机器学习笔记

# 机器学习基础

1. 概念：何为机器学习，将无序的数据转换成有用的信息
2. 数据获取：譬如可以在人们手机上装app，通过许多手机的磁力计得到信息
3. 术语：
   * 专家系统
   * 属性/特征
   * 分类
   * 目标变量(类别)
   * 训练数据和测试数据
4. 任务：
   1. 监督学习(知道预测什么，即目标变量的分类信息)
      * 分类：将实例数据划分到合适的分类中，譬如数据拟合曲线
      * 回归：主要用于预测数值型数据
   2. 无监督学习(数据无类别信息，不给目标值)
      * 聚类：数据集合分成类似对象组成的集合
      * 密度估计：寻找数据统计值的过程
      * 无监督学习还可以减少数据特征的维度，以便我们可以使用二维或三维图形展示数据
5. 开发步骤
   1. 收集数据
   2. 准备输入数据
   3. 分析输入数据
   4. 训练算法
   5. 测试算法
   6. 使用算法

# k-近邻算法(kNN)

## 概述

* 简单而言，k-近邻算法采用测量不同特征值之间的距离方法进行分类
* 工作原理：
  1. 存在一个样本数据集合(训练样本集)，其中每个数据存在标签
  2. 输入没有标签的新数据后，将新数据的每个特征与样本集中数据对应的特征进行比较
  3. 算法提取样本集中数据中最相似(最相邻)的分类标签，一般而言，我们只选择样本数据前k个最相似数据
  4. k即为k-近邻算法中k的出处，一般k取值不大于20

## 示例

### 场景一

收集到约会数据，保存在 'datingTestSet.txt' 中，每个样本数据占据一行，总共有1000行。 样本包含以下3种特征以及1种label：

* 每年获得的飞行常客里程数
* 玩视频游戏所耗时间百分比
* 每周消费冰淇淋公升数
* label分为三种：很喜欢，一般喜欢，不喜欢

#### 考虑

1. 算法采用 k 近邻算法
2. label可以依次用数字1,2,3代表很喜欢，一般喜欢以及不喜欢，假设这样的数据文件变为了 'datingTestSet2.txt'
3. 数据样本的前10%可以作为测试样本，剩余样本可以作为训练样本

#### 实现

1. 将数据读入，保存为一个 array 待用

* from numpy import \*  
    
   def file2matrix(filename):  
   fr = open(filename)  
   arrayOLines = fr.readlines()  
   m = len(arrayOLines)  
   returnMat = zeros((m,3))  
   classLabelVector = []  
   index = 0  
   for line in arrayOLines:  
   line = line.strip()  
   listFromLines = line.split('\t')  
   returnMat[index,:] = listFromLines[0:3]  
   classLabelVector.append(int(listFromLines[-1]))  
   index += 1  
   return returnMat, classLabelVector

1. 三种数据的数值范围差别过大，会导致对label的影响不一致，因此，需要对三种数据进行归一化

* from numpy import \*  
    
   def autoNorm(dataSet):  
   minVals = dataSet.min(0)  
   maxVals = dataSet.max(0)  
   ranges = maxVals - minVals  
   dataSetSize = dataSet.shape[0]  
   dataSet = dataSet - tile(minVals, (dataSetSize,1))  
   normDataSet = dataSet/tile(ranges, (dataSetSize,1))  
   return normDataSet, ranges, minVals

1. 算法实现

* from numpy import \*  
   import operator  
    
   def classify0(inX, dataSet, labels, k):  
   # 计算距离  
   m = dataSet.shape[0]  
   diffMat = tile(inX, (m, 1)) - dataSet  
   sqDiffMat = diffMat\*\*2  
   sqDistances = sqDiffMat.sum(axis=1)  
   distances = sqDistances\*\*0.5  
   # 将距离进行排序  
   # argsort() 函数，可以将 array 中的数值，按从小到大编号  
   sortedDistIndices = distances.argsort()  
   classCount = {}  
   for i in range(k):  
   # 前k个离测试点最近的label  
   voteIlabel = labels[sortedDistIndices[i]]  
   classCount[voteIlabel] = classCount.get(voteIlabel, 0) + 1  
   sortedClassCount = sorted(classCount.items(), key=operator.itemgetter(1), reverse=True)  
   return sortedClassCount[0][0]

1. 进行测试

* from numpy import \*  
    
   def datingClassTest():  
   # 前10%作为测试样本  
   hoRatio = 0.1  
    
   # 导入数据并归一  
   dataFile = '/home/curiousbull/Workspace/Python/machinelearninginaction/Ch02/datingTestSet2.txt'  
   datingDataMat, datingLabels = file2matrix(dataFile)  
   normMat, ranges, minVals = autoNorm(datingDataMat)  
    
   m = normMat.shape[0]  
   numTestVecs = int(m\*hoRatio)  
   errorCount = 0.0  
   for i in range(numTestVecs):  
   classifierResult = classify0(normMat[i,:], normMat[numTestVecs:m,:], \  
   datingLabels[numTestVecs:m], 3)  
   print("the classifier came back with: %d, the real answer is: %d" %(classifierResult, datingLabels[i]))  
   if(classifierResult != datingLabels[i]): errorCount += 1.0  
   print("the total error rate is: %f" %(errorCount/float(numTestVecs)))

### 场景二

目录digits下，分别有两个目录，testDigits和trainingDigits，代表用于测试的数据和训练的数据， 样本名包含需要识别的数字，也就是label，样本中是需要识别的手写数字数据集，每个单独文件都是32x32 的矩阵形式，包含 0 和 1

#### 考虑

1. 使用 k 近邻算法
2. 将文件名进行分割，得到对应的 labels 向量
3. 将每个文件内的32x32的0,1数字存储到一个vector，也就是一个1x1024的向量，作为识别的输入

#### 实现

1. 得到labels向量

* from numpy import \*  
   from os import listdir  
    
   def file2labels(dirName):  
   trainingFileList = listdir(dirName)  
   numberOfFiles = len(trainingFileList)  
   hwLabels = []  
   for i in range(numberOfFiles):  
   strTrainingFiles = trainingFileList[i].split('.')[0]  
   hwLabels.append(int(strTrainingFiles.split('\_')[0]))  
   trainingFileList[i] = dirName + trainingFileList[i]  
   return hwLabels, trainingFileList

1. 将img(32x32矩阵)转换为1x1024向量形式

* from numpy import \*  
   from os import listdir  
    
   def img2vec(filename):  
   fr = open(filename)  
   vecImg = zeros((1,1024))  
   for row in range(32):  
   strLine = fr.readline()  
   for col in range(32):  
   vecImg[0, 32\*row+col] = int(strLine[col])  
   return vecImg

1. 实现kNN算法

* from numpy import \*  
   from os import listdir  
   import operator  
    
   def classify0(inX, dataSet, labels, k):  
   dataSetSize = dataSet.shape[0]  
   diffMat = tile(inX, (dataSetSize, 1))-dataSet  
   sqDiffMat = diffMat\*\*2  
   sqDistances = sqDiffMat.sum(axis=1)  
   distances = sqDistances\*\*0.5  
   sortedDistIndicies = distances.argsort()  
   countLabels = {}  
   errorCount = 0.0  
   for i in range(k):  
   voteIlabel = labels[sortedDistIndicies[i]]  
   countLabels[voteIlabel] = countLabels.get(voteIlabel, 0) + 1  
   sortedCountLabels = sorted(countLabels.items(), key=operator.itemgetter(1),\  
   reverse=True)  
   return sortedCountLabels[0][0]

1. 测试样本

* def hwTest():  
   testDirName = 'digits/testDigits/'  
   trainingDirName = 'digits/trainingDigits/'  
   hwLabels, trainingFiles = file2labels(trainingDirName)  
   numberOfTrainingFiles = len(trainingFiles)  
   trainingDataSet = zeros((numberOfTrainingFiles, 1024))  
   for i in range(numberOfTrainingFiles):  
   trainingDataSet[i,:] = img2vec(trainingFiles[i])  
   testHwLabels, testFiles = file2labels(testDirName)  
   numberOfTestFiles = len(testFiles)  
   errorCount = 0.0  
   for i in range(numberOfTestFiles):  
   testVec = img2vec(testFiles[i])  
   classifierLabel = classify0(testVec, trainingDataSet, hwLabels, 3)  
   print("the classifer come back to: %d while the true value is: %d"\  
   %(classifierLabel, testHwLabels[i]))  
   if(classifierLabel != testHwLabels[i]): errorCount += 1.0  
   print("the total error classifiers number is: %f" %(errorCount/float(numberOfTestFiles)))  
   print("the error rate is: %f" %(errorCount/float(numberOfTestFiles)))

#### 注意点

1. 要使用 listdir() 函数，需要从 os 模块导入，但是不要将 os 模块中将所有内容都导入，以下代码

* from os import \*
* 会导致 open() 函数不可用

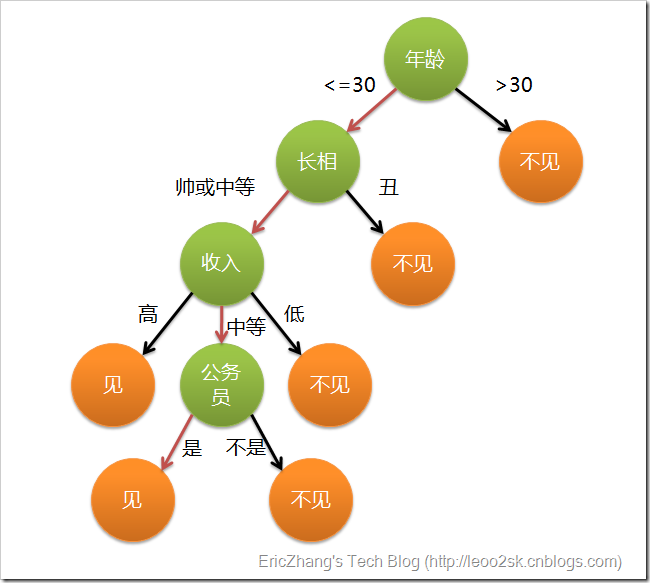
1. 注意 'labels' 从字符串中导入时，将字符串强转为 int 型

# 决策树

## 概述

#### 定义

决策树的分类思想类似姑娘找对象，相亲前，通过类似年龄、长相、收入，职业等将男人分为见或不见两类。类似下图：



从上图，可以总结出决策树的定义，即： 决策树（decision tree）是一个树结构（可以是二叉树或非二叉树）。其 每个非叶节点表示一个特征属性上的测试，每个分支代表这个特征属性在某个值域上的输出，而每个叶节点存放一个类别。 使用决策树进行决策的过程就是从根节点开始，测试待分类项中相应的特征属性，并按照其值选择输出分支，直到到达叶 子节点，将叶子节点存放的类别作为决策结果。

#### 构造

关键步骤在于分裂属性，即在某个节点处按照某一特征属性的不同划分构造不同的分支，目标是让各个分裂子集尽可能地 “纯”，即可能纯就是尽量让一个分裂子集中待分类项属于同一类别。分裂属性分三种情况：

1. 属性离散且不要求生成二叉树，此时用属性的每一个划分作为一个分支
2. 属性离散且要求生成二叉树，此时使用属性划分的一个子集作为测试，以“属于子集”和“不属于子集”分成两个分支
3. 属性值连续，确定一个值作为分裂点(split-point)，以 <split-point 和 >split-point 生成两个分支

#### ID3 算法(仅对于属性离散且不要求生成二叉树的情况讨论)

如果待分类的事物可能划分在多个分类中，譬如，设D为用类别对训练元组进行的划分

1. 符号 信息的定义：

* 其中表示第i个类别在整个训练元组出现的概率

1. D的熵(entropy)定义

* 熵的实际意义表示D中元组的类标号所需要的平均信息量。

1. 假设训练元组D按属性A划分，则该划分的熵计算如下：
2. 将D按属性A划分后，信息增益为两者的差值
3. 将D按所有属性进行划分，计算各种划分的信息增益，选择增益最大的属性进行划分，之后递归至
   * 划分后的所有分支内的元素都属于同一类别
   * 可用于划分的属性已使用完

## 示例

### 场景

有数据文件 'lenses.txt'，其中包含了如何预测患者需要佩戴的隐形眼镜类型数据。分为5列， 前四列为特征属性，最后一列为隐形眼镜的类型，也就是我们的 labels 特征依次为 '年龄(age)' '处方(prescript)' '是否散光(astigmatic)' '眼泪多少(tearRate)'

#### 考虑

1. 使用算法决策树解决问题
2. 数据导入(createDataSet())，返回数据(dataSet)和标签(labels)
3. 数据分割(splitDataSet())：按特征属性划分数据的时候，需要处理的数据(dataSet)是对应 该特征值的某个值的子dataSet
4. 考虑能够得到最大信息增益的特征，因此需要计算不同分割方法的熵(entropy)
5. 算法实现(createTree())：利用递归的方法，终止条件有两种：
   1. 按照当前特征划分的分支内所有元素都是同一类，即标签相同
   2. 用于划分的特征属性已经用完
6. 对应用于划分数据的特征属性用完后，如果某个分支内还有不同标签，处理方法可以效仿 KNN 算法，取该分支内占最多的label作为该分支的label
7. 测试

#### 实现

1. 数据导入 定义 createDataSet() 函数，返回 dataSet 和 labels 待用

* def createDataSet(filename):  
   fr = open(filename)  
   dataSet = [inst.strip().split('\t') for inst in fr.readlines()]  
   labels = ['age', 'prescript', 'astigmatic', 'tearRate']  
   return dataSet, labels

1. 数据分割 需要将数据中用于划分的列刨除

* def splitDataSet(dataSet, axis, value):  
   retDataSet = []  
   # 逐行输入  
   for featVec in dataSet:  
   if featVec[axis] == value:  
   reducedVec = featVec[0:axis]  
   reducedVec.extend(featVec[axis+1:])  
   retDataSet.append(reducedVec)  
   return retDataSet

1. 熵计算 对输入的dataSet，统计其所有事例数，统计不同label统计事例数 不同label统计事例数/总事例数=该label对应概率

* from math import log  
    
   def calcShannonEnt(dataSet):  
   numEntries = len(dataSet)  
   labelCount = {}  
   for featVec in dataSet:  
   currentLabel = featVec[-1]  
   labelCount[currentLabel] = labelCount.get(currentLabel, 0)+1  
   info = 0.0  
   for keys in labelCount.keys():  
   prob = float(labelCount[keys])/numEntries  
   info -= prob\*log(prob, 2)  
   return info

1. 挑选最佳分割路线，即确定最大信息增益的分割方式

* def chooseBestFeatToSplit(dataSet):  
   # 可用分割的特征数  
   numFeat = len(dataSet[0]) - 1  
    
   # 原始熵值  
   baseEntropy = calcShannonEnt(dataSet)  
    
   # 遍历可用特征，计算得到最大增益和与之匹配的特征编号  
   maxInfoGain = 0.0; bestFeat = -1  
   for i in range(numFeat):  
   featList = [example[i] for example in dataSet]  
   uniqueVals = set(featList)  
   newEntropy = 0.0  
   for value in uniqueVals:  
   subDataSet = splitDataSet(dataSet, i, value)  
   prob = len(subDataSet)/float(len(dataSet))  
   newEntropy += prob\*calcShannonEnt(subDataSet)  
   infoGain = baseEntropy - newEntropy  
   if infoGain > maxInfoGain:  
   maxInfoGain = infoGain  
   bestFeat = i  
    
   # 返回最大信息增益对应的特征编号  
   return bestFeat

1. 算法实现 利用熵计算，得到最佳的按特征分割方式，终止条件分为两种，如果是第二种终止条件，对该分支内 的元素采取类似kNN的做法，取多数的label作为该分支label
   1. 终止条件如果是用于分割数据的特征用完，需要定义一个处理分支内元素有不同label情况的函数
   * def majCnt(classList):  
      classCount = {}  
      for vote in classList:  
      classCount[vote] = classCount.get(vote, 0) + 1  
      # 注意，此处sortedLabelCount是一个tuple  
      sortedLabelCount = sorted(labelCount.items(), key=operator.itemgetter(1), reverse=True)  
      return sortedLabelCount[0][0]
   1. 利用递归实现 createTree() 方法，利用包含 dict 的 dict 的形式来存放 tree
   * def createTree(dataSet, labels):  
      # 首先应该考虑传入的数据是否还需要分割  
      # 获取数据集中所有标签  
      classList = [example[-1] for example in dataSet]  
      # 1. 分支内所有元素标签相同  
      if classList.count(classList[0]) == len(dataSet):  
      return dataSet  
       
      # 2. 所有可用分割特征使用结束  
      if len(dataSet[0]) == 1:  
      majCnt(classList)  
       
      # 以上终止条件不满足，对dataSet进行划分  
       
      # 最佳分割特征对应编号  
      bestFeat = chooseBestFeatToSplit(dataSet)  
      bestFeatLabel = labels[bestFeat]  
       
      # 利用特征定义树  
      myTree = {bestFeatLabel:{}}  
       
      # 将labels中此次最佳分割特征去除，在剩余特征中重新选择分割特征  
      del(labels[bestFeat])  
       
      # 该特征对应取值  
      featValues = [example[bestFeat] for example in dataSet]  
      uniqueValues = set(featValues)  
       
      # 按特征属性值进行分割  
      for value in uniqueValues:  
      subLabels = labels[:] # 尤其注意这句话的含义？  
      myTree[bestFeatLabel][value] = createTree(splitDataSet(dataSet, bestFeat, value), subLabels)  
       
      # 返回树  
      return myTree