

# NMDA 2019 第一次作业：线性模型与SVM

(推荐安装Anaconda软件包。每道题提交一个Jupyter Notebook文件，其中包括可直接运行的Python代码，以及注释和说明)

1. Iris flower dataset是机器学习领域的一个经典数据集，是英国统计学家Ronald Fisher于1936年整理建立的。数据集包括3种鸢尾花 (*setosa*, *virginica* 和 *versicolor*) 各50个样本，每个样本包含4个维度的特征 (Sepal length, Sepal width, Petal length, Petal width)。作业要求提交文件iris.ipynb。

- 1) 画出下面Wikipedia数据集链接中的散点图Scatterplot，观察讨论不同特征选择带来的不同可分性。
- 2) 采用线性回归Linear Regression方法，求得花瓣宽度 (Petal Width, x轴) 和花瓣长度 (Petal Length, y轴) 这两个特征之间的线性回归方程，并在这两维特征构成的样本平面上画出回归直线。以花萼宽度 (Sepal Width, x轴) 和花萼长度 (Sepal Length, y轴) 为两维特征，同样进行两者之间的线性回归和可视化。定量比较这两组线性回归的效果。
- 3) 选择Sepal length和 Sepal width这两个特征维度，采用Logistic Regression分类器实现Setosa和Versicolor (线性可分)，以及Versicolor和Virginica (线性不可分) 的二分类。要求可视化线性分类器迭代学习过程中的典型结果不少于5个，给出最后得到的分类器方程和分类准确率。

-数据集说明: [https://en.wikipedia.org/wiki/Iris\\_flower\\_data\\_set](https://en.wikipedia.org/wiki/Iris_flower_data_set)

-程序参考: [http://scikit-learn.org/stable/auto\\_examples/datasets/plot\\_iris\\_dataset.html](http://scikit-learn.org/stable/auto_examples/datasets/plot_iris_dataset.html)

2. 通过语音特征识别帕金森病的实验。请从下面的网址下载Parkinsons Data Set，其中包含31位受试者的语音数据23维特征。作业要求提交文件Parkinsons.ipynb。

- 1) 请仿照课堂上助教演示的乳腺癌检测的例子，把样本分成训练集和测试集，请分别采用Logistic Regression分类器 (L1范数作为正则项) 和线性SVM (推荐Scikit-learn中svm.SVC分类器) 对该数据进行二分类比较，比较这两种分类器的效果。并讨论两者之间的联系。
- 2) 采用线性核函数时，设法定量评估23维特征每一维对分类的贡献，挑选出对分类贡献最大的特征维。
- 3) 采用高斯核函数时，请通过GridSearch方法，寻找最有的SVM分类器参数Gamma和C，以及对应的分类准确率，并在二维坐标中把最优取值点的Gamma和C标出来。

-数据集说明: <https://archive.ics.uci.edu/ml/datasets/Parkinsons>

-程序参考网络学堂上的助教演示代码

3. 阅读Alan Turing写于1948年的论文Intelligent Machinery，翻译其摘要。并选择文中某一观点或者某一关于智能机器的设计进行评论 (300-500字)。