Technology Science Information Networks Computing

Lecturer: Ting Wang (王挺)

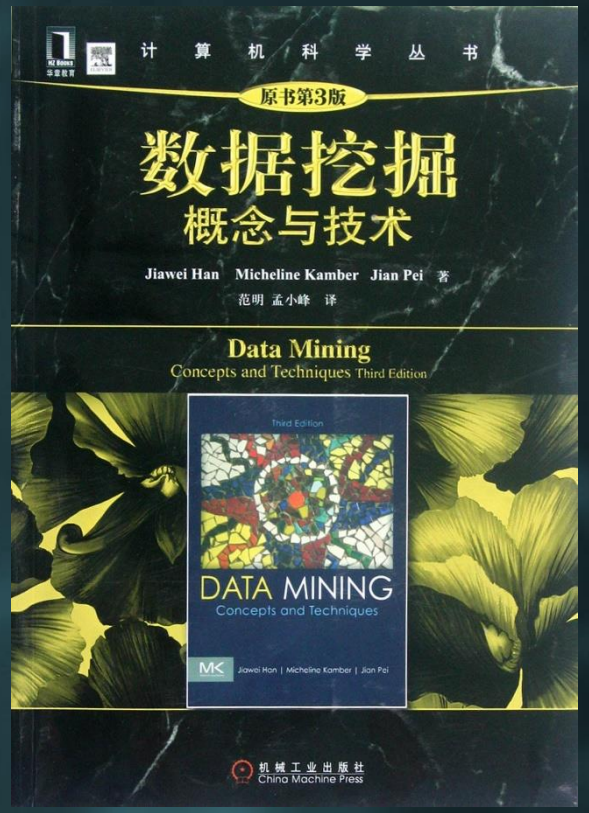
利物浦大学计算机博士

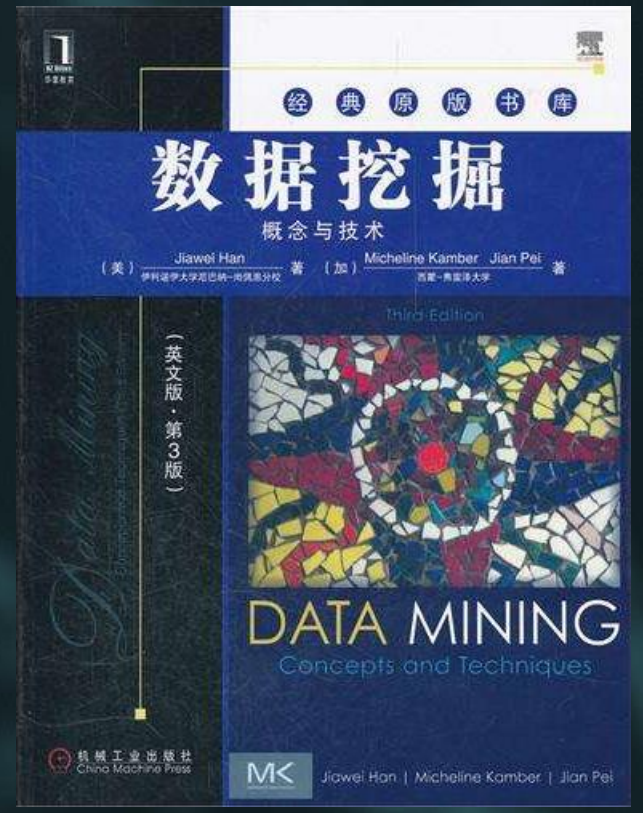
清华大学计算机博士后

电子信息技术高级工程师

上海外国语大学网络与新媒体副教授

浙江清华长三角研究院海纳认知与智能研究中心主任





Chapter 9

Classification: Advanced Methods

Chapter 9 Classification: Advanced Methods

1. **Bayesian belief networks (compared with** *Naive Bayesian classifiers*)

- do not assume class conditional independence
- allow the representation of dependencies among subsets of attributes

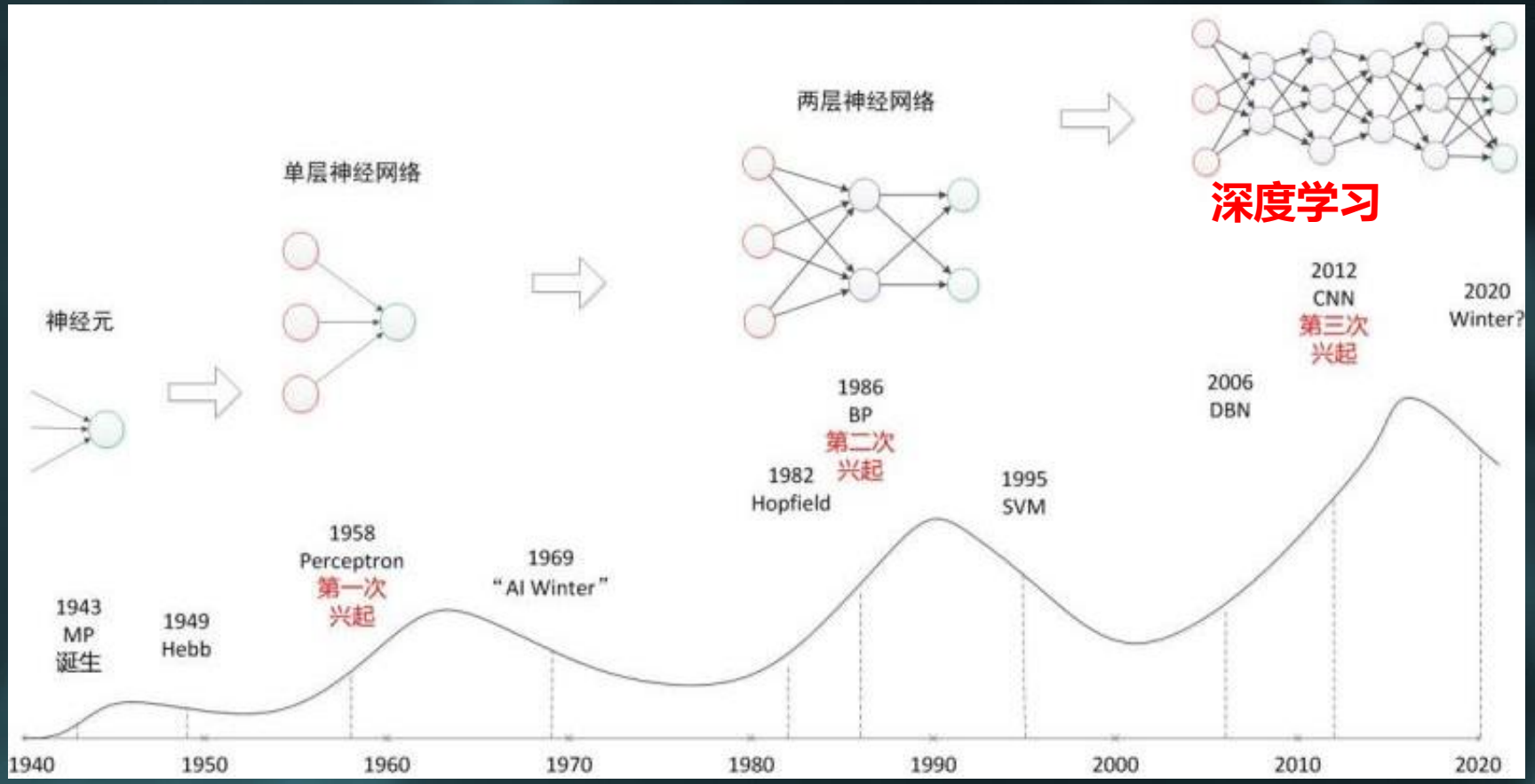
	贝叶斯网络	朴素贝叶斯
假设前提	各变量都是离散型的。各特征有依赖（不确定的因果推理）关系（变量无关）。 每一个节点在其直接前驱节点的值制定后，这个节点条件独立于其所有非直接前驱前辈节点。 P(v|par(v),x1,x2...,xn) = P(v|par(v)) 贝叶斯网络放宽了每个变量独立的假设。	各特征彼此独立。 朴素贝叶斯中对于若干条件概率值不存在的问题，一般通过将所有的概率值加1来解决。 且对被解释变量的影响一致，不能进行变量筛选。
应用案例	在信息不完备的情况下通过可以观察随机变量推断不可观察的随机变量解决文本分类时，相邻词的关系、近义词的关系	分类
缺点	不能对变量进行筛选，因为不能放宽对被解释变量影响一致的假设	彼此不独立的特征之间建立朴素贝叶斯，反而加大了模型复杂性
优点	贝叶斯原理和图论相结合，建立起一种基于概率推理的数学模型,对于解决复杂的不确定性和关联性问题有很强的优势 ● 对缺失数据不敏感 ● 可以学习因果关系，加深对数据的理解 ● 能将先验知识融入建模 ● 避免了过度拟合问题，不需要保留数据进行检验	简单，对于给出的待分类项，会选择条件概率最大的类别，这就是朴素贝叶斯的思想基础

Ref：https://blog.csdn.net/sinat_26917383/article/details/51569573

Chapter 9 Classification: Advanced Methods

2. Classification by Backpropagation

(1) neural network



Chapter 9 Classification

(2) Backpropagation

- Backpropagation is a neural network algorithm for classification that employs a method of gradient descent.
- It searches for a set of weights that can model the data so as to minimize **the mean-squared distance** between the network' s class prediction and the actual class label of data tuples.
- Rules may be extracted from trained neural networks to help improve the interpretability of the learned network.

Algorithm: Backpropagation. Neural network learning for classification or numeric prediction, using the backpropagation algorithm.

Input:

- D, a data set consisting of the training tuples and their associated target values;
- l, the learning rate;
- *network*, a multilayer feed-forward network.

Output: A trained neural network.

Method:

(1) Initialize all weights and biases in *network*;
(2) **while** terminating condition is not satisfied {
(3) **for** each training tuple X in D {
(4) // Propagate the inputs forward:
(5) **for** each input layer unit j {
(6) $O_j = I_j$; // output of an input unit is its actual input value
(7) **for** each hidden or output layer unit j {
(8) $I_j = \sum_i w_{ij} O_i + \theta_j$; //compute the net input of unit j with respect to the previous layer, i
(9) $O_j = \frac{1}{1+e^{-I_j}}$; } // compute the output of each unit j
(10) // Backpropagate the errors:
(11) **for** each unit j in the output layer
(12) $Err_j = O_j(1 - O_j)(T_j - O_j)$; // compute the error
(13) **for** each unit j in the hidden layers, from the last to the first hidden layer
(14) $Err_j = O_j(1 - O_j) \sum_k Err_k w_{jk}$; // compute the error with respect to the next higher layer, k
(15) **for** each weight w_{ij} in *network* {
(16) $\Delta w_{ij} = (l) Err_j O_i$; // weight increment
(17) $w_{ij} = w_{ij} + \Delta w_{ij}$; } // weight update
(18) **for** each bias θ_j in *network* {
(19) $\Delta \theta_j = (l) Err_j$; // bias increment
(20) $\theta_j = \theta_j + \Delta \theta_j$; } // bias update
(21) } }

Chapter 9 Classification: Advanced Methods

Chapter 9 Classification: Advanced Methods

基于深度学习神经网络的文章影响力预测

标题	代码	阅读量	转发	收藏	点赞	留言	新关注人数	取消关注人数	累计关注人数
啤酒大片的正确打开方式	8-1	5,679	318	4	37	1	124	493	60544
你离成功可能还差一座啤酒屋	8-2	897	31	0	15	3	124	493	60544
嘉士伯穿上红军外套，帅到没朋友！	8-3	13,848	644	2	72	17	160	511	60250
携手25周年，一起定格感动瞬间	8-4	766	34	0	14	2	160	511	60250
时光机启动，一场啤酒界“国王的演讲”即将开启	8-5	3,536	99	2	14	7	194	550	60347
红军捷报频传，25年老友助阵	8-6	445	14	0	4	0	194	550	60347
足球+啤酒，夏日最佳挚友	8-7	480	31	0	5	1	194	550	60347
足球解说员詹俊为“红军”深情打call！	8-8	3,228	200	3	14	5	261	516	60335
一罐披上红军外套啤酒引发的……	8-9	1,663	140	1	3	6	261	516	60335
“啤酒教父”竟是隐藏的哲学家	9-1	4,266	140	9	24	3	44	393	59781
和最爱的球队永不“分手”	9-2	4,202	189	3	16	6	42	373	59324
俊哥如此“多娇”，引无数球迷竞折腰	9-3	549	6	0	6	5	42	373	59324
世纪难题：足球和恋人，哪个更重要?	9-4	3,085	143	1	14	7	44	339	58793
球场恰似人生，“红军”继续雄起！	9-5	290	5	0	4	0	44	339	58793
丹麦“国民男神”到访天朝，快来围观！	9-6	3,547	105	3	28	4	53	299	58295
850+170=完美国庆假期	9-7	698	10	2	4	1	53	299	58295
狂欢之旅，从这个啤酒节开始！	10-1	3,448	97	0	25	0	36	356	57757
蓦然回首，你是哪年参军的?	10-2	564	8	0	6	3	36	356	57757
夜幕降至，嘉倍放肆，好胆你就来！	10-3	3,351	164	2	23	21	50	377	57219
“双红会”平局成瘾，冤家当道怎么破?	10-4	293	6	0	4	7	50	377	57219
啤酒花观看利物浦队球赛长大，嘉士伯酿造限量瓶啤酒	10-5	344	161	1	21	2	36	237	56719
利物浦球赛，嘉士伯啤酒花也爱看！	10-6	1,716	104	2	6	4	26	56	56365
走，去丹麦领事家赴宴！	11-1	2,361	104	1	23	2	26	116	61085
流淌着利物浦故事的啤酒是什么味?	11-2	398	15	0	3	0	26	116	61085
啤酒花拯救人类？一点没错！	11-3	3,620	93	2	16	7	23	299	60603
福布斯推出新榜，前10都有谁?	11-4	5,084	281	12	52	0	31	278	60006
萨拉赫连进两球，是不是你心中的Top3?	11-5	218	5	0	3	4	31	278	60006
没有啤酒，可能就没有感恩节?	11-6	3,454	121	6	33	45	28	265	59485
在北欧过冬——比你艰苦，但也比你会玩	11-7	299	3	1	5	1	28	265	59485
今天你Hygge了吗?	12-1	2,808	101	5	14	0	38	298	58944

3个月的训练数据

测试数据

预测分析结果

	阅读量	转发	收藏	点赞	留言	新关注人数	取消关注人数	累计关注人数
周1	2774	116	2	20	8	49	384	78169
周2	2755	116	2	20	8	49	381	77653
周3	2760	116	2	20	8	49	382	77772
周4	2751	115	2	20	8	49	381	77527
周5	2837	119	3	21	8	50	393	79951
周6	2734	115	2	20	7	49	378	77040
周7	2734	115	2	20	7	49	378	77040

说明：彩虹色从红到紫依次为文章发表时间的推荐程度

预测分析

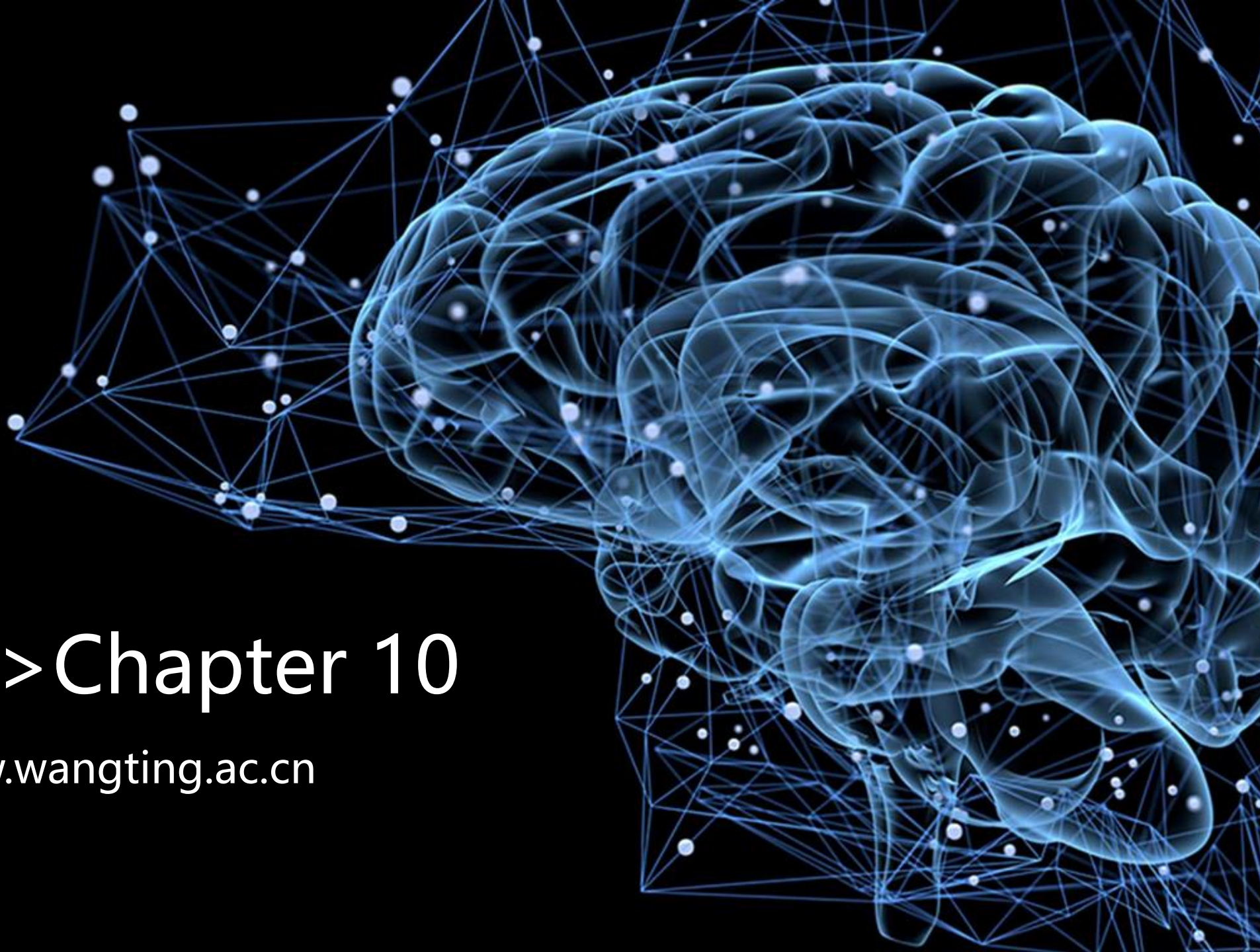
结果显示，周五发送最佳。除取消关注人数和累计关注人数过多外，其余各项基本符合预测需求。其中，阅读量预测值2837，实际值2808，误差在1%。

Chapter 9 Classification: Advanced Methods

3. SVM（支持向量机）

- an algorithm for the classification of both linear and nonlinear data.
- It transforms the original data into a higher dimension, from where it can find a hyperplane for data separation using essential training tuples called **support vectors**

- **High dimension**
- **Kernel function**



Next>>Chapter 10

www.wangting.ac.cn