machine learning 笔记

徐世桐

1 基础定义

二元分类：输出分类个数为 2
多元分类：输出分类个数不限
$one-versus-the-rest$ OvR：计算属于每一分类的可能性，取可能性最大的分类为输出分类
$one-versus-one$ OvO：对所有分类两两使用二元分类，每一分类器训练只需一部分数据
multilabel 多标签分类：目标检测，对一图像中的物体加 label
multioutput 多类分类：多标签分类，每一标签可包含多种信息
learning schedule：根据迭代次数更新学习率
early stopping：提早结束训练
对于每一 epoch，当验证集 MSE 值增高时，证明开始 overfit，停止训练
即在 epoch-error 图中泛化误差最低时停止训练
在训练中使用正则化代价函数，训练结束后测试中代价函数不使用正则化项

2 数学计算

MSE $= \frac{1}{m}\sum_{i=1}^{m}(x^{(i)} - y^{(i)})^2$
rigid regression：回归方法，$J(\theta) = MSE(\theta) + \frac{\alpha}{2}\sum_i \theta_i^2$
降低所有权重值
lasso regression：回归方法，$J(\theta) = MSE(\theta) + \alpha\sum_i |\theta_i|$
降低不重要的权重值
elastic net：回归方法，$J(\theta) = MSE(\theta) + \gamma\alpha\sum_i |\theta_i| + (1-\gamma)\frac{\alpha}{2}\sum_i \theta_i^2$
Normal Equation：$\hat{\theta} = (X^TX)^{-1}X^Ty$
直接得到权重 $\hat{\theta}$，适用于仅有一个输出值的模型
X 为 (批量大小, 参数个数) 输入矩阵，y 为 (批量大小,) 向量
当 X^TX 无逆矩阵时，用 psudo inverse$\hat{\theta} = X^+y$
pseudo inverse:
对矩阵 $X = USV^T$，pseudo inverse $X^+ = VS^+U^T$。S^+ 求法：
1. 对所有 S 元素，接近 0 的值赋为 0
2. 对所有非零元素取倒数
3. 取矩阵转置，得到 S^+

log loss：代价函数
$J(\theta) = -\frac{1}{|B|}\sum_{i=1}^{|B|}[y^{(i)}log(\hat{p}^{(i)}) + (1-y^{(i)})log(1-\hat{p}^{(i)})]$

标签值 $y^{(i)}$ 为离散 1/0 值，计算值 $\hat{p}^{(i)} \in [0,1]$

微分：** 推导 **

$\frac{dJ(\theta)}{d\theta_j} = \frac{1}{|B|}\sum_{i=1}^{|B|}(\hat{p}^{(i)} - y^{(i)})x_j^{(i)}$

Gaussian Radial Basis Function RBF：一种 similarity function

$\phi_\gamma(x,l) = exp(-\gamma||x-l||^2)$

l 为 landmark，即 ϕ_γ 由一样本 x_i 和一 landmark 的距离得来

Lagrange multipliers method 拉格朗日乘数法

将 有前提的多项式求最值 问题转化为 无前提多项式最值问题

定义：

对输入向量 X，$C(X) \geq 0$ 为 constrain。目标为在满足 $C(X) \geq 0$ 的前提下取 $f(X)$ 最值

Lagrange function $\mathcal{L}(X,\alpha) = f(X) + \alpha(C(X))$

α 为变量

计算：

对每一 X 的元素 和 α 取偏导，即向量 $\begin{bmatrix} \frac{d\mathcal{L}(X,\alpha)}{dx_1} \\ \frac{d\mathcal{L}(X,\alpha)}{dx_n} \\ ... \\ \frac{d\mathcal{L}(X,\alpha)}{dx_n} \\ \frac{d\mathcal{L}(X,\alpha)}{d\alpha} \end{bmatrix}$，计算向量 $=\vec{0}$ 时的 X, α 取值

3 分类模型

logistic regression:

判断输入符合每一输出类别的可能性，

前向计算：

$1.\hat{p} = \sigma(\theta^T x + b)$

$2.\hat{y} = 1(if\hat{p} \geq 0.5)$

$= 0(if\hat{p} < 0.5)$

代价函数为 log loss

SVM

找到分界，分离多种数据

support vector: 最靠近分界线的样本

hard margin classification 硬性分类：限制数据必须被分界隔开，同一类数据不可同时出现在分界 2 端

soft margin classification：与硬性分类相反，避免被 outlier 离群值影响

前向计算：$\hat{p} = f(x_1, x_2, ...)$，其余同 logistic regression

区别：f 可为 polynomial，非线性函数。可使用 kernel trick

线性分类训练：$\hat{p} = W^T x + b$

硬性分类：

$||W||_2$ 代表线性函数斜率

最小化 $\frac{1}{2}W^T W$，使得分界平面的斜率最小，最大化分界线和两种数据的距离

前提：对每一样本 i, $1.y^{(i)}\hat{p}^{(i)} \geq 1$，即标签和计算结果相同

求解：使用拉格朗日乘数法，其中 α 改为向量，非常数。$\mathcal{L} = \frac{1}{2}W^TW - \sum_{i=1}^{|B|}\alpha^{(i)}(y^{(i)}\hat{p}^{(i)} - 1)$

使偏导向量为 $\vec{0}$，得到 $2.W = \sum_{i=1}^{m}\alpha^{(i)}\hat{p}^{(i)}x^{(i)}$, $3.\sum_{i=1}^{m}\alpha^{(i)}\hat{p}^{(i)} = 0$

带入得 $\mathcal{L}(W,\alpha) = \frac{1}{2}\sum_{i=1}^{|B|}\sum_{j=1}^{|B|}\alpha^{(i)}\alpha^{(j)}\hat{p}^{(i)}\hat{p}^{(j)}x^{(i)T}x^{(j)} - \sum_{i=1}^{|B|}\alpha^{(i)}$，解 α

解 W：由 α 带入 1. 式计算

解 b：由于所有 support vector$x^{(i)}$ 满足 1. 式，则对所有 support vector 计算 b 取平均值

$b = E(\hat{p}^{(i)} - W^Tx^{(i)})$

软性分类：

最小化 $\frac{1}{2}W^TW + C\sum_{i=1}^{|B|}\zeta_i$

ζ_i 定义第 i 样本被忽视为误差样本的可能性，C 定义忽视率相对斜率的权重

前提：对每一样本 i，$y^{(i)}\hat{p}^{(i)} \geq 1 - \zeta^{(i)}$

非线性分类方法：

- 使用 polynomial 做 f

- 使用 similarity function：

选择多个 landmark$\mathcal{L} = l_1, l_2, ..., l_n$，对每一样本 x_i 计算其和每一 l_j 的 ϕ_γ 值 $\phi_\gamma(x_i, l_j)$

每个样本用新的向量 $x_i' = \begin{bmatrix}\phi_\gamma(x_i,l_1)\\ \phi_\gamma(x_i,l_2)\\ ...\\ \phi_\gamma(x_i,l_n)\end{bmatrix}$ 表示。新的向量组成训练集，进行 SVM 训练

kernel：

定义：能够从输入向量 a,b，不通过计算 $\phi(a),\phi(b)$ 直接得到点乘结果 $\langle\phi(a),\phi(b)\rangle$ 的函数

例：** 是否通过取 linear 为 phi 得到 kernel 函数 **

linear: $f(a,b) = a^Tb$

polynomial: $f(a,b) = (\gamma a^Tb + r)^d$

Gaussian RBF: $f(a,b) = exp(-\gamma||a-b||^2)$

Sigmoid: $f(a,b) = tanh(\gamma a^Tb + r)$

4 决策树

定义：

节点 N_i：

节点条件：判断样本进入哪一子节点，叶节点没有节点条件

sample 属性 S_i：有多少样本**进入** N_i **节点**，非满足 N_i 节点条件的样本个数

value 属性 $V_i = v_{i1}, ..., v_{in}$：$S_i$ 进入节点的样本中 v_i 个属于第 i 分类

gini 属性 G_i：数据混杂度，$G_i = 1 - \sum_{j=1}^{n}(\frac{v_{ij}}{S_i})^2$

子节点仅有 2 个，对应节点条件为 true/false 的情况

CART algorithm 创建决策树：

根节点初始化为叶节点，没有节点条件

对每一叶节点 S_i 选取一特征 k，一特征门槛 t_k，将样本集分为 2 组 S_{true}, S_{false}。

选取 (k, t_k) 方式：使代价函数 $J(k,t_k) = \frac{S_{true}}{S_i}G_{true} + \frac{S_{false}}{S_i}G_{false}$ 最小

直到决策树层数达到固定上限，或分组没有降低 G** 如何计算 2 子节点的 G **

5 分析结果

confusion matrix 困惑矩阵：分析二元/多元分类

$$\begin{bmatrix} TN & FP \\ FN & TP \end{bmatrix}$$

一行对应同一期望输出，一列对应同一计算输出

T/F: 此位置的计算输出是否和预计输出一致

P/N: 此位置的预计输出是否为真

precision $= \frac{TP}{TP+FP}$

即 $P($ 计算结果匹配 | 计算结果为正 $)$

recall $= \frac{TP}{TP+FN}$

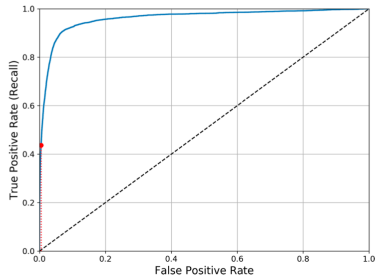
即 $P($ 计算结果匹配 | 预计结果为正 $)$

$F_1 = \frac{2}{\frac{1}{precision}+\frac{1}{recall}}$

precision 和 recall 的调和平均值

specificity $= \frac{TN}{TN+FN}$

ROC curve：分析二元/多元分类

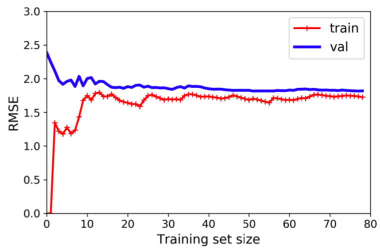


y 轴 recall 值，x 轴 false positive rate$FPR = \frac{FN}{FN+TN} = \frac{FN}{1-specificity}$

期望的 ROC curve 为 recall 从 0 快速增长到 1。并保持直到 FPR 为 1。

即期望曲线下方面积接近 1

learning curves：观察模型是否有 over underfit



x 轴为**一整次训练（包含多次 epoch）使用的训练集大小**，y 轴为 root MSE。

画出训练集 测试集在使用不同训练集大小后的 root MSE。

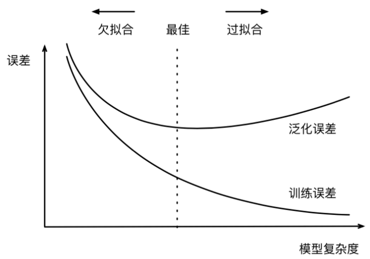
分析：

期望 2 曲线平缓值低且相近，

当 2 曲线平缓值差值较大，测试集平缓值较低，则过拟合

当 2 曲线平缓值较高，则欠拟合

模型复杂度-error epoch-error：



2 种图，形状类似，x 轴内容不同