第 1 章 第二次作业

本次作业包含三部分内容, 即基础题, 思考题和编程题. 作业提交形式为线上提交, 应使用 LaTeX 独立完成作业电子版.

Deadline: 2022 年 4 月 11 日 23:59.

1.1 基础题

说明：该部分的所有题目均为必做，所有题目及相关变种均可能出现在试卷上。

1.1.1 强凸函数性质

设 $f:\mathbb{R}^n\to\mathbb{R}$ 是连续可微函数, 并且 f 是强凸的, 即存在 $\mu>0$, 满足:

$$f(y)\geq f(x)+\nabla f(x)^{\mathrm{T}}(y-x)+\frac{\mu}{2}\|y-x\|^2.$$

证明以下结论:

1. 函数 $h(x)=f(x)-\frac{\mu}{2}\|x\|^2$ 是凸函数.
2. 梯度的强单调性:
$$(\nabla f(x)-\nabla f(y))^{\mathrm{T}}(x-y)\geq\mu\|x-y\|^2.$$
3. 如果 $f(x)$ 是二阶连续可微的, 则 $\nabla^2 f(x)\succeq\mu I$.

1.1.2 光滑函数性质

设 $f:\mathbb{R}^n\to\mathbb{R}$ 是二阶连续可微函数, 并且其梯度 $\nabla f(x)$ 是 L-Lipschitz 连续的, 即

$$\|\nabla f(x)-\nabla f(y)\|\leq L\|x-y\|,\forall x,y\in\mathbb{R}^n.$$

则有如下结论:

1. $\nabla^2 f(x)\preceq LI$.
2. 函数 $h(x)=\frac{L}{2}\|x\|^2-f(x)$ 是凸函数.

1.1.3 二次函数情形中梯度下降法的收敛性分析

给出二次函数

$$f(\boldsymbol{x})=\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}Q\boldsymbol{x}+\boldsymbol{b}^{\mathrm{T}}\boldsymbol{x}+c,$$

其中 Q 为严格正定的实对称矩阵，我们记

$$L\triangleq\lambda_{\max}=\|Q\|,\quad \mu\triangleq\lambda_{\min}=\|Q^{-1}\|.$$

使用线搜索 $\boldsymbol{x}_{t+1}=\boldsymbol{x}_t-\alpha_t\boldsymbol{d}_t$ 求解上述问题. 完成以下讨论.

1. 参考Wikipedia和文献 [1] 等，给出康托罗维奇（Kantorovich）不等式的形式;
2. 如果使用负梯度方向为搜索方向, 精确线搜索确定步长, 计算每次迭代 α_t.
3. 请基于 Kantorovich 不等式证明采用精准线搜索步长梯度下降法的收敛性满足
$$\frac{f(\boldsymbol{x}_{t+1})-f(\boldsymbol{x}^*)}{f(\boldsymbol{x}_t)-f(\boldsymbol{x}^*)} \leq \left(\frac{\lambda_{\max}-\lambda_{\min}}{\lambda_{\max}+\lambda_{\min}}\right)^2;$$
4. 如果采用固定步长 $\alpha = 1/L$，证明收敛性满足
$$\frac{f(\boldsymbol{x}_{t+1})-f(\boldsymbol{x}^*)}{f(\boldsymbol{x}_t)-f(\boldsymbol{x}^*)} \leq 1-\frac{\mu}{L};$$
5. 根据条件数（Condition Number）κ（LaTeX 符号为\kappa）的定义，对上面的收敛性结论进行改写.

1.1.4 Fisher 判别的简单公式推导

对于给定的 k 组 m 维数据 $\{X_i\}_{i=1}^N$，我们希望能将之投影到某一个（几个）方向上，使组与组之间尽可能地分开. 一种常见的做法是采用方差分析的思想来导出判别函数，我们不妨记投影方向为 w. 我们的目标是尽可能增大组间方差，减小组内方差，我们这里直接考虑组间平方和与组内平方和. 投影后的组间平方和是容易得到的：

$$\sum_{i=1}^{k} n_i w^{\mathrm{T}}\left(\overline{X^{(i)}}-\overline{X}\right)^2 = w^{\mathrm{T}}\left(\sum_{i=1}^{k} n_i\left(\overline{X^{(i)}}-\overline{X}\right)\left(\overline{X^{(i)}}-\overline{X}\right)^{\mathrm{T}}\right)w \triangleq w^{\mathrm{T}}S_{\mathrm{b}}w.$$

其中，$n_i, \overline{X^{(i)}}$ 分别是第 i 组数据的数量与组中心. 我们称 S_{b} 为组间离差阵（between）. 上式的含义其实就是把一个组看成一个整体，按组中数据量赋予权重进行求和，最后进行了一个写法上的简化.

同样的，有组内平方和，我们把对应的组内离差阵（within）记为 S_{w}，对应推导如下：

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(w^{\mathrm{T}}\left(X_j^{(i)}-\overline{X^{(i)}}\right)\right)^2 = w^{\mathrm{T}}\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_j^{(i)}-\overline{X^{(i)}}\right)\left(X_j^{(i)}-\overline{X^{(i)}}\right)^{\mathrm{T}}\right)w \triangleq w^{\mathrm{T}}S_{\mathrm{w}}w.$$

因此，一维 Fisher 判别的优化目标为

$$\max_{w}\frac{w^{\mathrm{T}}S_{\mathrm{b}}w}{w^{\mathrm{T}}S_{\mathrm{w}}w}.$$

对于多维的 Fisher 判别，有多种的不同的目标，我们这里主要介绍一种通过 trace 构造的. 对于向多个方向投影，我们记最终选定的投影方向为 $\{w_i\}_{i=1}^p$. 我们把离差阵统一记为 S，那对应的平方和可以写作

$$\sum_{i=1}^{p} w_i^{\mathrm{T}}Sw_i = \sum_{i=1}^{p}\sum_{j=1}^{p}\delta_{ij}w_i^{\mathrm{T}}Sw_j = \mathrm{tr}\left(W^{\mathrm{T}}SW\right).$$

此处 δ_{ij} 为 Kronecker-δ，$W = \begin{bmatrix} w_1 & w_2 & \cdots & w_p \end{bmatrix}$. 一种令人困惑的思路是采用行列式进

行构造，据称解释是因为行列式的值实际上是矩阵特征值的积，一个特征值可以表示在该特征向量上的发散程度。（如果你能想到一个好的解释，请务必指点一下愚蠢的 TA）.

结合上述介绍，与部分参考资料（机器学习幼儿园 · 线性模型、机器学习幼儿园 · 降维与度量学习、Fisher’s Linear Discriminant: Intuitively Explained、Linear discriminant analysis : a detailed tutorial）完成以下问题：

1. 对于一维 Fisher 判别的优化目标

$$\max_{w} \frac{w^{\mathrm{T}} S_{\mathrm{b}} w}{w^{\mathrm{T}} S_{\mathrm{w}} w}$$

进行求解.

2. 对于 p 维 Fisher 判别的优化目标

$$\max_{W} \frac{\operatorname{tr}\left(W^{\mathrm{T}} S_{\mathrm{b}} W\right)}{\operatorname{tr}\left(W^{\mathrm{T}} S_{\mathrm{w}} W\right)}$$

进行求解.

3. 总结 PCA 与 LDA 间的异同.

1.1.5 主成分的简单计算

设 p 元总体 X 的协方差阵为

$$\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix} \quad (0 < \rho \leq 1).$$

1. 试证明总体的第一主成分 $Z_1 = \frac{1}{\sqrt{p}} (X_1 + X_2 + \cdots + X_p)$;
2. 试求第一主成分的贡献率.

1.1.6 指数组函数与 PCA 的简单性质

指数族分布的形式如下

$$p(x|\eta) = h(x) \exp\left(\eta^{\mathrm{T}} T(x) - A(\eta)\right).$$

其中，η 为自然参数（natural parameter），$T(x)$ 为充分统计量（sufficient statistic），$h(x)$ 为测度（underlying measure），$A(\eta)$ 为对数归一化因子（log normalizer）. 相关内容可以参考 [2] 与 [3].

1. 请推导指数族分布的 moment matching 性质：

$$\nabla_{\eta} A(\hat{\eta}) = \mathbb{E}_{\hat{\eta}} \left(T(x)\right).$$

2. 基于极大似然参数估计准则，进行进一步推导：

$$\nabla_\eta A(\hat{\eta}) = \frac{1}{N}\sum_{i=1}^{N} T\left(x^{(i)}\right).$$

3. 基于带约束问题的拉格朗日乘子优化方法，推导指数族分布为满足观测到的充分统计量的分布中，假设最少的模型，即最大熵模型.

$$\text{Constraints: } \mathbb{E}(f_k) = F_k,\ 1 \le k \le n$$

4. 从充分统计量的角度谈谈 PCA 的应用场景.

1.2 思考题

说明：该部分的所有题目均可以不做，给出精彩的解答可以弥补在基础题区域的失分，特别精彩的解答可能可以折算为平时成绩的额外加分（待定）。部分题目及相关变种均可能出现在试卷上。

本次思考题以调研为主，请做好文献引用。对于给出的两个方向，优质地完成一个即可加分（多选多加分）。每一个方向可以进入加分的最低篇幅为 3 页（不含参考文献列表）。

1.2.1 网络初始化的简单技巧

1. 从网络退化的角度解释为什么不能在初始化时将所有参数设置为相同数值;
2. 介绍 Xavier、MSRA，并选择至少一种初始化方法进行深入调研.

1.2.2 梯度消失与梯度爆炸的初步调研

1. 就梯度消失与梯度爆炸进行调研;
2. 就梯度消失与梯度爆炸的缓解方法进行调研.

1.3 编程题

说明：该部分注明为“思考题”的题目均可以不做，给出精彩的解答可以弥补在基础题区域的失分，特别精彩的解答可能可以折算为平时成绩的额外加分（待定）。

非“思考题”部分必做，主要考察思考过程，不会因为给出算法的效率差等扣分（不排除后期用 OJ 收代码，可能部分性能极差的代码会扣分）。

除特殊要求的题目外，所有题目的编程语言不限；除特殊申明，不限库函数的使用；担心提交代码无法编译的，可以备注开发环境。

代码可能会抽样查重，发现代码抄袭现象时（尤其是逐字符相同、最后修改日期不正常等），一切处理手段解释保留。

1.3.1 梯度法求解 LASSO 问题

阅读如下使用梯度法解 LASSO 问题的实例和对应的 MATLAB 代码:

- LASSO 问题的梯度法求解
- LASSO 问题的 Huber 光滑化梯度法
- LASSO 问题的连续化策略.

将上述实例中所有涉及的 MATLAB 程序转写成 Python 程序, 并实现上述实例. 注意, MATLAB 的 eig 函数可以由 numpy.linalg.eig 实现; MATLAB 的画图使用 matplotlib 实现.

1.3.2 多项式回归

使用多项式回归等方法, 拟合 data.m 中的数据, 作图并撰写报告. 要求实现以下模型: 最小二乘法, ℓ_2 范数的岭回归, ℓ_1 范数的 LASSO, 并使用梯度法寻找拟合参数, 不得直接调用 sklearn 的 fit 拟合 (但是可以阅读参考 fit 函数的代码).

1.3.3 二分法的简单应用

文件 ‘2020_icpc_shanghai_statement.pdf’ 中有十三道题，针对其中的 D 题，考虑基于二分法的算法，要求可以 AC第 45 届国际大学生程序设计竞赛（ICPC）亚洲区域赛（上海）D.Walker，具体如下：

1. 参考逆十字的答案，给出 C++ 代码实现.
2. 通常，我们认为 C++ 算法的需要进行 10^8 量级的运算可以在 1 秒中得出结果（性能高的平台可能可以达到 5×10^9）. 请结合二分法的收敛率，分析本题中实现 10^{-6} 次的精度迭代 100 次是一个合理的数值.
3. （思考题）参考五点共圆的答案，谈谈你对于闭式解于数值解的理解.

1.3.4 牛顿分形的简单介绍

参考【官方双语】（牛顿本人都不知道的）牛顿分形，学习牛顿分形的基本由来，并完成如下要求：

1. 在 3B1B 提供的Newton's Fractal交互平台上调出若干个自己觉得好看的分形，用 subfigure 环境提交.
2. 参考给大家看点好玩的，用 Python 画牛顿分形，用 Python 绘制不少于 4 个牛顿分形并提交分形与代码，要求如下：
   - 配色好看（可以参考 3B1B 的源码）；
   - （思考题）采用 CUDA 加速.

1.3.5 梯度下降法的简单变种

参考梯度下降法实现以下算法：

- Momentum 动量法
- Adagrad 法
- Adadelta 法
- RMSprop 法
- Adam 法

要求用 algorithm2e 包给出 Adam 法的伪代码，并自己实现各算法下降过程的可视化.

Bibliography

[1] Jianguo Huang and Jie-yong Zhou. “A direct proof and a generalization for a Kantorovich type inequality”. In: Linear Algebra and its Applications 397 (2005), pp. 185–192 (cit. on p. 2).

[2] David M. Blei. Exponential Families. Website. https://www.cs.princeton.edu/courses/archive/fall11/cos597C/lectures/exponential-families.pdf (cit. on p. 3).

[3] Michael I. Jordan. The Exponential Family and Generalized Linear Models. Website. https://people.eecs.berkeley.edu/~jordan/courses/281A-fall04/lectures/lec-10-7.pdf (cit. on p. 3).