第一章（绪论）作业

【2022.2.22 → 2022.2.28】

1.1 通过网上资料粗略调研以下系统之一的技术原理，分析其中哪些部分属于模式识别问题范畴，哪些部分采用了模式识别和机器学习方法：

（1） 手机拍照或视频直播中的自动美颜

（2） 无人驾驶汽车

（3） DeepCode

（4） DeepFake

（5） 疫情期间商城等场所使用的自动摄像测温设备

（6） 其他你感兴趣的系统

将调研和分析结果写成不超过 1 页的文字报告，如发现系统中未采用模式识别和机器学习方法也做出相应的分析。（说明：请从问题和任务性质角度分析是否属于模式识别和机器学习问题和方法，不要仅依据其中用到了某种你听说过的某个名词或方法而做出判断。）

第二章（分类器性能评估）作业

【2022.2.22 → 2022.2.28】

2.1 假设有两种检测 COVID-19 新冠病毒的技术，技术 A 的灵敏度是 80%，特异度是 100%；技术 B 的灵敏度是 90%，特异度是 99%。

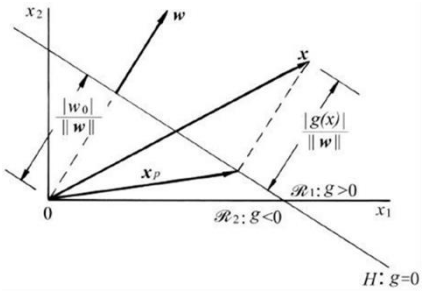
（1）设某城市现有人口 100 万人，近期确诊病例为 10,000 人，因此估计染病率大约在 1%左右。若 Tom 用技术 A 检测结果为阳性，Bob 用技术 A 检测结果为阴性，John 用技术 B 检测结果为阳性，Don 用技术 B 检测结果为阴性，请计算四人真正染病的可能性分别是多少。

（2）若该城市人口中确诊病例数仅 1,000 人，估计染病率在 0.1%左右，四人真正染病的可能性分别是多少？

（3）若四人均为现有病例的密切接触者，假设根据流行病学调查得知密切接触者的染病率大约在 0.5 左右，那么四人真正染病的可能性分别是多少？

（4）从以上结果中体会到了什么？试讨论这些认识对于类似检验的方法选择和措施设计有什么指导意义。

2.2 （下一章预热作业）写出 d 维空间中线性分类面的数学形式，证明空间中任意一点到分类面的距离和原点到分类面的距离。



第三章（线性学习机器）作业

【2022.3.1 → 2022.3.7】

3.1 写出 Fisher 线性判别最优投影方向的完整证明。（如参考教材之外的文献资料，须注明出处。）

3.2 回归问题中的性能评估指标：课上学习了分类器性能评估指标，本作业要求同学们课后自学回归模型的主要性能评估指标，简要说明以下名词的含义：

误差平方和、决定系数（R^2）、MAE 和 MAPE、MSE 和 RMSE

并讨论不同指标的作用、哪种指标最能体现回归模型的准确性。试证明线性回归中的R^2与皮尔森相关系数 r 的关系为$R^2 = r^2$。（可查阅文献资料，但须注明出处。）

3.3 证明 MSE 方法在选两类样本的 b 值分别为 N/N_1和 N/N_2时得到的解等价于 Fisher 线性判别的解且$w_0 = -\hat{m}$，其中N, N_1, N_2分别为样本总数和两类样本的数目。（可查阅文献资料，但须注明出处。）

3.4 试写出完整的罗杰斯特回归梯度下降算法的推导过程。（可查阅文献资料，但须注明出处。）

【2022.3.1 → 2022.3.14】

3.5 计算机小实验 1：线性回归练习

请对附件中的数据（prostate_train.txt 和 prostate_test.txt）使用前四个临床数据（即 lcavol，lweight，lbph，svi）对前列腺特异抗原水平（lpsa）进行预测。所给出的文件中，前 4 列每列代表一个临床数据（即特征），最后一列是测量的前列腺特异抗原水平（即预测目标的真实值）；每行代表一个样本。

作业要求：

（1） 在不考虑交叉项的情况下，利用 Linear Regression 对 prostate_train.txt 的数据进行回归，给出回归结果(R^2,RSS)，并对 prostate_test.txt 文件中的患者进行预测，给出结果评价(R^2,RSS)。

（2） 如考虑交叉项是否会有更好的预测结果？请分析理由（合理即可）。

数据名词解释：

lcavol：log cancer volume

lweight：log prostate weight

lbph：log of the amount of benign prostatic hyperplasia

svi：seminal vesicle invasion

lpsa：level of prostate-specific antigen

3.6 计算机小实验 2：线性分类器练习

请对附件中的数据（mammographic.txt），使用罗杰斯特回归和 Fisher 线性判别设计分类器，实现良性和恶性乳腺癌的分类和预测，要求进行十折交叉验证计算出各次实验正确率以及平均正确率，并对这两种方法做出比较。

附件乳腺癌诊断数据集是961×6维的矩阵，6维的特征信息如下。

Attribute Domain

```
1.  BI-RADS      [0,6]
2.  Age          [18,96]
3.  Shape        [1,4]
4.  Margin       [1,5]
5.  Density      [1,4]
6.  Severity     {0, 1}  (0 for benign, 1 for malignant)
@inputs BI-RADS, Age, Shape, Margin, Density
@outputs Severity
```

说明：

1. Logistic 回归可以调用函数，Fisher 线性判别请**自行编写程序实现**。
2. 请用适当的方法处理数据中的**缺失值**并在作业中注明**具体步骤**。

第四章（人工神经网络）作业

【2022.3.8 → 2022.3.14】

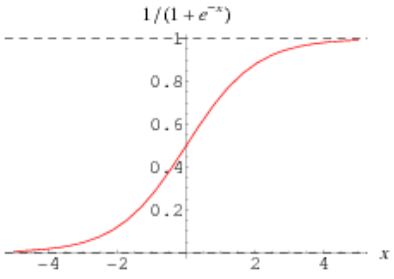
4.1 试证明，多层感知器中节点的激活函数如果采用线性函数，网络无法实现非线性映射。

4.2 (1) 对罗杰斯特函数（即 Sigmoid 函数）$\theta(s)=\frac{1}{1+e^{-s}}$，证明以下关系成立：

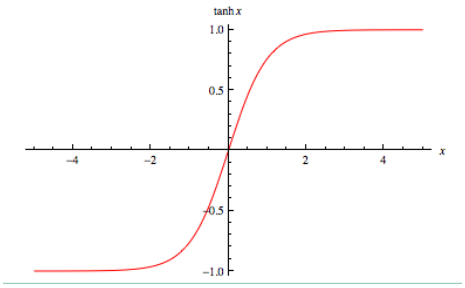
(i) $\theta(s)=\frac{1}{1+e^{-s}}=\frac{e^{s}}{e^{s}+1}$ ，

(ii) $\theta(-s)=1-\theta(s)$ ，

(iii) $\theta'(s)=\theta(s)\big(1-\theta(s)\big)$.



(2) 学习双曲正切函数$f(s)=\tanh s$，推导它与θ(s)的关系，推导$f'(s)$。



（3）如采用 tanh()作为多层感知器中隐节点的激活函数，试推导 BP 算法，并讨论为什么多层感知器一般不常用 tanh()作为激活函数。

【2022.3.8 → 2022.3.21】

4.3 计算机小实验：用 Softmax 回归进行人脸识别

在课上我们已经学习了如何使用 Logistic Regression 进行二分类。请大家阅读 Softmax Regression 并回答以下问题。

（1） Softmax Regression 是线性还是非线性分类器？请说明你的理由（提示：可以从位于分界面上的点入手分析）。

（2） 在附件中我们给出了 10 个人的人脸图像（数据来自 VGGface2，附件为 pictures.rar）。请用 Softmax Regression 设计分类器，实现以下要求：

- 请随机取出 2 个人的图像，75%作为训练集，25%作为测试集，给出 Softmax 的测试集正确率；同时，计算出 TPR, FPR, TNR, FNR, sensitivity, specificity, FDR；绘制 ROC 曲线，计算 AUC。
- 请随机取出 5 个人的图像，75%作为训练集，25%作为测试集，给出 Softmax 的测试集正确率。
- 请使用所有人的图像，75%作为训练集，25%作为测试集，给出 Softmax 的测试集正确率。

以上的测试中你的正确率是如何变化的，总结变化并给出合理解释。

提示：

本题提供的图片是彩色图片，请大家在进行分类前将图片转化为灰度图片，即大家在分类问题中处理的图片是 48x48 像素的灰度图片。如果你读入程序的图片是 3x48x48，请按照以下的提示对图片重新进行处理。

在 Python 中，你可以使用下面的方法来进行转换。Matlab 下请参考 rgb2gray 函数。你也可以手动编程实现，原理请参考该问题。其它语法请自行查询。

```
# 直接调用函数，在读取图片时将图片读取为灰度图片
from skimage import io
img = io.imread('image.png', as_grey=True)
```

在 Python 中，softmax 的实现函数是 sklearn 中的 LogisticRegression。设置 multi_class 参数为“multinomial”即可以使用 softmax 函数对每类的概率进行预测。

4.4 计算机小实验：用多层感知器进行人脸识别

请使用多层感知器进行作业 4.3 中同样的实验。讨论实验中多层感知器设计和训练中的因素对结果的影响(要求至少选择**三个**要素进行分析，如隐层节点数、隐层层数、激活函数、学习率等)，将本次实验结果与上次实验结果进行对比，简要分析两种分类器的异同。

第五章（支持向量机与统计学习理论）作业

【2022.3.15 → 2022.3.21】

5.1 写出考虑分类错误情况下广义最优分类面的原问题，推导出它的对偶问题。

5.2 【选做】查阅文献，讨论如果 SVM 使用不满足 Mercer 条件的核函数会带来什么问题。（可查阅文献资料，但须注明出处。）

5.3 试证明，R^d空间中的不加约束的线性分类函数（超平面）的 VC 维是 d+1。

【2022.3.15 → 2022.3.28】

5.4 计算机小实验：用 SVM 进行人脸识别

在作业 4.3 和 4.4 中，我们分别用 Softmax Regression 和多层感知器设计了人脸分类器，在本章中，请使用 SVM 进行同样的实验。讨论实验中核函数及其参数选择对结果的影响，将本次实验结果与之前的实验结果进行对比，简要分析三种分类器的异同。