MLaPP读书会

第一章

机器学习

- 有监督学习
  - 有标注的数据
  - 学习一个函数，由特征得出类别(或者一个值)
  - 分类问题，回归问题

有监督学习

- 基本框架
  - 原始问题数学建模，特征，类别(数值)
  - 训练集，校验集，测试集
  - 假设空间(函数的基本形式，模型选择)
  - hyper参数，可学习参数
  - 损失函数
  - 训练(梯度下降，或者其他方法)
  - 通过校验集，调整hyper参数
  - 重复如上，得到最后模型，通过测试集看效果

有监督学习

- 问题讨论
  - 刚才那个框架对吗？
  - 校验集的作用，和测试集的区别
  - hyper参数是否可以变成可学习参数(正则化?)
  - 假设空间的优劣性，是否存在最优的，万能的假设空间(比如神经网络，顺序问题rnn???)No free lunch theorem
  - ml问题的重点是否在于降低训练难度，而不在于假设空间
  - 其他问题？？？

有监督学习

- 从概率角度来看

$$\hat{y} = \hat{f}(\mathbf{x}) = \operatorname*{argmax}_{c=1}^{C} p(y = c|\mathbf{x}, \mathcal{D})$$

无监督学习

- 没有标注的数据
- The goal is to discover “interesting structure” in the data
- 学习数据之间的关系
  - 聚类
  - 向量化
  - 维度压缩，提取隐含信息
  - Discovering graph structure(没看懂???)
  - Matrix completion(信息平滑???)
  - 其他???

机器学习

- 补充内容
  - Parametric vs non-parametric models
  - non-parametric classifier: K-nearest neighbors
  - The curse of dimensionality(重点理解)
  - 线性回归和Logistic回归(为什么是回归呢？)

$$p(y|\mathbf{x}, \boldsymbol{\theta}) = \mathcal{N}(y|\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}), \sigma^2)$$

$$p(y|\mathbf{x}, \mathbf{w}) = \mathrm{Ber}(y|\mathrm{sigm}(\mathbf{w}^T\mathbf{x}))$$

第二章

- 复习概率论
  - 两种观点
    - frequentist interpretation
    - Bayesian interpretation(重点了解)
  - 重要概念
    - 随机变量，离散的，连续的
    - 分布(实际分布 vs 数学上的分布 vs 语文上的分布)

概率论

- 一些公式(看书)
  - 条件概率
  - 链式法则
  - 贝叶斯公式
- 独立性
  - 条件独立(重点了解)

$$X \perp Y|Z \iff p(X,Y|Z) = p(X|Z)p(Y|Z)$$

  - 作用(更容易建模世界)

概率论

- 这个看不懂

Theorem 2.2.1. $X \perp Y | Z$ *iff there exist function* g *and* h *such that*

$$p(x, y|z) = g(x, z)h(y, z)$$

for all x, y, z *such that* $p(z) > 0$.

概率论

- 连续随机变量
  - cumulative distribution function(累积分布函数)
  - probability density function(概率密度函数)(重点理解)
  - Quantiles(有点绕，但比较好理解)
  - 期望
  - 方差

$$\mathbb{E}[X] \triangleq \sum_{x \in \mathcal{X}} x \; p(x)$$

$$\mathrm{var}[X] \triangleq \mathbb{E}\left[(X-\mu)^2\right]$$

几个重要的离散分布

- 二项分布

$$\mathrm{Bin}(k|n,\theta) \triangleq \binom{n}{k} \theta^k (1-\theta)^{n-k}$$

where

$$\binom{n}{k} \triangleq \frac{n!}{(n-k)!k!}$$

期望 $n\theta$

方差 $\mathrm{var} = n\theta(1-\theta)$

- 伯努利分布(只进行一次实验的二项分布)

$$\mathrm{Ber}(x|\theta) = \begin{cases} \theta & \text{if } x = 1 \\ 1-\theta & \text{if } x = 0 \end{cases}$$

几个重要的离散分布

- 多项分布(通过组合排列来了解)

$$\mathrm{Mu}(\mathbf{x}|n,\boldsymbol{\theta}) \triangleq \binom{n}{x_1 \ldots x_K} \prod_{j=1}^{K} \theta_j^{x_j}$$

where θ_j is the probability that side j shows up, and

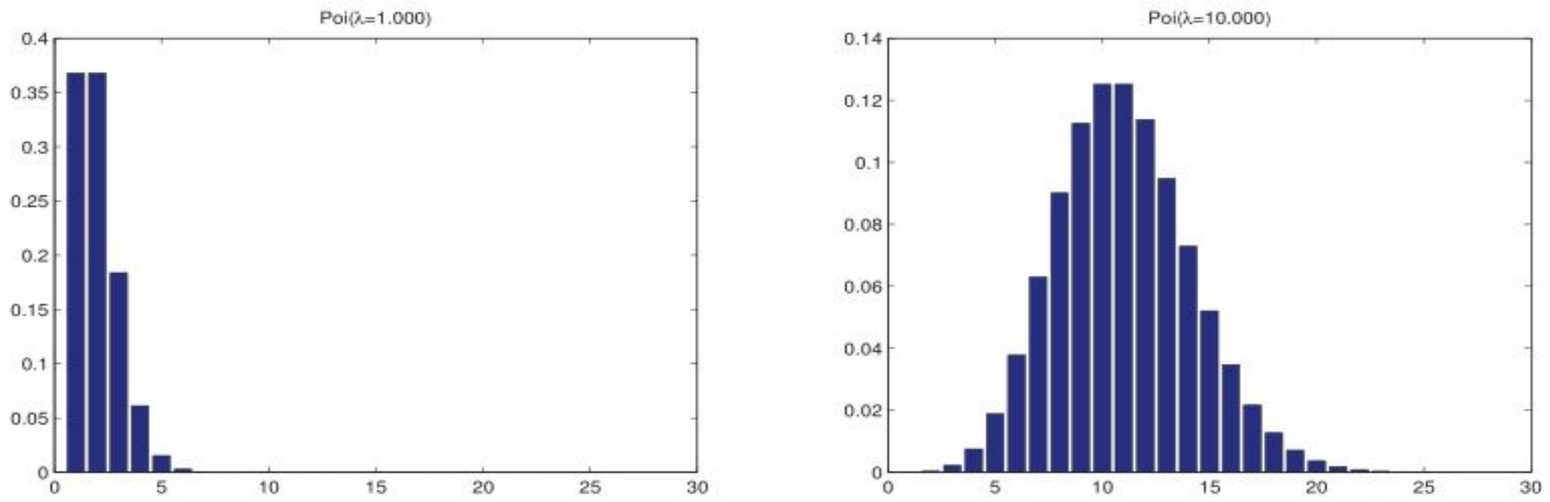
$$\binom{n}{x_1 \ldots x_K} \triangleq \frac{n!}{x_1!x_2! \cdots x_K!}$$

x是一个离散向量，不是一个数，那么期望和方差呢？？？

- Application: DNA sequence motifs(没看懂)

几个重要的离散分布

- 泊松分布(理解) $$\mathrm{Poi}(x|\lambda) = e^{-\lambda}\frac{\lambda^x}{x!} \qquad \lambda > 0$$



(a) (b)

- 泊松分布的期望和方差均为 λ

几个重要的离散分布

$$P(\xi = m) = \frac{\lambda^m}{m!}e^{-\lambda}, (m = 0,1,2,\cdots)$$

$$E\xi = \sum_{m=0}^{\infty} m\frac{\lambda^m}{m!}e^{-\lambda} = \lambda e^{-\lambda}\sum_{m=1}^{\infty}\frac{\lambda^{m-1}}{(m-1)!}$$

$$= \lambda e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} = \lambda e^{-\lambda}e^{\lambda} = \lambda$$

几个重要的离散分布

再计算$E\xi^2$，

$$E\xi^2=\sum_{m=0}^{\infty}m^2\frac{\lambda^m}{m!}e^{-\lambda}=\lambda e^{-\lambda}\sum_{m=1}^{\infty}\frac{m\lambda^{m-1}}{(m-1)!}$$

$$=\lambda e^{-\lambda}\sum_{k=0}^{\infty}\frac{(k+1)\lambda^k}{k!}=\lambda e^{-\lambda}[\sum_{k=1}^{\infty}\frac{\lambda^k}{(k-1)!}+\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}]$$

$$=\lambda e^{-\lambda}(\lambda e^{\lambda}+e^{\lambda})=\lambda(\lambda+1)$$

$$D\xi=E\xi^2-(E\xi)^2=\lambda(\lambda+1)-\lambda^2=\lambda$$

故 $E\xi=D\xi=\lambda$

几个重要的离散分布

- The empirical distribution(经验分布)
  - 应该是通过样本来模拟逼近整体

几个重要的连续分布

- 高斯分布(重点理解)

$$\mathcal{N}(x|\mu,\sigma^2) \triangleq \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

- 本质上就是对称的平方指数函数？？？(对比拉普拉斯分布)
- 这是个方便的建模工具(讨论理解)
- 中心极限定理(重点理解)
- has maximum entropy????

几个重要的连续分布

- 高斯分布的变种
  - Degenerate pdf
  - the Student t distribution (很难，没看懂，重点讨论理解)

几个重要的连续分布

- The Laplace distribution

$$\mathrm{Lap}(x|\mu, b) \triangleq \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right)$$

- 本质上是对称的指数函数(对比高斯分布)
- long tail问题，对outliers的鲁棒性？？？？

- mode是众数

$$\mathrm{mean} = \mu, \ \mathrm{mode} = \mu, \ \mathrm{var} = 2b^2$$

几个重要的连续分布

- 其他一些连续分布(都没看懂，可以讨论)
  - The gamma distribution
  - The beta distribution
  - Pareto distribution
    - 本质上是指数函数？？？
- 技巧:抛开那些用来保证分布sum=1的项，focus on 那些有意义的项

联合分布

- 协方差 covariance
  - 协方差矩阵(对多个变量)

$$\mathrm{cov}\left[X,Y\right] \triangleq \mathbb{E}\left[(X-\mathbb{E}\left[X\right])(Y-\mathbb{E}\left[Y\right])\right]=\mathbb{E}\left[XY\right]-\mathbb{E}\left[X\right]\mathbb{E}\left[Y\right]$$

- 相关系数

  对协方差进行归一化

$$\mathrm{corr}\left[X,Y\right] \triangleq \frac{\mathrm{cov}\left[X,Y\right]}{\sqrt{\mathrm{var}\left[X\right]\mathrm{var}\left[Y\right]}}$$

讨论理解其背后表达的意义

理解“独立则不相关，但不相关不一定独立”

联合分布

- 多个变量联合的一些分布
  - The multivariate Gaussian
  - Multivariate Student t distribution
  - Dirichlet distribution
- 似乎就是从数到向量的推广，但细节真的没看懂，求大神带=_=

随机变量的变换

- Linear transformations
  - 可以方便地计算出变换后的期望和方差，对高斯分布来说足够，但对其他的不行。
- General transformations
  - 离散变量可以枚举累加
  - 连续变量，通过cdf连接新旧变量，再求微得到pdf
  - 推广到多个变量，Jacobian matrix J(没太看懂)

蒙特卡罗近似

- 一种近似的方法来计算X变换后Y的分布
  - 从X分布中采样得到x1,x2,x3….
  - 计算样本变化后的值y1,y2,y3….
  - 再从变换后的样本值y近似分布Y

信息论

- 信息熵(如何理解公式)

$$\mathbb{H}(X) \triangleq -\sum_{k=1}^{K} p(X=k)\log_2 p(X=k)$$

- KL di
  - 用来计算两个分布之间的差异性
  - 交叉熵(用一个分布的编码形式来编码另一个分布所要编码数)
  - uniform distribution满足最大熵
  - 不对称?????

信息论

One way to measure the dissimilarity of two probability distributions, p and q, is known as the **Kullback-Leibler divergence** (**KL divergence**) or **relative entropy**. This is defined as follows:

$$\mathbb{KL}(p||q) \triangleq \sum_{k=1}^{K} p_k \log \frac{p_k}{q_k} \tag{2.110}$$

where the sum gets replaced by an integral for pdfs.[10] We can rewrite this as

$$\mathbb{KL}(p||q) = \sum_k p_k \log p_k - \sum_k p_k \log q_k = -\mathbb{H}(p) + \mathbb{H}(p,q) \tag{2.111}$$

where $\mathbb{H}(p,q)$ is called the **cross entropy**,

$$\mathbb{H}(p,q) \triangleq -\sum_k p_k \log q_k \tag{2.112}$$

信息论

- 互信息Mutual information

$$\mathbb{I}(X;Y) \triangleq \mathbb{KL}\left(p(X,Y) \| p(X)p(Y)\right) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

$$\mathbb{I}(X;Y) = \mathbb{H}(X) - \mathbb{H}(X|Y) = \mathbb{H}(Y) - \mathbb{H}(Y|X)$$

- 连续变量的计算方式(没太看懂)

$$\mathbb{H}(Y|X) = \sum_x p(x)\mathbb{H}(Y|X=x)$$

结束

- thanks