Lecture2 Mathematical Background

1. 数学符号定义

数和数组

符号	解释
a	标量 (整数或实数)
$\boldsymbol{a}$	向量
$\boldsymbol{A}$	矩阵
$\boldsymbol{\mathsf{A}}$	张量
$\boldsymbol{I}_n$	n 行 n 列的单位矩阵
$\boldsymbol{I}$	维度蕴含于上下文的单位矩阵
$\boldsymbol{e}^{(i)}$	标准基向量 $[0,\dots,0,1,0,\dots,0]$，其中索引 i 处值为 1
$diag(\boldsymbol{a})$	对角方阵，期中对角元素由 $\boldsymbol{a}$ 给定
a	标量随机变量
$\mathbf{a}$	向量随机变量
$\mathbf{A}$	矩阵随机变量

集合和图

符号	解释
$\mathbb{A}$	集合
$\mathbb{R}$	实数集
$\{0,1\}$	包含 0 和 1 的集合
$\{0,1,\dots,n\}$	包含 0 到 n 之间所有整数的集合
$[a,b]$	包含 a 和 b 的实数区间
$(a,b]$	不包含 a 但包含 b 的实数区间
$\mathbb{A}\backslash\mathbb{B}$	差集，即其元素包含于 $\mathbb{A}$ 但不包含于 $\mathbb{B}$
$\mathcal{G}$	图
$Pa_{\mathcal{G}}(\mathrm{x}_i)$	图 $\mathcal{G}$ 中 x_i 的父节点

索引

符号	解释
a_i	向量 $\boldsymbol{a}$ 的第 i 格元素，其中索引从 1 开始
a_{-i}	除了第 i 个元素，$\boldsymbol{a}$ 的所有元素
$A_{i,j}$	矩阵 $\boldsymbol{A}$ 的 i,j 元素
$A_{i,:}$	矩阵 $\boldsymbol{A}$ 的第 i 行
$A_{:,j}$	矩阵 $\boldsymbol{A}$ 的第 j 列
$\mathsf{A}_{i,j,k}$	3 维张量 $\mathbf{\mathsf{A}}$ 的 (i,j,k) 元素
$\mathbf{\mathsf{A}}_{:,:,i}$	3 维张量的 2 维切片
a_i	随机向量 $\mathbf{a}$ 的第 i 个元素

线性代数

符号	解释
$\boldsymbol{A}^T$	矩阵 $\boldsymbol{A}$ 的转置
$\boldsymbol{A}^+$	矩阵 $\boldsymbol{A}$ 的 Moore-Penrose 伪逆
$\boldsymbol{A} \odot \boldsymbol{B}$	$\boldsymbol{A}$ 和 $\boldsymbol{B}$ 的逐元素乘积（Hadamard 乘积）
$det(\boldsymbol{A})$	$\boldsymbol{A}$ 的行列式

微积分

符号	解释
$\frac{dy}{dx}$	y 关于 x 的导数
$\frac{\partial y}{\partial x}$	y 关于 x 的偏导
$\nabla_{\boldsymbol{x}} y$	y 关于 $\boldsymbol{x}$ 的偏导
$\nabla_{\boldsymbol{X}} y$	y 关于 $\boldsymbol{X}$ 的矩阵倒数
$\nabla_{\mathbf{X}} y$	y 关于 $\mathbf{X}$ 的求导后的张量
$\frac{\partial f}{\partial \boldsymbol{x}}$	$f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ 的 Jacobian 矩阵 $\boldsymbol{J} \in R^{m \times n}$
$\nabla_x^2 f(\boldsymbol{x})$ or $\boldsymbol{H}(f)(\boldsymbol{x})$	f 在点 $\boldsymbol{x}$ 处的 Hessian 矩阵
$\int f(\boldsymbol{x}) d\boldsymbol{x}$	$\boldsymbol{x}$ 整个域上的定积分
$\int_{\mathbb{S}} f(\boldsymbol{x}) d\boldsymbol{x}$	集合 $\mathbb{S}$ 上关于 $\boldsymbol{x}$ 的定积分

概率和信息论

符号	解释
$\mathrm{a} \perp \mathrm{b}$	y 关于 x 的导数
$\mathrm{a} \perp \mathrm{b} \mid \mathrm{c}$	给定 c 后条件独立
$P(\mathrm{a})$	离散变量上的概率分布
$p(\mathrm{a})$	连续变量（或变量类型未指定）上的概率分布
$\mathrm{a} \sim P$	具有分布 P 的随机变量 a
$\mathbb{E}_{\mathrm{x} \sim P}[f(x)]$ or $\mathbb{E} f(x)$	$f(x)$ 关于 $P(x)$ 的期望
$\mathrm{Var}(f(x))$	$f(x)$ 在分布 $P(x)$ 下的方差
$\mathrm{Cov}(f(x), g(x))$	$f(x)$ 和 $g(x)$ 在分布 $P(x)$ 下的协方差
$H(\mathrm{x})$	随机变量 x 的香农熵
$D_{\mathrm{KL}}(P \\| Q)$	P 和 Q 的 KL 散度
$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$	均值为 $\boldsymbol{\mu}$，协方差为 $\boldsymbol{\Sigma}$，$\boldsymbol{x}$ 上的高斯分布

函数

符号	解释
$f : \mathbb{A} \to \mathbb{B}$	定义域 $\mathbb{A}$ 值域为 $\mathbb{B}$ 的函数 f
$f \circ g$	f 和 g 的组合
$f(\boldsymbol{x};\boldsymbol{\theta})$	由于 $\boldsymbol{\theta}$ 的参数化，关于 $\boldsymbol{x}$ 的函数 有时候为了简化表示，忽略 $\boldsymbol{\theta}$，记为 $f(\boldsymbol{x})$
$\log x$	x 的自然对数
$\sigma(x)$	Logistic sigmoid $\frac{1}{1+exp(-x)}$
$\zeta(x)$	Softplus $log(1+exp(x))$
$\|\boldsymbol{x}\|_p$	$\boldsymbol{x}$ 的 L^p 范数
$\|\boldsymbol{x}\|$	$\boldsymbol{x}$ 的 L^2 范数
x^+	x 的正数部分，即 $\max(x,0)$
$\mathbf{1}_{\text{condition}}$	如果条件为真则为 1，否则为 0

有时候我们使用函数 f，它的参数是一个标量，但应用到一个向量、矩阵或张量：$f(\boldsymbol{x}), f(\boldsymbol{X})$ 或者 $f(\mathbf{X})$，这表示逐元素地将 f 应用于这些向量、矩阵或张量

- 例如 $\mathbf{C} = \sigma(\mathbf{X})$，则对于所有合法的 i, j, k，$C_{i,j,k} = \sigma(X_{i,j,k})$

数据集和分布

符号	解释
p_{data}	数据生成分布
$\hat{p}_{data}$	由训练集定义的经验分布
$\mathbb{X}$	训练样本的集合
$\boldsymbol{x}^{(i)}$	数据集的第 i 个样本（输入）
y^ior $\boldsymbol{y}^{(i)}$	监督学习中于 $\boldsymbol{x}^{(i)}$ 关联的目标
$\boldsymbol{X}$	$m \times n$ 的矩阵，其中行 $\boldsymbol{X}_{i,:}$为输入样本 $\boldsymbol{x}^{(i)}$

2. 线性代数 Linear algebra

基本概念

标量 Scalars

- 标量是一个单一的数
- 整数、实数、有理数等等
- 我们用斜体表示它：a, n, x

向量 Vectors

- 向量是一组数字的一维数组

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

矩阵 Matrices

矩阵 $\mathbf{A} \in \mathbb{R}^{m \times n}$，是一个二维的数组

$$\mathbf{A} = \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,2} & A_{2,2} \end{bmatrix}$$

张量 Tensors

张量是一组数字，可能有

- 0 维：标量
- 1 维：向量
- 2 维：矩阵
- 或者更高的维度

矩阵操作

矩阵转置 Matrix Transpose

$$\boldsymbol{A} = \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \\ A_{3,1} & A_{3,2} \end{bmatrix} \Rightarrow \boldsymbol{A}^\top = \begin{bmatrix} A_{1,1} & A_{2,1} & A_{3,1} \\ A_{1,2} & A_{2,2} & A_{3,2} \end{bmatrix}$$

- $(A^T)_{i,j} = A_{j,i}$
- $(AB)^T = B^T A^T$

矩阵加法/减法 Matrix addition and subtraction

只有当两个矩阵 $\mathbf{A}$ 和 $\mathbf{B}$ 的大小相等时，可以对进行加减运算

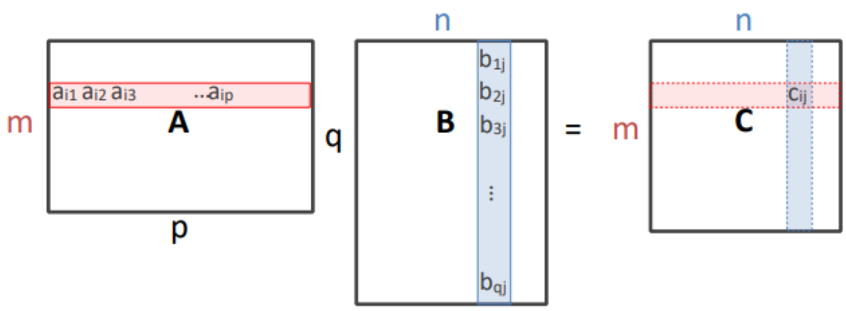
$$\mathbf{A}+\mathbf{B}=\begin{bmatrix}1 & 2\\ 3 & 4\\ 5 & 6\end{bmatrix}+\begin{bmatrix}-1 & 2\\ 4 & -8\\ -16 & 32\end{bmatrix}=\begin{bmatrix}0 & 4\\ 7 & -4\\ -11 & 38\end{bmatrix}=\mathbf{C}$$

标量乘法 Scalar multiplication

一个矩阵与一个标量 t 相乘的结果是得到一个大小相同的矩阵，它的每一项都乘以 t

$$\mathbf{B}=t\mathbf{A}=t\begin{bmatrix}1 & 2\\ 3 & 4\\ 5 & 6\end{bmatrix}=\begin{bmatrix}1t & 2t\\ 3t & 3t\\ 5t & 6t\end{bmatrix}=\mathbf{C}$$

矩阵乘法 Matrix multiplication



$$\mathbf{C}=\mathbf{A}\times\mathbf{B}=>c_{ij}=\sum_{k=1}^{p}a_{ik}b_{kj}=a_{i1}b_{ij}+a_{i2}b_{2j}+\cdots+a_{ip}b_{pj}$$

矩阵的性质

单位矩阵 Identity matrix

单位矩阵是对角线上为 1，其他地方为 0 的对角矩阵

$$\mathbf{I}=\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}$$

如果 $\mathbf{A}$ 是一个 $m\times n$ 维的矩阵，那么

$$\mathbf{A}\mathbf{I}_n=\mathbf{A}\qquad\mathbf{I}_m\mathbf{A}=\mathbf{A}$$

矩阵的逆 Inverse of a matrix

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_n$$

对称矩阵 Symmetric Matrix

$$\mathbf{A} = \mathbf{A}^T$$

正交矩阵 Orthogonal matrix

$$\boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{I}$$
$$\boldsymbol{A}^{-1} = \boldsymbol{A}^\top$$

矩阵的迹 Trace of a matrix

一个矩阵的迹等于它的对角元素的和

$$Tr(\mathbf{A}) = \sum_i \mathbf{A}_{i,i}$$
$$Tr(\mathbf{ABC}) = Tr(\mathbf{CAB}) = Tr(\mathbf{BCA})$$

线性方程组 Linear systems of equations

一个线性方程组 $\mathbf{Ax} = \mathbf{b}$ 可以是

- 没有解
- 很多个解
- 只有一个解（仅当 $\mathbf{A}$ 是可逆的时候）

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$$
$$\boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b}$$
$$\boldsymbol{I}_n\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b}$$

范数 Norms

测量一个向量有多“大”的函数

类似于 0 到由向量表示的点之间的距离

L_p 范数：

$$||\boldsymbol{x}||_p = \left(\sum_i |x_i|^p\right)^{\frac{1}{p}}$$

- L_1 范数：$||\mathbf{x}||_1 = \sum_i |x_i|$
- 最大的范数 $||\mathbf{x}||_\infty = max_i|x_i|$

范数的计算 $f(x)$ 满足如下性质

- $f(\mathbf{x}) = 0 \rightarrow \mathbf{x} = \mathbf{0}$
- $f(\mathbf{x}+\mathbf{y}) \leq f(\mathbf{x}) + f(\mathbf{y})$（三角不等性）
- $\forall \alpha \in \mathbb{R}, f(\alpha\mathbf{x}) = |\alpha| f(\mathbf{x})$

矩阵的分解

矩阵特征值分解 Matrix eigendecomposition

方阵 $\mathbf{A}$ 的**特征向量 eigenvector** 是指与 A 相乘后等于对该向量进行缩放的非零向量 $\mathbf{v}$

$$\boldsymbol{A}\boldsymbol{v} = \lambda\boldsymbol{v}$$

- 标量 λ 被称为这个特征向量对应的**特征值(eigenvalue)**
- 如果 $\mathbf{v}$ 是 $\mathbf{A}$ 的特征向量，那么任何缩放后的向量 $s\mathbf{v}(s \in R，s \neq 0)$ 也是 $\mathbf{A}$ 的特征向量
- 通常我们只考虑单位特征向量

对方阵 $\mathbf{A}$ 的特征值分解

- 假设矩阵 $\mathbf{A}$ 有 n 个线性无关的特征向量 $\{v(1), \ldots, v(n)\}$，对应着特征值 $\{\lambda_1, \ldots, \lambda_n\}$
- 将特征向量连接成一个矩阵，使得每一列是一个特征向量 $\boldsymbol{V} = \left[\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(n)}\right]$
- 也可以将特征值连接成一个向量 $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_n]^\top$

A 的**特征分解(eigendecomposition)** 可以记为

$$\mathbf{A} = \mathbf{V} diag(\lambda)\mathbf{V}^{-1}$$

- 不是每一个矩阵都可以分解成特征值和特征向量
- 在某些情况下，特征分解存在，但是会涉及复数而非实数

每个**实数对称矩阵**都可以分解成实特征向量和实特征值

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T$$

- 其中 $\mathbf{Q}$ 是 $\mathbf{A}$ 的特征向量组成的正交矩阵

- $\mathbf{\Lambda}$ 是对角矩阵

矩阵奇异值分解 Singular value decomposition

奇异值分解是一个能适用于**任意的矩阵**（不一定是方阵）的一种分解的方法

$$\boldsymbol{A}=\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\top}$$

每个实数矩阵都有一个奇异值分解，但不一定都有特征分解

矩阵 $\mathbf{U}$ 和 $\mathbf{V}$ 都定义为**正交矩阵**，而矩阵 $\mathbf{D}$ 定义为**对角矩阵**

- $\mathbf{U}$ 是 $\mathbf{A}\mathbf{A}^T$的特征向量 **左奇异向量(left singular vector)**
- $\mathbf{V}$ 是 $\mathbf{A}^T\mathbf{A}$的特征向量 **右奇异向量(right singular vector)**
- $\mathbf{D}$ 是 $\mathbf{A}^T\mathbf{A}$ 的特征向量平方根 **奇异值(singular values)**

3. 概率论 Probability

概率质量函数 Probability mass function

P 的定义域必须是 x 的所有可能状态的集合

$$\forall x\in \mathrm{x}, 0\leq P(x)\leq 1$$

- 不可能事件的概率为 0，没有状态的概率小于 0
- 保证发生的事件的概率为 1，任何状态的发生概率都不可能比它大

$$\sum_{x\in \mathrm{x}} P(x)=1$$

- 我们称这个性质为**标准化 normalized**
- 如果没有这个特性，我们可以通过计算众多事件中的一个发生的概率来获得大于1的概率

分布模型

均匀分布 Uniform Distribution

$$P(x=x_i)=\frac{1}{k}$$
$$u(x;a,b)=\frac{1}{b-a}$$

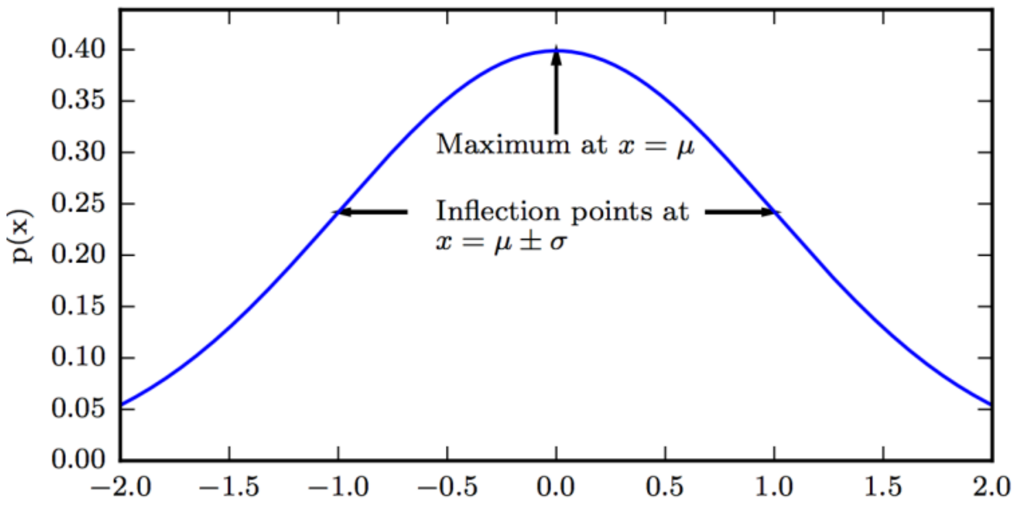
边际概率（离散和连续）Marginal probabilities

$$\forall x \in \mathrm{x}, P(\mathrm{x}=x)=\sum_{y} P(\mathrm{x}=x, \mathrm{y}=y)$$
$$p(x)=\int p(x, y) d y$$

伯努利分布 Bernoulli distribution

$$P(\mathrm{x}=1)=\phi$$
$$P(\mathrm{x}=0)=1-\phi$$
$$P(\mathrm{x}=x)=\phi^{x}(1-\phi)^{1-x}$$
$$\mathbb{E}_{\mathrm{x}}[\mathrm{x}]=\phi$$
$$\operatorname{Var}_{\mathrm{x}}(\mathrm{x})=\phi(1-\phi)$$

高斯分布 Gaussian distribution



基于方差的分布

$$\mathcal{N}\left(x ; \mu, \sigma^{2}\right)=\sqrt{\frac{1}{2 \pi \sigma^{2}}} \exp \left(-\frac{1}{2 \sigma^{2}}(x-\mu)^{2}\right)$$

基于精度的分布

$$\mathcal{N}\left(x ; \mu, \beta^{-1}\right)=\sqrt{\frac{\beta}{2 \pi}} \exp \left(-\frac{1}{2} \beta(x-\mu)^{2}\right)$$

多维高斯分布

$$\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})=\sqrt{\frac{1}{(2\pi)^n\det(\boldsymbol{\Sigma})}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

指数分布 Exponential

$$p(x;\lambda)=\lambda\mathbf{1}_{x\geq 0}exp(-\lambda x)$$

拉普拉斯分布 Laplacian

$$Laplace(x;\mu,\gamma)=\frac{1}{2\gamma}exp(-\frac{|x-\mu|}{\gamma})$$

狄拉克分布 Dirac

$$p(x)=\delta(x-\mu)$$

分布的特点

条件概率 Conditional probability

$$P(\mathrm{y}=y\mid\mathrm{x}=x)=\frac{P(\mathrm{y}=y,\mathrm{x}=x)}{P(\mathrm{x}=x)}$$

条件概率链式法则

$$P\left(\mathrm{x}^{(1)},\ldots,\mathrm{x}^{(n)}\right)=P\left(\mathrm{x}^{(1)}\right)\Pi_{i=2}^{n}P\left(\mathrm{x}^{(i)}\mid\mathrm{x}^{(1)},\ldots,\mathrm{x}^{(i-1)}\right)$$

独立性 Independence

$$\forall x\in\mathrm{x},y\in\mathrm{y},p(\mathrm{x}=x,\mathrm{y}=y)=p(\mathrm{x}=x)p(\mathrm{y}=y)$$

条件独立

$$\forall x\in\mathrm{x},y\in\mathrm{y},z\in\mathrm{z},p(\mathrm{x}=x,\mathrm{y}=y\mid\mathrm{z}=z)=p(\mathrm{x}=x\mid\mathrm{z}=z)p(\mathrm{y}=y\mid\mathrm{z}=z)$$

概率统计模型特征

期望 Expectation

- 对于离散随机变量： $\mathbb{E}_{\mathrm{x}\sim P}[f(x)]=\sum_x P(x)f(x)$
- 对于连续随机变量：$\mathbb{E}_{\mathrm{x}\sim p}[f(x)]=\int p(x)f(x)dx$
- 期望的线性特征：$\mathbb{E}_{\mathbf{x}}[\alpha f(x)+\beta g(x)]=\alpha\mathbb{E}_{\mathbf{x}}[f(x)]+\beta\mathbb{E}_{\mathbf{x}}[g(x)]$

方差与协方差 Variance and Covariance

- 方差：$\mathrm{Var}(f(x))=\mathbb{E}\left[(f(x)-\mathbb{E}[f(x)])^2\right]$
- 协方差：$\mathrm{Cov}(f(x),g(y))=\mathbb{E}[(f(x)-\mathbb{E}[f(x)])(g(y)-\mathbb{E}[g(y)])]$
- 给定一个随机向量 $\mathbf{x}$ 的协方差矩阵 $\mathrm{Cov}(\mathbf{x})_{i,j}=\mathrm{Cov}\left(\mathbf{x}_i,\mathbf{x}_j\right)$

4. 真实情况下的计算问题

- 算法通常用实数来表示
  - 不能用有限数量的比特来实现，这个算法还有效吗？
- 函数输入的小变化会导致输出的大变化吗？
  - 舍入/测量误差，噪声，可以引起很大的变化
  - 迭代搜索最佳输入是困难的

舍入和截断误差 Rounding and truncation errors

在计算机中，我们使用 float32 或者相似的体系来描述实数

一个实数 x 会被估计成 x + delta（一些小的 delta）

- Overflow：大的 x 会被替换成 inf
- Underflow：小的 x 会被替换成 0

将一个非常小的数字加到一个较大的数字上可能没有任何效果，这可能会导致下游发生很大的变化

二次效应

假设我们有一个代码计算 x-y

- 假设 x 溢出成了 inf
- 假设 y 溢出成了 inf

那么 x - y = inf - inf = NaN

exp

大的 x 会造成 exp(x) 的溢出

- 例如：在 float32 中，exp(89) 都会造成溢出

非常负的 x 会造成 exp(x) Underflow

- 当它作为分母、对数的参数的时候，可能是灾难性的

log 和 sqrt

- log(0) = -inf
- log(某个负数) = NaN
- sqrt(0) = 0，但是它的倒数被除以 0

寻找 bug 的策略

- 如果你提高了你的学习速度，而损失却被卡住了，你可能会在某个地方将梯度舍入到零：可能会使用概率而不是对数来计算交叉熵
- 对于正确执行的损耗，过高的学习率通常会导致爆炸
- 如果看到爆炸（NaN，非常大的值），立即怀疑
    - log
    - exp
    - sqrt
    - division
- 总是怀疑那些最近更改的代码