1 数学基础

https://www.cnblogs.com/pinard/archive/2004/01/13/10791506.html

1.1 矩阵

1.1.1 迹相关性质

- 转置不变性: $tr(A^T) = tr(A)$
- 迹的循环相等性: $tr(ABC) = tr(BCA) = tr(CAB)$
- 加减分配: $tr(X \pm Y) = tr(X) \pm tr(Y)$
- 乘法和迹交换: $tr((A \odot B)^T C) = tr(A^T(B \odot C))$

1.2 矩阵导数

1.2.1 导数布局

行头对列头求导	标量y	m维向量$\boldsymbol{y}$	$p \times q$维矩阵$\boldsymbol{Y}$
标量x	1×1	$m \times 1$	$\frac{\partial \boldsymbol{Y}}{\partial x}$, $p \times q$
n维向量$\boldsymbol{x}$	$\frac{\partial y}{\partial \boldsymbol{x}}$, $n \times 1$	$n \times m$	
$r \times s$维矩阵$\boldsymbol{X}$	$\frac{\partial y}{\partial \boldsymbol{X}}$, $r \times s$		

1.2.2 标量对张量求导

标量值函数的矩阵微分定义为:

$$df = \sum_{i,j} \frac{\partial f}{\partial X_{ij}} dX_{ij} = tr((\frac{\partial f}{\partial X})^T dX) = tr((\frac{\partial f}{\partial X}) dX^T)$$

并且具有如下常用性质:

1. $d(X \pm Y) = dX \pm dY$, $d(XY) = (dX)Y + X(dY)$, $d(X \odot Y) = (dX) \odot Y + X \odot (dY)$
2. $d(X^T) = (dX)^T$, $d(tr(X)) = tr(dX)$
3. $dX^{-1} = -X^{-1}(dX)X^{-1}$
4. $d|X| = |X| tr(X^{-1}dX)$
5. 逐元素求导: $d\sigma(X) = \sigma'(X) \odot dX$

例题1. 求$\frac{\partial tr(ABA^T)}{\partial A}$
$d(tr(ABA^T)) = tr((dA)BA^T + AB(dA^T)) = tr(AB^T(dA)^T) + tr(AB(dA)^T) = tr(A(B^T + B)(dA)^T)$
故有$\frac{\partial tr(ABA^T)}{\partial A} = A(B^T + B)$

例题2. $y = \boldsymbol{a}^T exp(X\boldsymbol{b})$, 求$\frac{\partial y}{\partial X}$
$dy = tr(dy) = tr(\boldsymbol{a}^T dexp(X\boldsymbol{b})) = tr(\boldsymbol{a}^T exp(X\boldsymbol{b}) \odot d(X\boldsymbol{b})) = tr((\boldsymbol{a} \odot exp(X\boldsymbol{b}))^T d(X\boldsymbol{b})) = tr(b(a \odot exp(X\boldsymbol{b}))^T dX)$
故$\frac{\partial y}{\partial X} = (a \odot exp(X\boldsymbol{b}))b^T$

1.2.3 链式法则

- 比如$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}$, 本质上就是对中间变量的每个分量对$\boldsymbol{x}$求导乘上$\boldsymbol{y}$对中间变量的每个分量求导, 然后求和.

1.2.4 对向量求导链式法则

- 向量对向量的链式法则:
$$\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{y}} \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}$$

- 标量对向量的链式法则:
$$\frac{\partial z}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} \frac{\partial z}{\partial \boldsymbol{y}}, \qquad ((len(\boldsymbol{x}), len(\boldsymbol{y})) \cdot (len(\boldsymbol{y}), 1))$$
$$\frac{\partial z}{\partial \boldsymbol{x}} = (\frac{\partial \boldsymbol{y_n}}{\partial \boldsymbol{y_{n-1}}} \frac{\partial \boldsymbol{y_{n-1}}}{\partial \boldsymbol{y_{n-2}}} \cdots \frac{\partial \boldsymbol{y_2}}{\partial \boldsymbol{y_1}}) \frac{\partial z}{\partial \boldsymbol{y_n}}, \qquad ((len(\boldsymbol{y_n}), len(\boldsymbol{y_1})) \cdot (len(\boldsymbol{y}), 1))$$

例题1. 最小二乘法损失函数的求导, 即求$\frac{\partial l}{\partial \theta}$, 这里$l = (X\boldsymbol{\theta} - \boldsymbol{y})^T (X\boldsymbol{\theta} - \boldsymbol{y})$
令$\boldsymbol{z} = X\boldsymbol{\theta} - \boldsymbol{y}$
则$\frac{\partial l}{\partial \boldsymbol{\theta}} = (\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{\theta}}) \frac{\partial l}{\partial \boldsymbol{z}} = X^T \frac{\partial \boldsymbol{z}^T \boldsymbol{z}}{\partial \boldsymbol{z}}$
而$tr(d(\boldsymbol{z}^T \boldsymbol{z})) = tr(\boldsymbol{z}^T (d\boldsymbol{z}) + (d\boldsymbol{z}^T)\boldsymbol{z}) = tr(\boldsymbol{z}^T (d\boldsymbol{z}) + (\boldsymbol{z} d\boldsymbol{z}^T)) = tr(2\boldsymbol{z})$
所以$\frac{\partial l}{\partial \boldsymbol{\theta}} = X^T 2\boldsymbol{z} = 2X^T (X\boldsymbol{\theta} - \boldsymbol{y})$

1.2.5 特殊求导

- 梯度: $\nabla f(\boldsymbol{x})_i = \frac{\partial f(\boldsymbol{x})}{\partial x_i}$
- 海森矩阵: $\nabla^2 f(\boldsymbol{x})_{ij} = \frac{\partial^2 f(\boldsymbol{x})}{\partial x_i \partial x_j}$
- Frobenius范数求导: $\frac{\partial \|A\|_F^2}{\partial A} = \frac{\partial A^T A}{\partial A} = 2A$

1.3 矩阵分解

1.3.1 特征值分解

- 特征方程: $A\boldsymbol{v} = \lambda \boldsymbol{v}$
- 特征值分解: $A = V diag(\boldsymbol{\lambda}) V^{-1}$, V的每一列为特征向量$\boldsymbol{v_i}$
- 特征值为正(非负)的矩阵为**正定(半正定)矩阵**
- 可特征值分解(可对角化)的充分条件: 实对称矩阵
- 手算分解方法: 求出特征向量就行了.

1.3.2 奇异值分解

- 设矩阵$A \in C^{m \times n}$, $A^T A\boldsymbol{v} = \lambda \boldsymbol{v}$,
  令$A\boldsymbol{v} = \sigma \boldsymbol{u}$, 则
  - 两边同乘AA^T得到, $AA^T \boldsymbol{u} = \frac{\lambda}{\sigma} A\boldsymbol{v} = \lambda \boldsymbol{u}$, $\boldsymbol{u}$对应AA^T的特征值为λ的特征向量
  - 两边同左乘$\boldsymbol{u}^T$得到, $\boldsymbol{u}^T A\boldsymbol{v} = \sigma$
  - 两边同左乘$\boldsymbol{v}^T A^T$得到, $\frac{\lambda}{\sigma} = \boldsymbol{v}^T A^T \boldsymbol{u} = \boldsymbol{u}^T A\boldsymbol{v}$

因此$\lambda = \sigma^2$
写成矩阵形式，有$AV = U\Sigma$即$A = U\Sigma V^T$，其中$U \in C^{m\times m}$，为AA^T的特征向量构成矩阵；$\Sigma \in C^{m\times n}$；$V \in C^{n\times n}$，为A^TA的特征向量构成矩阵

- 任意矩阵都可以作奇异值分解
- 截断奇异值分解：$A \approx U_k\Sigma_k V_k^T$，$U_k \in C^{m\times k}$即取前k列；$\Sigma_k \in C^{k\times k}$；$V_k \in C^{n\times k}$，即取前k行

1.4 概率分布

- 多元正态分布
  - $\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\Sigma) = \frac{1}{\sqrt{(2\pi)^n det(\Sigma)}} = exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$
  - $\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\beta}^{-1}) = \sqrt{\frac{det(\boldsymbol{\beta})}{(2\pi)^n}} = exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\beta}(\boldsymbol{x}-\boldsymbol{\mu})\right)$
- 贝塔分布
  - 概率密度函数：
  $$Beta(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$$
  - $E[\mu] = \frac{a}{a+b}$
  - $Var[\mu] = \frac{ab}{(a+b)^2(a+b+1)}$
- 狄利克雷分布(是贝塔分布的多元扩展)
  - 多个连续变量$\mu_i \in [0,1]$的分布，满足$\sum \mu_i = 1$
  - $Dir(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_i \alpha_i)}{\Pi_i \alpha_i}\Pi_i \mu_i^{\alpha_i - 1}$
  - $E[\mu_i] = \frac{\alpha_i}{\sum_i \alpha_i}$
- 伽马分布($\tau > 0$)
  - $Gam(\tau|a,b) = \frac{1}{\Gamma(a)}b^a\tau^{a-1}e^{-b\tau}$
  - $E[\tau] = \frac{a}{b}$
  - $Var[\tau] = \frac{a}{b^2}$

1.5 K-L散度

- 衡量两个分布的差异
  - K-L散度定义：
  $$D_{KL}(P\|Q) = E_{X\sim P}\left[log\frac{P(x)}{Q(x)}\right]$$
  即
  $$D_{KL}(P\|Q) = E_{X\sim P}logP(x) - E_{X\sim P}Q(x)$$
  $E_{X\sim P}logP(x)$为P的熵，$E_{X\sim P}Q(x)$为P和Q的交叉熵
  - 非负，P=Q时为0
  - $D_{KL}(P\|Q) \neq D_{KL}(Q\|P)$，但理论上最小值都是P=Q

1.6 优化

- 标准优化问题:
$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) s.t. \begin{cases} g_i(\boldsymbol{x}) \leq 0, i=1,2,\cdots,m \\ h_j(\boldsymbol{x}) = 0, j=1,2,\cdots,n \end{cases}$$

- 凸函数
  - 定义(零阶条件): 对于$f: \mathbb{R}^n \mapsto \mathbb{R}$, 若$\forall t \in [0,1], f(t\boldsymbol{x}+(1-t)\boldsymbol{y}) \leq tf(\boldsymbol{x})+(1-t)f(\boldsymbol{y})$, 则$f$为凸函数
  - 一阶条件: $f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T(\boldsymbol{y}-\boldsymbol{x})$
  - 二阶条件: $\nabla^2 f(\boldsymbol{x}) \geq 0$, 即海森矩阵半正定

- 凸优化问题:
$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) s.t. \begin{cases} g_i(\boldsymbol{x}) \leq 0, i=1,2,\cdots,m \\ \boldsymbol{a}_i^T\boldsymbol{x} = b, j=1,2,\cdots,n \end{cases}$$

- 无约束优化: 梯度下降: $x_{t+1} \leftarrow x_t - \alpha \nabla f(x_t)$

- 有约束优化: 拉格朗日乘子法
  - 拉格朗日函数: $L(\boldsymbol{x},\boldsymbol{u},\boldsymbol{v}) = f(\boldsymbol{x}) + \sum_i u_i g_i(\boldsymbol{x}) + \sum_j v_j h_j(\boldsymbol{x})$, 其中$u_i \geq 0$
  - $\forall \boldsymbol{u} \geq 0, \boldsymbol{v}$和可行解$\boldsymbol{x}$, 有
  $$L(\boldsymbol{x},\boldsymbol{u},\boldsymbol{v}) \leq f(\boldsymbol{x})$$
  - 对偶问题:
  $$\max_{\boldsymbol{u},\boldsymbol{v}} g(\boldsymbol{u},\boldsymbol{v}),\ s.t.\ \boldsymbol{u} \geq 0$$
  其中$g(\boldsymbol{u},\boldsymbol{v}) = \min_{\boldsymbol{x}} L(\boldsymbol{x},\boldsymbol{u},\boldsymbol{v})$
  - 弱对偶性: 设可行解集C, 原问题最优解f^*, 则
  $$f^* \geq \min_{\boldsymbol{x}\in C} L(\boldsymbol{x},\boldsymbol{u},\boldsymbol{v}) \geq \min_{\boldsymbol{x}} L(\boldsymbol{x},\boldsymbol{u},\boldsymbol{v}) = g(\boldsymbol{u},\boldsymbol{v})$$
  从而得到
  $$f^* \geq g^* = \max_{\boldsymbol{u},\boldsymbol{v}} g(\boldsymbol{u},\boldsymbol{v})$$

2 模型评估与选择

2.1 误差与过/欠拟合

- 误差: 预测与实际之差异
- 训练误差/经验误差: 在训练集上的误差
- 泛化误差: 在新样本上的误差
- 过拟合/欠拟合

2.2 评估方法

2.2.1 数据集划分

- 留出法:
  - 注意分层采样以保留比例
  - 一般采用若干次随即划分取平均值作为留出法的评估结果.
  - 一般将$2/3 \sim 4/5$的样本用于训练
- k折交叉验证法
  - 最常用k=10
  - 当k=m(即数据集D中样本数)时, 称为留一法. 准确(未必)但开销大.

- 自助法: 采样到D'后放回, 则约有$\lim\limits_{m\to\infty}\left(1-\frac{1}{m}\right)^m = \frac{1}{e} \approx 0.368$ 的样本未出现在采样的数据集D'中, 可将其(D/D')作为测试集
  - 数据集小难以划分时有用, 但改变了初始数据集的分布会引入估计偏差

2.2.2 调参

- 网格法/随机法
- 验证集: 从训练集中选出用来验证超参数

2.3 性能度量

- 均方误差(MSE)/根均方误差(RMSE)
- 绝对误差(MAE)

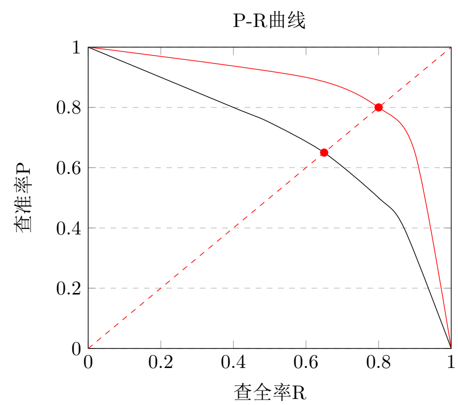
2.3.1 分类任务

- 基本定义:
  - 错误率: 分错样本数/样本总数
  - 精度(accuracy): 分对样本数/样本总数
  - 混淆矩阵:

真实情况	预测结果	
	正例	反例
正例	TP(真正例)	FN(假反例)
反例	FP(假正例)	TN(真反例)

  - 查准率$P = \frac{TP}{TP+FP}$
  - 查全率$R = \frac{TP}{TP+FN}$
  - 真正例率$TPR = \frac{TP}{TP+FN}$
  - 假正例率$FPR = \frac{FP}{FP+TN}$

- 度量方法
  - P-R图法: 查全率为横轴, 查准率为纵轴, 作出曲线.
    * 用**P-R曲线下面积**大小度量
    * 用**平衡点**度量(平衡点在直线P=R上)



  - **F1度量**:

$$F1 = \frac{2 \times P \times R}{P + R} = \frac{2}{\frac{1}{P} + \frac{1}{R}}$$

  - F_β**度量**:

$$F_\beta = \frac{1 + \beta^2}{\frac{1}{P} + \frac{\beta^2}{R}}$$

    $\beta > 1$时, 查全率影响更大; $\beta < 1$时, 查准率影响更大
  - **宏F1**: 由于多次重复训练测试, 有多个混淆矩阵. 可用:

$$macro_P = \frac{1}{n}\sum_{i=1}^{n} P_i$$
$$macro_R = \frac{1}{n}\sum_{i=1}^{n} R_i$$
$$macro_{F1} = \frac{2 \times macro_P \times macro_R}{macro_P + macro_R}$$

  - **微F1**:

$$micro_P = \frac{\bar{TP}}{\bar{TP}+\bar{FP}}$$
$$micro_R = \frac{\bar{TP}}{\bar{TP}+\bar{FN}}$$
$$micro_{F1} = \frac{2 \times micro_P \times micro_R}{micro_P + micro_R}$$

- **ROC与AUC**: 衡量排序(分类任务常先预测出一个数值)质量好坏
  * ROC曲线: 给定不同阈值, 会有一个假正例率FPR和一个真正例率TPR, 以这些点画图. 得到以假正例率为横轴, 真正例率为纵轴的图
  * AUC: Area Under ROC Curve. 有公式:

$$AUC = \frac{1}{2}\sum_{i=1}^{m-1}(x_{i+1} - x_i) \cdot (y_i + y_{i+1})$$

  * 排序损失:

$$\begin{aligned} l_{rank} &= \frac{1}{m^+m^-}\sum_{x^+\in D^+}\sum_{x^-\in D^-}\left(\mathbb{I}(f(x^+) < f(x^-)) + \frac{1}{2}\mathbb{I}(f(x^+) = f(x^-))\right) \\ &= 1 - AUC \end{aligned}$$

- **代价敏感错误率与代价曲线**
  * 代价敏感错误率(cost-sensitive):

$$E(f;D;cost) = \frac{1}{m}\left(\sum_{x_i\in D^+}\mathbb{I}(f(x_i) \neq y_i) \times cost_{01} + \sum_{x_i\in D^-}\mathbb{I}(f(x_i) \neq y_i) \times cost_{10}\right)$$

  * 代价曲线: 正例概率代价为横轴, 归一化代价为纵轴. ROC上每一点对应代价曲线上一条线段. 可以在非均等代价下衡量期望总体代价(ROC曲线不行)
    · 正例概率代价

$$P(+)cost = \frac{p \times cost_{01}}{p \times cost_{01} + (1-p) \times cost_{10}}$$

    · 归一化代价

$$cost_{norm} = \frac{FNR \times p \times cost_{-1} + FPR \times (1-p) \times cost_{10}}{p \times cost_{01} + (1-p) \times cost_{10}}$$

2.4 比较检验

2.4.1 假设检验

- 二项检验:若测试错误率$\hat{\epsilon}$小于临界值$\bar{\epsilon}$, 则可认为在α显著度下, 假设"$\epsilon \leq \epsilon_0$不能被拒绝.

$$\bar{\epsilon} = \min \epsilon \ s.t. \sum_{i=\epsilon\times m+1}^{m}\begin{pmatrix} m \\ i \end{pmatrix}\epsilon_0^i(1-\epsilon_0)^{m-1} < \alpha$$

- t检验: 多次重复留出法/交叉验证法, 得到多个测试错误率$\hat{\epsilon}_1 \cdots \hat{\epsilon}_k$. 则($\mu$和$s^2$是测试错误率的样本均值和样本方差)

$$\tau_t = \frac{\sqrt{k}(\mu - \epsilon_0)}{s} \sim t_{k-1}$$

- 成对交叉验证t检验: 两两同折作差得到$\Delta_1 \cdots \Delta_k$,其样本均值和样本方差为μ, σ^2,则

$$\tau_t = \left|\frac{\sqrt{k}\mu}{\sigma}\right| \sim t_{k-1}$$

- 问题: 采样可能不独立, 可以用"5×2交叉验证"(书41)

2.4.2 McNemar检验

算法B	算法A	
	正确	错误
正确	e_{00}	e_{01}
错误	e_{10}	e_{11}

假设性能相同应该有$e_{01}=e_{10}$, $|e_{01}-e_{10}|\sim N$, 从而

$$\tau_{\chi^2}=\frac{(|e_{01}-e_{10}|-1)^2}{e_{01}+e_{10}}\sim\chi_1^2$$

2.4.3 Friedman检验与Nemenyi后续检验

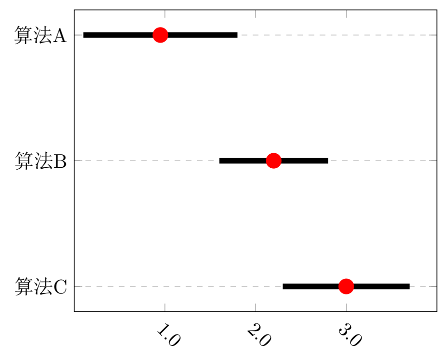
- 定义序值,对每个算法在不同数据集序值求均值, 则有

$$\tau_{\chi^2}=\frac{k-1}{k}\frac{12N}{k^2-1}\sum_{i=1}^{k}(r_i-\frac{k+1}{2})^2\sim\chi_{k-1}^2$$

- 另有

$$\tau_F=\frac{(N-1)\tau_{\chi^2}}{N(k-1)-\tau_{\chi^2}}\sim F_{k-1,(k-1)(N-1)}$$

- 后续检验(拒绝了算法性能相同后的后续比较)
  - 临界值域: $CD=q_\alpha\sqrt{\frac{k(k+1)}{6N}}$
  - 有交叠就无明显差异, 否则有显著优性.(如图中算法A显著优于算法C)



Friedman检验图

2.5 偏差与方差

- 学习算法的期望预测:

$$\bar{f}(x)=E_D[f(x;D)]$$

- 方差:

$$var(x)=E_D\left[(f(x;D)-\bar{f}(x))^2\right]$$

- 噪声(数据标签就标错了):

$$\epsilon^2=E_D\left[(y_D-y)^2\right]$$

- 偏差(期望输出与真实标记的差别):

$$bias^2(x)=(\bar{f}(x)-y)^2$$

- 偏差-方差分解:

$$E(f;D)=bias^2(x)+var(x)+\epsilon^2$$