机器学习与深度学习中的数学

SIGAI 雷明

2019.11.12

个人简介

清华大学出版社《机器学习-原理、算法与应用》作者

2009年毕业于清华大学计算机系，研究方向为计算机视觉/机器学习，发表论文数篇

曾就职于百度，任软件工程师/项目经理；zmodo/meshare，任CTO（创业）

2018年创立SIGAI，致力于研发机器视觉、深度强化学习框架，用标准化的算法为各个行业赋能，目前已经应用于物流，商业，国防等领域

内容提要

需要哪些数学知识

微积分

线性代数与矩阵论

概率论

信息论

最优化方法

随机过程

图论

需要哪些数学知识

现状分析

数学是给机器学习、深度学习的初学者和进阶者造成困难的主要原因之一

国内本科数学教学方式、学生学习质量上存在的不足 - 过于抽象，偏重于计算，忽视了对数学思维、建模能力的培养 - 清华大学换用国外线性代数教材事件，如果结合一些具体的例子来讲解会好很多

某些数学知识超出了本科一般理工科专业的范畴 - 矩阵论/矩阵分析，信息论，最优化方法，随机过程，图论

通常情况下，高校、其他机构在教《机器学习》、《深度学习》之前不会为学生把这些数学知识补齐

学生普遍对数学存在一种恐惧心理，数学自信的人只占少部分

究竟需要哪些数学知识？

1.微积分 - 一元函数微积分，多元函数微积分，是整个高等数学的基石

2.线性代数与矩阵论 - 矩阵论本科一般不讲

3.概率论 - 内容基本已经覆盖机器学习的要求

4.信息论 - 一般专业不会讲，如果掌握了概率论，理解起来并不难

5.最优化方法 - 学了这门课的学生非常少，但对机器学习、深度学习非常重要，几乎所有算法归结为求解优化问题

6.随机过程 - 本科一般不学，但在机器学习中经常会使用，如马尔可夫过程，高斯过程，后者应用于贝叶斯优化

7.图论 - 计算机类专业本科通常会学，但没有学谱图理论

第1部分-微积分

为什么需要微积分？

研究函数的性质 - 单调性，凹凸性

求解函数的极值

概率论、信息论、最优化方法等的基础

一元函数微积分

极限 - 微积分的基石，数列的极限，函数的极限

函数的连续性与间断点

上确界与下确界

Lipschitz连续性

导数，一阶导数，高阶导数，导数的计算 - 符号微分，数值微分，自动微分

导数与函数的性质，单调性，极值，凹凸性

泰勒公式

不定积分及其计算

定积分及其计算

广义积分及其计算

常微分方程的基本概念

常系数线性微分方程的求解

基本函数	求导公式
幂函数	$\left(x^{a}\right)' = ax^{a-1}$
指数函数	$\left(e^{x}\right)' = e^{x}$
指数函数	$\left(a^{x}\right)' = a^{x}\ln a$
三角函数	$\left(\sin x\right)' = \cos x$
三角函数	$\left(\cos x\right)' = -\sin x$
三角函数	$\left(\tan x\right)' = \sec^{2} x$
三角函数	$\left(\cot x\right)' = -\csc^{2} x$
对数函数	$\left(\ln x\right)' = \frac{1}{x}$
对数函数	$\left(\log_{a} x\right)' = \frac{1}{\ln a}\frac{1}{x}$
反三角函数	$\left(\arcsin x\right)' = \frac{1}{\sqrt{1-x^{2}}}$
反三角函数	$\left(\arccos x\right)' = -\frac{1}{\sqrt{1-x^{2}}}$
反三角函数	$\left(\arctan x\right)' = \frac{1}{1+x^{2}}$

基本函数的求导公式

基本运算	求导公式
加法	$(f(x)+g(x))' = f'(x)+g'(x)$
减法	$(f(x)-g(x))' = f'(x)-g'(x)$
乘法	$(f(x)g(x))' = f'(x)g(x)+f(x)g'(x)$
除法	$\left(\frac{f(x)}{g(x)}\right)' = \frac{f'(x)g(x)-f(x)g'(x)}{g^2(x)}$

四则运算的求导公式

$$(f(g(x)))' = f'(g)g'(x)$$

复合函数的求导公式

类型	激活函数	导数
sigmoid	$f(x)=\dfrac{1}{1+\exp(-x)}$	$f'(x)=f(x)(1-f(x))$
tanh	$f(x)=\dfrac{\exp(2x)-1}{\exp(2x)+1}$	$f'(x)=1-(f(x))^2$
BNLL	$f(x)=\log(1+\exp(x))$	$f'(x)=\dfrac{\exp(x)}{1+\exp(x)}$
power	$f(x)=(\alpha x+\beta)^{\gamma}$	$f'(x)=\alpha\gamma(\alpha x+\beta)^{\gamma-1}$
ReLU	$f(x)=\max(0,x)$	$\mathrm{ReLU}'(x)=\begin{cases}1 & x>0\\ 0 & x\le 0\end{cases}$
ELU	$f(x)=\begin{cases}x & x>0\\ \alpha(e^x-1) & x\le 0\end{cases}$	$f'(x)=\begin{cases}1 & x>0\\ f(x)+\alpha & x\le 0\end{cases}$
PReLU	$f(x)=\begin{cases}x & x>0\\ \alpha x & x\le 0\end{cases}$	$f'(x)=\begin{cases}1 & x>0\\ \alpha & x\le 0\end{cases}$
exp	$f(x)=\gamma^{\alpha x+\beta}$	$f'(x)=\gamma^{\alpha x+\beta}(\ln\gamma)\alpha$
log	$f(x)=\log_{\gamma}(\alpha x+\beta)$	$f'(x)=\dfrac{\alpha}{\ln\gamma}\dfrac{1}{\alpha x+\beta}$

激活函数的导数

类型	损失函数	导数
欧氏距离	$L=\frac{1}{2n}\sum_{i=1}^{n}\left\Vert \hat{\mathrm{y}}_i-\mathrm{y}_i\right\Vert^2$	$\nabla_{\hat{\mathrm{y}}}L=\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mathrm{y}}_i-\mathrm{y}_i\right)$ $\nabla_{\mathrm{y}}L=-\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mathrm{y}}_i-\mathrm{y}_i\right)$
softmax 交叉熵	$y_i^*=\frac{\exp\left(x_i\right)}{\sum_{j=1}^{k}\exp\left(x_j\right)}$ $L=-\frac{1}{n}\sum_{i=1}^{n}\mathrm{y}_i^{\mathrm{T}}\log \mathrm{y}_i^*$	$\nabla_{\mathrm{x}}L=\frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{y}_i^*-\mathrm{y}_i\right)$
sigmoid 交叉熵	$\hat{p}_i=\frac{1}{1+e^{-x_i}}$ $L=-\frac{1}{n}\sum_{i=1}^{n}\left(p_i\log\left(\hat{p}_i\right)+\left(1-p_i\right)\log\left(1-\hat{p}_i\right)\right)$	$\frac{\partial L}{\partial x_i}=\frac{1}{n}\left(\hat{p}_i-p_i\right)$
对比损失	$d_i=\left\Vert \mathrm{a}_i-\mathrm{b}_i\right\Vert_2$ $L=\frac{1}{2n}\sum_{i=1}^{n}\left(y_i d_i^2+\left(1-y_i\right)\max\left(m-d_i^2,0\right)\right)$	$\nabla L_{\mathrm{a}_i}=\frac{1}{n}\left(y_i\left(\mathrm{a}_i-\mathrm{b}_i\right)\right)$ $\nabla L_{\mathrm{b}_i}=-\frac{1}{n}\left(y_i\left(\mathrm{a}_i-\mathrm{b}_i\right)\right)$
合页损失	$L=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}\left(\left\Vert\max\left(0,1-\delta\left\{l_i=j\right\}t_{ij}\right)\right\Vert^p\right)$	$\frac{\partial L}{\partial t_{ij}}=\begin{cases}-\frac{1}{n}\delta\left\{l_i=j\right\}\\ 0\end{cases}$
信息增益	$L=-\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}h_{ij}\log\left(\hat{p}_{ij}\right)$	$\frac{\partial L}{\partial \hat{p}_{ij}}=-\frac{1}{n}\frac{h_{ij}}{\hat{p}_{ij}}$
多项式 logistic	$L=-\sum_{i=1}^{k}I\left(y=i\right)\log\left(p_i\right)$	$\frac{\partial L}{\partial p_i}=-\sum_{i=1}^{k}I\left(y=i\right)\frac{1}{p_i}$

$$f(x)=f(a)+\frac{f^{'}(a)}{1!}(x-a)+\frac{1}{2}f^{''}(a)(x-a)^{2}+\ldots+\frac{1}{n!}f^{(n)}(a)(x-a)^{n}+R_{n}(x)$$

一元函数的泰勒公式-连接一元函数微分学各知识点的桥梁

多元函数微积分

偏导数的定义与计算

梯度的定义与性质

方向导数的定义与性质

高阶偏导数的计算

链式法则 - 熟练计算多元函数的偏导数

雅克比矩阵 - 链式法则的矩阵形式

Hessian矩阵与多元函数的极值，凹凸性

向量与矩阵求导公式

多元函数的泰勒公式

重积分　二重积分，三重积分，n重积分，多重积分的坐标变换

偏微分方程的基本概念

$$z = f\left(y_1, ..., y_m\right)$$

$$y_j = g_j\left(x_1, ..., x_n\right), j = 1, ..., m$$

$$\begin{bmatrix} \frac{\partial z}{\partial x_1} \\ ... \\ \frac{\partial z}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{m} \frac{\partial z}{\partial y_j} \frac{\partial y_j}{\partial x_1} \\ ... \\ \sum_{j=1}^{m} \frac{\partial z}{\partial y_j} \frac{\partial y_j}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & ... & \frac{\partial y_m}{\partial x_1} \\ ... & ... & ... \\ \frac{\partial y_1}{\partial x_n} & ... & \frac{\partial y_m}{\partial x_n} \end{bmatrix} \begin{bmatrix} \frac{\partial z}{\partial y_1} \\ ... \\ \frac{\partial z}{\partial y_m} \end{bmatrix}$$

$$= \left(\frac{\partial \mathbf{y}}{\partial \mathbf{x}}\right)^{\mathrm{T}} \begin{bmatrix} \frac{\partial z}{\partial y_1} \\ ... \\ \frac{\partial z}{\partial y_m} \end{bmatrix}$$

链式法则的矩阵形式

函数	求导公式
$y = \mathbf{w}^{\mathrm{T}}\mathbf{x}$	$\nabla \mathbf{w}^{\mathrm{T}}\mathbf{x} = \mathbf{w}$
$y = \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$	$\nabla \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} = \left(\mathbf{A} + \mathbf{A}^{\mathrm{T}}\right)\mathbf{x}$
$y = \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$	$\nabla^2 \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} = \mathbf{A} + \mathbf{A}^{\mathrm{T}}$

重要的向量和矩阵求导公式

$$f(\mathbf{x}) = f(\mathbf{a}) + \left(\nabla f(\mathbf{a})\right)^{\mathrm{T}} (\mathbf{x}-\mathbf{a}) + \frac{1}{2}(\mathbf{x}-\mathbf{a})^{\mathrm{T}} \mathbf{H} (\mathbf{x}-\mathbf{a}) + o\left(\|\mathbf{x}-\mathbf{a}\|^{2}\right)$$

多元函数的泰勒公式-连接多元函数微分学各知识点的桥梁

第2部分-线性代数与矩阵论

为什么需要线性代数？

机器学习算法的输入、输出、中间结果，通常为向量，矩阵，张量

简化问题的表达

与微积分结合，研究多元函数的性质，也是概率论中随机向量的基础

在图论中亦有应用 - 图的拉普拉斯矩阵

在随机过程中同样有应用 - 状态转移矩阵

向量的定义与基本运算，向量的范数

线性相关性

向量空间

矩阵的定义及其运算

矩阵的范数

线性变换

行列式的定义与计算

线性方程组 齐次，非齐次

特征值与特征值向量

广义特征值

Rayleigh商

谱与条件数

二次型与标准型

Cholesky分解

特征值分解

奇异值分解

$$\mathbf{u}^{(l)} = \mathbf{W}^{(l)}\mathbf{x}^{(l-1)} + \mathbf{b}^{(l)}$$

$$\mathbf{x}^{(l)} = f\left(\mathbf{u}^{(l)}\right)$$

正向传播算法

$$\boldsymbol{\delta}^{(l)} = \left(\mathbf{W}^{(l+1)}\right)^{\mathrm{T}} \boldsymbol{\delta}^{(l+1)} \odot f'\left(\mathbf{u}^{(l)}\right)$$

$$\nabla_{\mathbf{w}^{(l)}} L = \boldsymbol{\delta}^{(l)} \left(\mathbf{x}^{(l-1)}\right)^{\mathrm{T}}$$

$$\nabla_{\mathbf{b}^{(l)}} L = \boldsymbol{\delta}^{(l)}$$

反向传播算法

$$\mathbf{S}\mathbf{e} = \lambda\mathbf{e}$$

主成分分析

$$\mathbf{L}\mathbf{f} = \lambda\mathbf{D}\mathbf{f}$$

拉普拉斯特征映射

$$\mathbf{X}\mathbf{L}\mathbf{X}^{\mathrm{T}}\mathbf{a} = \lambda\mathbf{X}\mathbf{D}\mathbf{X}^{\mathrm{T}}\mathbf{a}$$

局部保持投影

第3部分-概率论

为什么需要概率论？

将机器学习算法的输入、输出看作随机变量/向量，用概率论的观点进行建模

对不确定性进行建模

挖掘变量之间的概率依赖关系

随机算法 - 蒙特卡洛算法，遗传算法

随机数生成 - 基本随机数生成，采样算法

随机事件与概率

条件概率

全概率公式

贝叶斯公式

条件独立

离散型随机变量

连续型随机变量

数学期望与方差，标准差

Jesen不等式

Hoeffding不等式

常用概率分布 均匀分布，伯努利分布，二项分布，多项分布，正态分布，狄拉克分布，t分布

随机变量函数

逆变换算法

离散型随机向量

连续型随机向量

联合期望

协方差

常用概率分布 均匀分布，正态分布

分布变换

极限定理 切比雪夫不等式，大数定律，中心极限定理

参数估计 最大似然估计，最大后验概率估计，贝叶斯估计，核密度估计

随机算法 基本随机数生成，遗传算法，蒙特卡洛算法

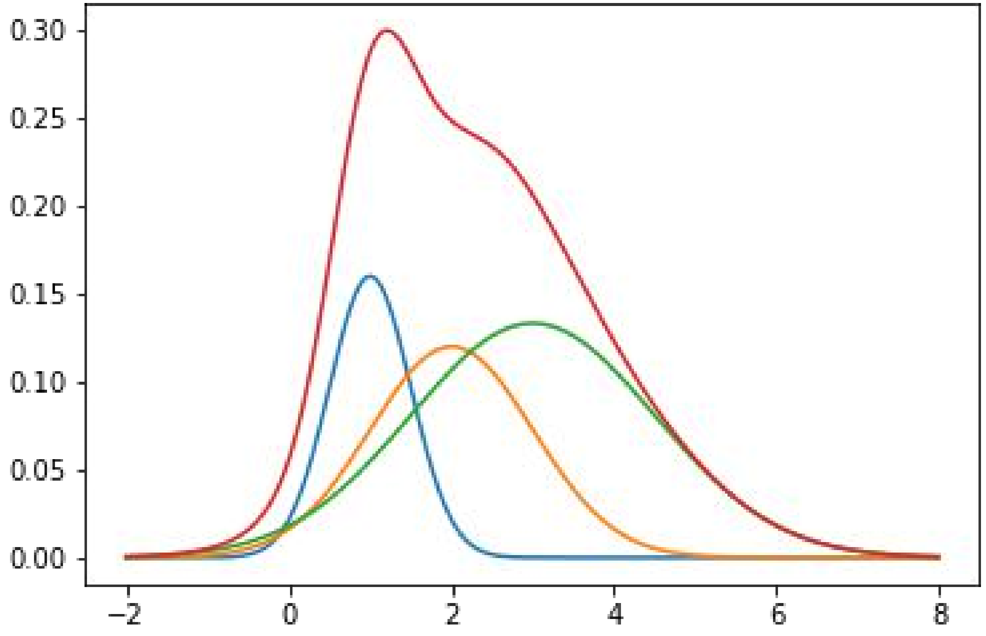
采样算法 拒绝采样，重要性采样

$$p\left(y \mid \mathbf{x}\right)=\frac{p\left(\mathbf{x} \mid y\right)p\left(y\right)}{p\left(\mathbf{x}\right)}$$

$$\arg\max_{y}\, p\left(\mathbf{x} \mid y\right)p\left(y\right)$$

贝叶斯分类器

$$p(\mathbf{x}) = \sum_{i=1}^{k} w_i N(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$



高斯混合模型

第4部分-信息论

香浓熵

交叉熵

KL散度

JS散度

联合熵

互信息

条件熵

$$\prod_{i=1}^{l}\left(\prod_{j=1}^{k}\left(\frac{\exp\left(\boldsymbol{\theta}_j^{\mathrm{T}}\mathbf{x}_i\right)}{\sum_{t=1}^{k}\exp\left(\boldsymbol{\theta}_t^{\mathrm{T}}\mathbf{x}_i\right)}\right)^{y_{ij}}\right)$$

$$\sum_{i=1}^{l}\sum_{j=1}^{k}\left(y_{ij}\ln\frac{\exp\left(\boldsymbol{\theta}_j^{\mathrm{T}}\mathbf{x}_i\right)}{\sum_{t=1}^{k}\exp\left(\boldsymbol{\theta}_t^{\mathrm{T}}\mathbf{x}_i\right)}\right)$$

softmax回归

$$p_{j|i} = \frac{\exp\left(-\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\left\|\mathbf{x}_i - \mathbf{x}_k\right\|^2 / 2\sigma_i^2\right)}$$

$$q_{j|i} = \frac{\exp\left(-\left\|\mathbf{y}_i - \mathbf{y}_j\right\|^2\right)}{\sum_{k \neq i} \exp\left(-\left\|\mathbf{y}_i - \mathbf{y}_k\right\|^2\right)}$$

$$L\left(\mathbf{y}_i\right) = \sum_{i=1}^{l} KL\left(P_i \middle| Q_i\right) = \sum_{i=1}^{l}\sum_{j=1}^{l} p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}$$

流形学习-SNE降维

$$\min_G \max_D V(D,G) = \mathrm{E}_{\mathbf{x}\sim p_{data}(\mathbf{x})}\left[\ln D(\mathbf{x})\right] + \mathrm{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}\left[\ln\left(1 - D\left(G(\mathbf{z})\right)\right)\right]$$

$$\begin{aligned}
C(G) &= -\ln 4 + \ln 4 + \mathrm{E}_{\mathbf{x}\sim p_{data}(\mathbf{x})}\left[\ln \frac{p_{data}(\mathbf{x})}{p_{data}(\mathbf{x}) + p_g(\mathbf{x})}\right] + \mathrm{E}_{\mathbf{z}\sim p_g(\mathbf{z})}\left[\ln \frac{p_g(\mathbf{x})}{p_{data}(\mathbf{x}) + p_g(\mathbf{x})}\right] \\
&= -\ln 4 + \mathrm{E}_{\mathbf{x}\sim p_{data}(\mathbf{x})}\left[\ln \frac{2p_{data}(\mathbf{x})}{p_{data}(\mathbf{x}) + p_g(\mathbf{x})}\right] + \mathrm{E}_{\mathbf{z}\sim p_g(\mathbf{z})}\left[\ln \frac{2p_g(\mathbf{x})}{p_{data}(\mathbf{x}) + p_g(\mathbf{x})}\right] \\
&= -\ln 4 + D_{\mathrm{KL}}\left(p_{data} \left\| \frac{p_{data} + p_g}{2}\right.\right) + D_{\mathrm{KL}}\left(p_g \left\| \frac{p_{data} + p_g}{2}\right.\right) \\
&= -\ln 4 + 2D_{\mathrm{JS}}\left(p_{data} \| p_g\right)
\end{aligned}$$

生成对抗网络

第5部分-最优化方法

基本概念 问题定义，迭代法的基本思想

梯度下降法

最速下降法

梯度下降法的各种改进 AdaGrad，AdaDelta，Adam

随机梯度下降法

牛顿法

拟牛顿法 DFP，BFGS，L-BFGS

分治法 坐标下降法，分阶段优化

凸优化 定义与性质

拉格朗日乘数法

拉格朗日对偶

KKT条件

多目标优化 基本概念，求解算法

泛函与变分

Euler-Lagrange方程

$$L(W)=\frac{1}{2m}\sum_{i=1}^{m}\left\|h(\mathbf{x}_i)-\mathbf{y}_i\right\|^2$$

$$W_{t+1}=W_t-\eta\nabla_W L(W_t)$$

神经网络的训练

$$\max_{m} \mathrm{ACC}(m) \times \left[\frac{\mathrm{LAT}(m)}{T}\right]^{w}$$

$$w = \begin{cases} \alpha, \mathrm{LAT}(m) \leq T \\ \beta, \mathrm{LAT}(m) > T \end{cases}$$

多目标神经结构搜索

$$F[y]=\int_a^b\sqrt{1+y'^2}\,dx$$

$$\frac{d}{dx}\frac{y'}{\sqrt{1+y'^2}}=0$$

$$y(x)=\frac{C}{\sqrt{1-C^2}}x+C'$$

证明两点之间直线最短

第6部分-随机过程

马尔可夫性

马尔可夫链

平稳分布

细致平稳条件

马尔可夫链采样算法

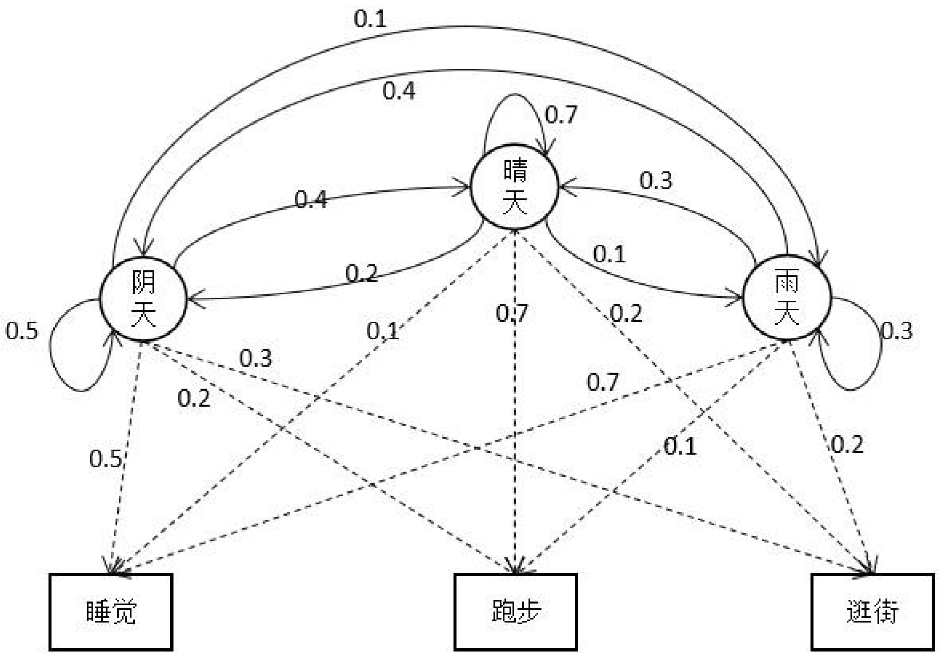
Metropolis-Hastings算法

Gibbs采样

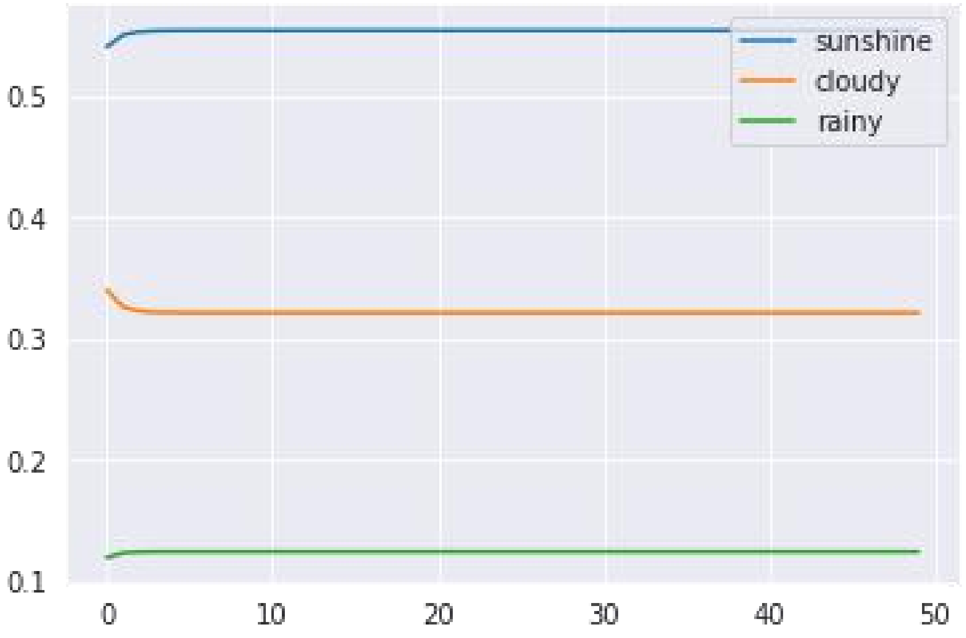
高斯过程

高斯过程回归

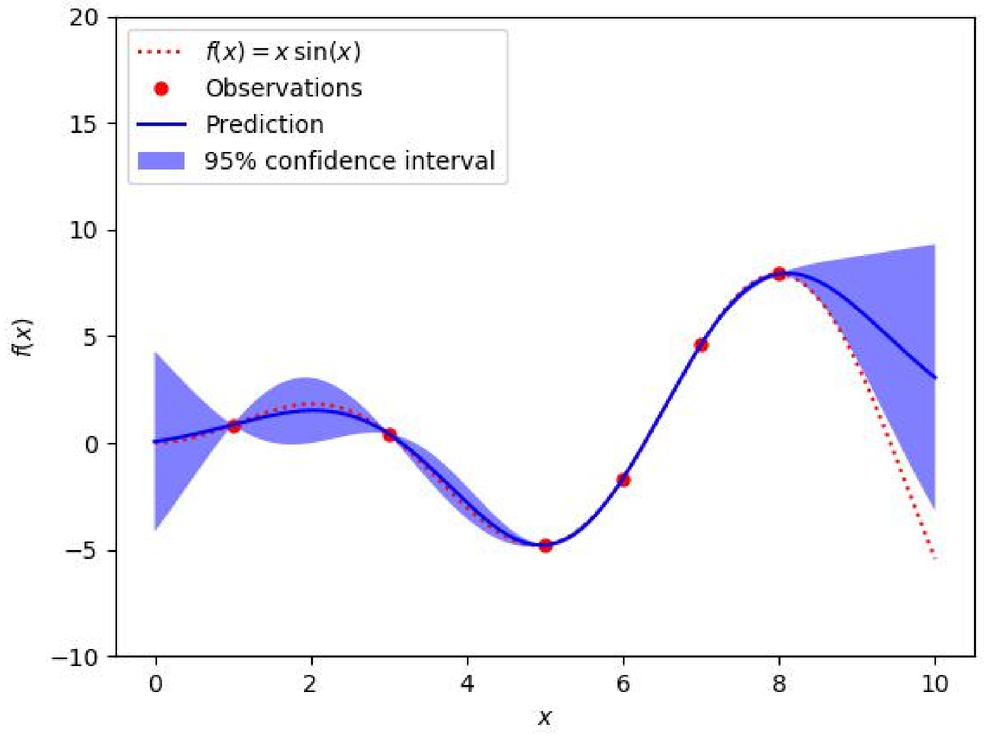
贝叶斯优化



隐马尔可夫模型



平稳分布



$f(x) = x\sin(x)$
Observations
Prediction
95% confidence interval
20
15
10
5
0
−5
−10
$f(x)$
0
2
4
6
8
10
x

高斯过程

第7部分-图论

基本概念

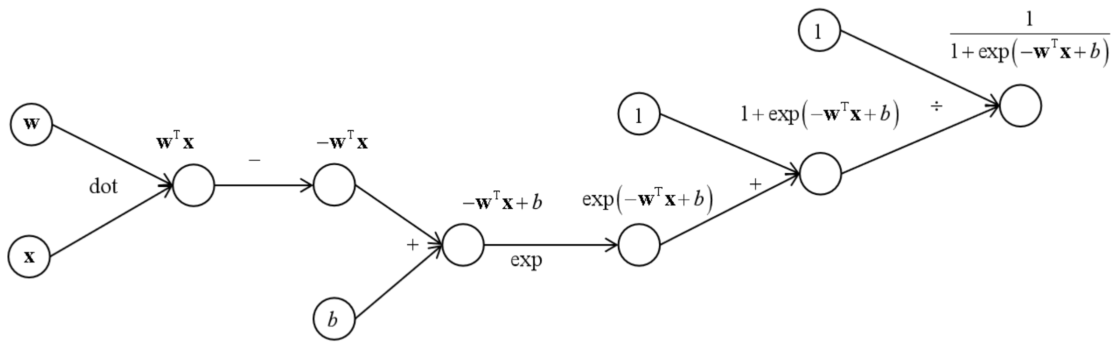
图的矩阵表示

特殊的图 联通图，二部图，有向无环图

Dijkstra算法

拉普拉斯矩阵

归一化拉普拉斯矩阵



logistic回归的计算图

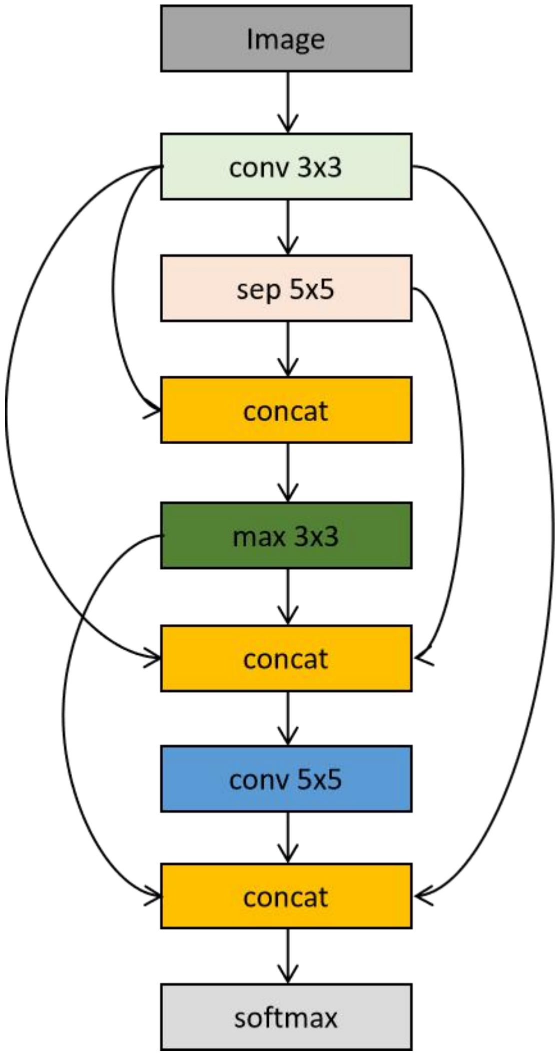
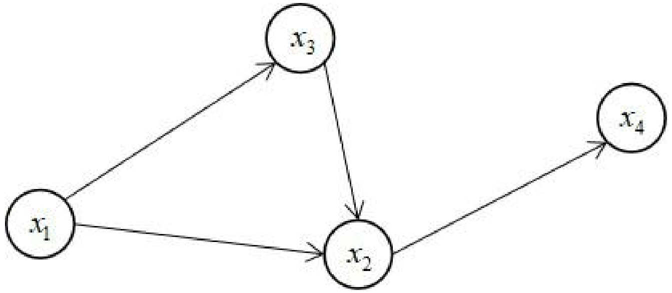


Image
conv 3x3
sep 5x5
concat
max 3x3
concat
conv 5x5
concat
softmax

神经网络的拓扑结构图



概率图模型