）下列哪一个是 $h_2(x)$？（C）

A.sign(x_1)　　B. sign(x_1-2)　　C. sign(x_1-4)　　D. sign(x_1-6)

6）上面学长的弱分类器错误率是多少？（C）

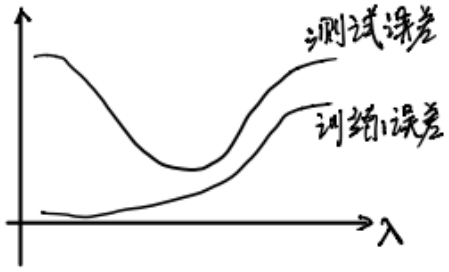
A.1/5　　B. 1/4　　C.2/5　　D. 3/8

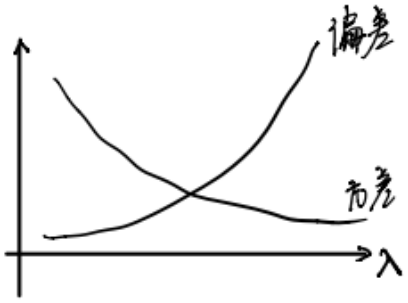
7）$h_2(x)$的权值为（　）A/D

A.$\frac{1}{2}ln2$　　B. $ln2$　　C.$\frac{1}{2}ln3$　　D. $ln3$

3. 选择题

10.假设正则化的经验风险为 $L(y,h(x;\theta))=\frac{1}{m}\sum l(y,h\,h(x;\theta))+\lambda$，$\lambda \geqslant 0$ 是正则化系数，请在下图左图中画出测试误差、训练误差与 λ 的关系曲线。在右图中画出方差和偏差与 λ 的关系曲线。





11.一下情况发生了过拟合的是（D）

A.训练误差很大，测试误差很小　　B.训练误差与测试误差都很小

C.训练误差与测试误差都很大 欠拟合　　D. 训练误差很小，测试误差很大

12.在模型发生过拟合和欠拟合情况下分别可以采取哪些对策？

过拟合:（A、B、G、E、D）；　　欠拟合:（C、F、H）；

A.增加训练样本；　B.增大正则化系数；　C.减少正则化系数；　D.采用模型平

E.采用 Boosting 方法；　F.增加特征；　G.减少特征；　H.采用更复杂的模型

4.

6. 假设R^2上定义的核函数为$K(x,z)=x^Tz+1$,（$x=(x_1,x_2)^T$，$z=(z_1,z_2)^T$）请推导出该核函数隐式定义的特征映射函数 $\phi(x)$。

$K(x,z)=x^Tz+1$

$=[x_1,x_2]\begin{bmatrix}z_1\\z_2\end{bmatrix}+1$

$=x_1z_1+x_2z_2+1$

$=\langle(x_1,x_2,1)\rangle,\langle(z_1,z_2,1)\rangle$

(一)(告诉)

$\therefore (x_1,x_2)\Rightarrow(x_1,x_2,1)$

(二)

$x=(x_1,x_2)^T$，$z=(z_1,z_2)^T$ 代入 $K(x,z)$:

$\psi(x)=K(x,z)=[(x_1,x_2)^T]^T(z_1,z_2)^T+1$

$=(x_1,x_2)(z_1,z_2)^T$

$=x_1z_1+x_2z_2+1$ ✓

模拟试卷

一、选择题

二、填空题

1. 最常用的两种监督学习任务是，。
2. 列举出四种常见的有监督式学习的任务，，和。
3. 垃圾邮箱检测的问题是监督学习还是无监督学习 。
4. 在梯度下降中，沿着负梯度方向进行下一步探索，前进距离为：；这种参数属于 。
5. L1正则化在原来损失函数的基础上加上___。
6. L2正则化在原来损失函数的基础上加上___。
7. 减少过拟合可以提升模型的能力。

三、计算题

1. 已知正例点x1 = (1, 2)T，x2 = (2, 3)T ，负例点x3 = (2, 1)T，试求最大间隔分离超平面和分类决策函数，并在图上画出分离超平面、间隔边界及支持向量。

四、综合题

1. (1) 在聚类分析中，传统的K-means算法都有哪些局限性？有哪些相应的改进方法？

   (2) 请简要描述聚类与关联分析的主要相似点和不同点。

   (3) 请举出一个采用聚类作为主要的数据挖掘方法的实际应用例子。

2. 当模型出现过拟合或者欠拟合时，应该采用什么方法来解决？

五、答案

填空题：https://github.com/OracleClubAI/Test/edit/master/doc/