标题

作者
jerryyanglei@gmail.com
2014年10月25日

标题

摘　　要

摘要的正文

关键词： 关键字

目　录

第一章 章节层次

§1.1 章节层次示例

1. **章节层次示例**
2. **章节层次示例**
3. **章节层次示例**
4. **章节层次示例**

§1.2 章节层次示例

§1.3 章节层次示例

1.3.1 章节层次示例

1.3.2 章节层次示例

1.3.3 章节层次示例

1.3.4 章节层次示例

1. **章节层次示例**
2. **章节层次示例**
3. **章节层次示例**
4. **章节层次示例**

第二章　数学公式示例

§2.1　各种公式实例

1. 指数、指标

x^{2n}　x^{y_1}　$A_{j_{m,n}^{2n}}^{x_i^2}$

2. 分数

$\frac{1}{x+y}$　$\frac{a^2-b^2}{a-b}$

3. 根式

$\sqrt[k]{x^2+y^3}$　$\sqrt[n]{\frac{x^n-y^n}{1+u^{2n}}}$

4. 求和与积分

在正文公式中求和$\sum_{i=1}^{n}$ 积分$\int_a^b$
在显示公式中求和

$$\sum_{i=1}^{n}$$

积分

$$\int_a^b$$

5. 连续点、省略号

ldots . . . cdots $\cdots$ vdots $\vdots$ ddots $\ddots$

§2.2　各种字符

1. 带圈字符

①②③④⑤⑥⑦⑧⑨㊽

2. 希腊字母

alpha α	theta θ	beta β	gamma γ	delta δ	eta η
varepsilon ε	lambda λ	mu μ	sigma σ	rho ρ	xi ξ
pi π	psi ψ	phi ϕ	tau τ	omega ω	nu ν
Gamma Γ	Lambda Λ	Sigma Σ	Psi Ψ	Delta Δ	Omega Ω
Theta Θ	Pi Π	Phi Φ			

大写斜体$\mathit{\Gamma\Pi\Phi}$

3. 数学符号

$\otimes$

$$\frac{\pm\begin{vmatrix} x_1-x_2 & y_1-y_2 & z_1-z_2 \\ l_1 & m_1 & n_1 \\ l_2 & m_2 & n_2 \end{vmatrix}}{\sqrt{\begin{vmatrix} l_1 & m_1 \\ l_2 & m_2 \end{vmatrix}^2+\begin{vmatrix} m_1 & n_1 \\ m_2 & n_2 \end{vmatrix}^2+\begin{vmatrix} n_1 & l_1 \\ n_2 & l_2 \end{vmatrix}^2}}$$

$$\underbrace{a+\overbrace{b+\cdots+y}^{123}+z}_{\alpha\beta\gamma}$$

§2.3 各种矩阵

pmatrix $\begin{pmatrix} 1 & 2 & 3 \\ 6 & 3 & 5 \\ 7 & 98 & 78 \end{pmatrix}$ bmatrix $\begin{bmatrix} 1 & 2 & 3 \\ 6 & 3 & 5 \\ 7 & 98 & 78 \end{bmatrix}$ vmatrix $\begin{vmatrix} 2 & 3 & 4 \\ 6 & 3 & 5 \end{vmatrix}$ Vmatrix $\begin{Vmatrix} 2 & 3 & 4 \\ 6 & 3 & 5 \end{Vmatrix}$

§2.4 用户定义命令(有参数值的命令)—自定义公式

1. 文本中的公式1

$a_i,\cdots,a_j$；$x_1,\cdots,x_n$；$x_i,\cdots,x_j$；$A_{ij},\cdots,A_{lk}$.

2. 文本中的公式2

$x_i, \cdots, x_j$; $A_{ij}, \cdots, A_{lk}$.

3. 居中的公式(不带编号)

$$a_i, \cdots, a_j$$

$$A_{ij}, \cdots, A_{lk}$$

4. 自动编号的单个的公式

$$A_{ij}, \cdots, A_{lk} \tag{2.1}$$

$$A_{ij}, \cdots, A_{lk} \tag{2.2}$$

$$A_{ij}, \cdots, A_{lk} \tag{2.3}$$

5. 自动编号的公式组

$$\begin{aligned} x_i, \cdots, x_j &= x_i, \cdots, x_j \end{aligned} \tag{2.4}$$

$$\begin{aligned} x_i, \cdots, x_j &= x_i, \cdots, x_j \end{aligned} \tag{2.5}$$

第三章 范例

在基于线性回报函数的假设下，线性回报函数基于回报函数的线性组合：

$$R(s) = \sum_{i=1}^{d} w_i \phi_i(s) \tag{3.1}$$

在式3.1中，$\phi_1, ...\phi_d$是确定的状态特征函数，用来描述每一个状态的特征，d为特征的个数；$w_1, ..., w_d$是各个状态特征函数的权值向量，也称为回报权重，这是学徒学习所试图还原的参数，通过$w = [w_1, ..., w_d]$就可以还原出回报函数，进而可以就可以通过加强学习的算法来求解最优策略，从而模拟出专家演示路径。

本文所讨论的2种算法都是建立在IRL框架上的，这一框架的特点在于它假设专家是基于一个能产生最优或者近似最优策略的回报函数来进行演示的。

Ng和Russell 提出逆向增强学习后[1]，Abbeel和Ng将增强学习引入学徒学习[2]，它通过最大化专家演示策略和其他策略的差别，还原出一个能得出和专家演示相似策略的回报函数。策略π对应的值函数可以表示成：

$$V_w(\pi) = w^T \dot{E}[\sum_{t=0}^{\infty} \gamma^t \varphi(s_t)|\pi] \tag{3.2}$$

式3.2中，γ为折扣因子，$\mu = E[\sum_{t=0}^{\infty} \gamma^t \varphi(s_t)|\pi$为特征期望(feature expectation)，根据之前提到的对专家路径所做的最优假设，μ就作为一个代表专家路径的”最优值“，也可以被用作衡量策略之间相似程度的标准。

逆向增强学习通过使执行专家演示策略和次优策略时获得的回报值的差最大来求得各特征之间的权值w，因此，该学习问题可以归结为以下的最优化问题：

$$\max_{\tau, w:||w||_2 \leq 1} \tau, s.t. V_w(\pi_E) \geq V_w(\pi_i) + \tau, i = 1, ..., t-1 \tag{3.3}$$

式3.3中：π_E为专家演示策略，π_i为第i次迭代产生的策略。当前，用于解决该问题比较成熟的算法有：Abbel提出的边际最大算法(Max-margin)，投影法(Projection)[2]，Ziebart提出的基于逆向增强学习的最大熵算法（Maximum Entropy）[3]，还有一些效果一般但速度很快的算法，比如在线学徒学习算法（Online Apprentice Learning）。

为了将该算法应用到实际情况中，Grimes和Rao等人探讨了在不确定环境下的学徒学习系统设计[4]。目前，基于回报函数学习的单专家学徒学习已经被应用到如小型直升机在空中自主完成一系列复杂动作[5]，并且取得了良好的效果。

相比于单专家的学徒学习，多专家的学徒学习具有更明显的现实意义，是一个融合了聚类和学徒学习的问题，这里聚类的意思是说假定归属于同一类的专家演示都是由同一个专家生成的，那么问题就变成了：既要能够推测出每个路径所属的类又要能够为每个类生成回报函数。

致　谢

本模板参考清华大学学术论文LaTeX中文模板。感谢为LaTeX默默奉献的前辈们。

参考文献

[1] Andrew Y. Ng and Stuart Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2000.

[2] Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2004.

[3] Maas Andrew Bagnell J. Andrew and Dey Anind K. Ziebart, Brian D. Maximum entropy inverse reinforcement learning. In *Proceedings of the 23rd National Conference on Artificial Intelligence*, pages 1433–1438, 2008.

[4] Daniel R. Rashid David B. Grimes and Rajesh P. N. Rao. Learning nonparametric models for probabilistic imitation. In *Proceedings of Neural Information Processing Systems*, pages 521–528, 2007.

[5] Morgan Quigley Andrew Y. Ng Pieter Abbeel, Adam Coates. An application of reinforcement learning to aerobatic helicopter flight. In *Proceedings of Neural Information Processing Systems*, pages 1–8, 2007.

[6] K Subramanian ML Littman M Babes, V Marivate. Apprenticeship learning about multiple intentions. In *Proceedings of the 28th International Conference on Machine Learning*, 2011.

[7] Jaedeug Choi and Kee-Eung Kim. Nonparametric bayesian inverse reinforcement learning for multiple reward functions. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States.*, pages 314–322, 2012.

[8] Jaedeug Choi and Kee-Eung Kim. MAP inference for bayesian inverse reinforcement learning. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings*

of a meeting held 12-14 December 2011, Granada, Spain., pages 1989–1997, 2011.

[9] George H. John. When the best move isn't optimal: Q-learning with exploration. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, page 1464, 1994.

[10] Andrew Y. Ng J. Zico Kolter, Mike Rodgers. A complete control architecture for quadruped locomotion over rough terrain. In *IEEE International Conference on Robotics and Automation*, 2008.

[11] Deepak Ramachandran and Eyal Amir. Bayesian inverse reinforcement learning. In *Proceedings of IJCAI*, 2007.

[12] CHEN Shen-yi ZHU Miao-liang JIN Zhuo-jun, QIAN Hui. Survey of apprenticeship learning based on reward function learning. *CAAI Transactions on Intelligent System*, 4(3):209–212, June 2009.

[13] Moore A W kaelbling L P, Littman M L. Reinforcement learning: a survey. *Journals of Artificial Intelligence Research*, 4:237–285, 1996.

[14] Laird N. M. Dempster, A. P. and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

[15] W.K. Hastings. Monte carlo samping methods using markov chains and their applications. *Biometrika*, 57:97–109, 1970.