
目 录

前言

符号表

第一部分 概述	1
第 1 章 绪论	2
1.1 样例: 自动驾驶	3
1.2 模式识别与机器学习	5
1.2.1 一个典型的模式识别流程	5
1.2.2 模式识别 vs. 机器学习	8
1.2.3 评估、部署和细化	9
1.3 本书的结构	9
习题	12
第 2 章 数学背景知识	14
2.1 线性代数	14
2.1.1 内积、范数、距离和正交性	14
2.1.2 角度与不等式	15
2.1.3 向量投影	16
2.1.4 矩阵基础	17
2.1.5 矩阵乘法	18
2.1.6 方阵的行列式与逆	19
2.1.7 方阵的特征值、特征向量、秩和迹	20
2.1.8 奇异值分解	22
2.1.9 (半) 正定实对称矩阵	22
2.2 概率	23
2.2.1 基础	23
2.2.2 联合分布、条件分布与贝叶斯定理	25
2.2.3 期望与方差/协方差矩阵	26
2.2.4 不等式	27
2.2.5 独立性与相关性	28
2.2.6 正态分布	29
2.3 优化与矩阵微积分	30

2.3.1	局部极小、必要条件和矩阵微积分	30
2.3.2	凸优化与凹优化	31
2.3.3	约束优化和拉格朗日乘子法	33
2.4	算法复杂度	34
2.5	阅读材料	35
	习题	35
第 3 章	模式识别系统概述	39
3.1	人脸识别	39
3.2	一个简单的最近邻分类器	40
3.2.1	训练或学习	40
3.2.2	测试或预测	40
3.2.3	最近邻分类器	41
3.2.4	k -近邻	42
3.3	丑陋的细节	43
3.4	制定假设并化简	46
3.4.1	设计工作环境 vs. 设计复杂算法	46
3.4.2	假设与简化	47
3.5	一种框架	51
3.6	阅读材料	51
	习题	53
第 4 章	评估	55
4.1	简单情形中的准确率和错误率	55
4.1.1	训练与测试误差	56
4.1.2	过拟合与欠拟合	56
4.1.3	使用验证集来选择超参数	58
4.1.4	交叉验证	59
4.2	最小化代价/损失	61
4.2.1	正则化	62
4.2.2	代价矩阵	62
4.2.3	贝叶斯决策理论	63
4.3	不平衡问题中的评估	64
4.3.1	单个类别内的比率	64
4.3.2	ROC 曲线下的面积	65
4.3.3	查准率、查全率和 F 值	66
4.4	我们能达到 100% 的准确率吗?	68
4.4.1	贝叶斯错误率	68
4.4.2	真实标记	69
4.4.3	偏置-方差分解	70

4.5 对评估结果的信心	73
4.5.1 为什么要取平均?	73
4.5.2 为什么要报告样本标准差?	74
4.5.3 比较两个分类器	75
4.6 阅读材料	79
习题	79
第二部分 与领域知识无关的特征提取	83
第 5 章 主成分分析	84
5.1 动机	84
5.1.1 维度与内在维度	84
5.1.2 降维	86
5.1.3 PCA 与子空间方法	86
5.2 PCA 降维到零维子空间	86
5.2.1 想法-形式化-优化实践	87
5.2.2 一个简单的优化	87
5.2.3 一些注释	88
5.3 PCA 降维到一维子空间	88
5.3.1 新的形式化	88
5.3.2 最优性条件与化简	89
5.3.3 与特征分解的联系	90
5.3.4 解	91
5.4 PCA 投影到更多维度	91
5.5 完整的 PCA 算法	92
5.6 方差的分析	93
5.6.1 从最大化方差出发的 PCA	94
5.6.2 一种更简单的推导	95
5.6.3 我们需要多少维度呢?	95
5.7 什么时候使用或不用 PCA 呢?	96
5.7.1 高斯数据的 PCA	96
5.7.2 非高斯数据的 PCA	96
5.7.3 含异常点数据的 PCA	98
5.8 白化变换	98
5.9 特征分解 vs. SVD	98
5.10 阅读材料	99
习题	99
第 6 章 Fisher 线性判别	103
6.1 用于二分类的 FLD	104

6.1.1	想法: 什么是隔得很远呢?	104
6.1.2	翻译成数学语言	105
6.1.3	散度矩阵 vs. 协方差矩阵	107
6.1.4	两种散度矩阵以及 FLD 的目标函数	108
6.1.5	优化	108
6.1.6	等等, 我们有一条捷径	109
6.1.7	二分类问题的 FLD	109
6.1.8	陷阱: 要是 S_W 不可逆呢?	110
6.2	用于多类的 FLD	111
6.2.1	稍加修改的符号和 S_W	111
6.2.2	S_B 的候选	111
6.2.3	三个散度矩阵的故事	112
6.2.4	解	113
6.2.5	找到更多投影方向	113
6.3	阅读材料	113
	习题	114

第三部分 分类器与其他工具.....119

第 7 章 支持向量机 120

7.1	SVM 的关键思想	120
7.1.1	简化它! 简化它! 简化它!	120
7.1.2	查找最大 (或较大) 间隔的分类器	121
7.2	可视化并计算间隔	122
7.2.1	几何的可视化	123
7.2.2	将间隔作为优化来计算	124
7.3	最大化间隔	124
7.3.1	形式化	125
7.3.2	各种简化	125
7.4	优化与求解	127
7.4.1	拉格朗日函数与 KKT 条件	127
7.4.2	SVM 的对偶形式	128
7.4.3	最优的 b 值与支持向量	129
7.4.4	同时考虑原始形式与对偶形式	131
7.5	向线性不可分问题和多类问题的扩展	131
7.5.1	不可分问题的线性分类器	132
7.5.2	多类 SVM	134
7.6	核 SVM	134
7.6.1	核技巧	135

7.6.2	Mercer 条件与特征映射	136
7.6.3	流行的核函数与超参数	137
7.6.4	SVM 的复杂度、权衡及其他	138
7.7	阅读材料	139
	习题	139
第 8 章	概率方法	144
8.1	思考问题的概率路线	144
8.1.1	术语	144
8.1.2	分布与推断	145
8.1.3	贝叶斯定理	145
8.2	各种选择	146
8.2.1	生成式模型 vs. 判别式模型	146
8.2.2	参数化 vs. 非参数化	147
8.2.3	该如何看待一个参数呢?	148
8.3	参数化估计	148
8.3.1	最大似然	148
8.3.2	最大后验	150
8.3.3	贝叶斯	151
8.4	非参数化估计	153
8.4.1	一个一维的例子	153
8.4.2	直方图近似中存在的问题	155
8.4.3	让你的样本无远弗届	156
8.4.4	核密度估计	157
8.4.5	带宽选择	158
8.4.6	多变量 KDE	158
8.5	做出决策	159
8.6	阅读材料	159
	习题	160
第 9 章	距离度量与数据变换	163
9.1	距离度量和相似度度量	163
9.1.1	距离度量	164
9.1.2	向量范数和度量	164
9.1.3	ℓ_p 范数和 ℓ_p 度量	165
9.1.4	距离度量学习	167
9.1.5	均值作为一种相似度度量	168
9.1.6	幂平均核	170
9.2	数据变换和规范化	171
9.2.1	线性回归	172

9.2.2 特征规范化	173
9.2.3 数据变换	175
9.3 阅读材料	177
习题	177
第 10 章 信息论和决策树	182
10.1 前缀码和霍夫曼树	182
10.2 信息论基础	183
10.2.1 熵和不确定性	184
10.2.2 联合和条件熵	184
10.2.3 互信息和相对熵	185
10.2.4 一些不等式	186
10.2.5 离散分布的熵	187
10.3 连续分布的信息论	187
10.3.1 微分熵	188
10.3.2 多元高斯分布的熵	189
10.3.3 高斯分布是最大熵分布	191
10.4 机器学习和模式识别中的信息论	192
10.4.1 最大熵	192
10.4.2 最小交叉熵	193
10.4.3 特征选择	194
10.5 决策树	195
10.5.1 异或问题及其决策树模型	195
10.5.2 基于信息增益的结点划分	197
10.6 阅读材料	198
习题	199
第四部分 处理变化多端的数据	203
第 11 章 稀疏数据和未对齐数据	204
11.1 稀疏机器学习	204
11.1.1 稀疏 PCA?	204
11.1.2 使用 ℓ_1 范数诱导稀疏性	205
11.1.3 使用过完备的字典	208
11.1.4 其他一些相关的话题	210
11.2 动态时间规整	212
11.2.1 未对齐的时序数据	212
11.2.2 思路(或准则)	213
11.2.3 可视化和形式化	214
11.2.4 动态规划	215

11.3 阅读材料	218
习题	218
第 12 章 隐马尔可夫模型	222
12.1 时序数据与马尔可夫性质	222
12.1.1 各种各样的时序数据和模型	222
12.1.2 马尔可夫性质	224
12.1.3 离散时间马尔可夫链	225
12.1.4 隐马尔可夫模型	227
12.2 HMM 学习中的三个基本问题	228
12.3 α 、 β 和评估问题	229
12.3.1 前向变量和算法	230
12.3.2 后向变量和算法	231
12.4 γ 、 δ 、 ψ 和解码问题	234
12.4.1 γ 和独立解码的最优状态	234
12.4.2 δ 、 ψ 和联合解码的最优状态	235
12.5 ξ 和 HMM 参数的学习	237
12.5.1 Baum-Welch: 以期望比例来更新 λ	238
12.5.2 如何计算 ξ	238
12.6 阅读材料	240
习题	241
第五部分 高阶课题	245
第 13 章 正态分布	246
13.1 定义	246
13.1.1 单变量正态分布	246
13.1.2 多元正态分布	247
13.2 符号和参数化形式	248
13.3 线性运算与求和	249
13.3.1 单变量的情形	249
13.3.2 多变量的情形	250
13.4 几何和马氏距离	251
13.5 条件作用	252
13.6 高斯分布的乘积	253
13.7 应用 I: 参数估计	254
13.7.1 最大似然估计	254
13.7.2 贝叶斯参数估计	255
13.8 应用 II: 卡尔曼滤波	256
13.8.1 模型	256

13.8.2	估计	257
13.9	在本章中有用的数学	258
13.9.1	高斯积分	258
13.9.2	特征函数	259
13.9.3	舒尔补 & 矩阵求逆引理	260
13.9.4	向量和矩阵导数	262
	习题	263
第 14 章	EM 算法的基本思想	266
14.1	GMM: 一个工作实例	266
14.1.1	高斯混合模型	266
14.1.2	基于隐变量的诠释	267
14.1.3	假若我们能观测到隐变量, 那会怎样?	268
14.1.4	我们可以模仿先知吗?	269
14.2	EM 算法的非正式描述	270
14.3	期望最大化算法	270
14.3.1	联合非凹的不完整数据对数似然	271
14.3.2	(可能是) 凹的完整数据对数似然	271
14.3.3	通用 EM 的推导	272
14.3.4	E 步和 M 步	274
14.3.5	EM 算法	275
14.3.6	EM 能收敛吗?	275
14.4	EM 用于 GMM	276
14.5	阅读材料	279
	习题	279
第 15 章	卷积神经网络	281
15.1	预备知识	281
15.1.1	张量和向量化	282
15.1.2	向量微积分和链式法则	283
15.2	CNN 概览	283
15.2.1	结构	283
15.2.2	前向运行	285
15.2.3	随机梯度下降	285
15.2.4	误差反向传播	286
15.3	层的输入、输出和符号	287
15.4	ReLU 层	288
15.5	卷积层	290
15.5.1	什么是卷积?	290
15.5.2	为什么要进行卷积?	291

15.5.3	卷积作为矩阵乘法	293
15.5.4	克罗内克积	295
15.5.5	反向传播: 更新参数	296
15.5.6	更高维的指示矩阵	297
15.5.7	反向传播: 为前一层准备监督信号	298
15.5.8	用卷积层实现全连接层	300
15.6	汇合层	301
15.7	案例分析: VGG-16 网络	303
15.7.1	VGG-Verydeep-16	303
15.7.2	感受野	304
15.8	CNN 的亲身体验	305
15.9	阅读材料	305
	习题	305
	参考文献	309
	英文索引	325
	中文索引	332