woaidapaopao的博客

面试笔试整理4：机器学习面试问题准备（进阶）

2017年09月07日 16:35:20 阅读数：1746

这部分主要是针对上面问题的一些更细节的补充，包括公式的推倒思路、模型的基本构成、细节问题的分析等

一、问题

1、PCA的第一主成分

第二个主成分是第一成分方向正交的差异性次大方向。

2、什么时候使用集成的学习模型

只有当各个模型之间没有相关性的时候组合起来是最好用的。但是一般来说，弱相关的模型组合比较好用。

3、多重共线性

多重共线性是指两个特征的相关性很大的时候，会对参数模型造成非常大的影响。可以用相关分析判断多重共线性的存在性。

4、什么时候使用L2优于L1

如果多个变量是一些具有小尺度或中等尺度影响的时候用L2比较好，如果个别变量影响很大的时候用L1。其实也可以结合起来使用。

5、交叉验证的作用：选择问题

我们通常进行参数选择的时候都是用网格法做的，但是这样其实也是有弊端的，索性可以用随机取样的方式逼近最优参数。

6、如果缺失值超过30%要怎么办？

可以把缺失值单独组成一类。

二、模型流程和公式推导

1、PCA传统计算流程：

1. 去除均值
2. 计算协方差矩阵
3. 计算特征值和特征向量
4. 特征值从大到小排序
5. 保留前N个特征向量
6. 投影重构（记得吧去除的均值还回去）

或者干脆去均值后用SVD计算

2、离散数据下的生成模型

（1）贝叶斯概念

我们都知道概率学派和贝叶斯学派的不同，现在我们从贝叶斯的角度上考虑问题。对于一个问题，通常要考虑其先验概率，这是因为对于某些数据不足或有某些问题的情况下，单纯考虑似然函数是不够的，还需要引入假设先验给一个主观的先验概率，而且在真正分析的时候应该引入假设空间D的概念（满足要求的所有假设），后验就相当于给定假设空间D下的其中某一个假设D的概率P(h|D)。

其实本质上最大后验估计MAP是等价于最大似然估计的，即数据点足够多的时候会淹没先验。

利用得到的后验进行预测需要后验预测分布（Posterior pordictive distribution），方法是对每一个独立假设的加权均值（称之为Bayes model averaging）

我们使用MAP的时候都要对先验进行一些假设，而这些假设对应的先验函数和似然函数通常是**共轭**的，这样方便计算，关于共轭分布的概念其实很简单，常用的几个了解就可以。

（2）朴素贝叶斯分类器

朴素贝叶斯是最简单的分类器之一了，根本是假设各个特征之间是独立同分布的，也就是说P(X|y)=P(x1|y)*…P(xn|y)。我们可以假设特征x的分布，比如：在特征为real-value的时候，可以假设特征分布为高斯分布、在特征为二元特征的时候假设为Bernoulli分布、在类别特征的时候假设为multinoulli分布（我们通常见到的）。通常我们看到的Laplace平滑实际上是对参数的先验分布（但是这个先验可以只看出一个附加条件）。

3、Gaussian高斯模型的高斯判别分析
对于多元高斯分布来说，他的共轭分布也是多元高斯分布，关于多元高斯分布的最大似然结果可以自己查查资料。这里主要说的是高斯判别分析。
高斯判别分析假设$p(X,y=c,\theta)= N(X|\mu,\Sigma)$服从多元高斯分布，当$\Sigma$为对角矩阵的时候起始就是上面说的朴素贝叶斯了。我们通常说到的Linear discriminant analysis（LDA）其实就是高斯判别模型的一种，假设所有类别的协方差矩阵都是相同的，这时求解后验分布的时候得到的就是LDA。当然协方差矩阵不同的时候对应的QDA（Quadratic discriminant analysis,二次判别分析）。这个相当于我们对于通常定义LDA**最大化类间距最小化类内距离**实际上是等价的。

4、**Logistic regression**和指数分布族
这里将会从两个角度看一下逻辑回归的推导过程。
（1）逻辑回归推导
这个很简单，网上随便找一个都有，就是求解MLE而已。但是除了二元的逻辑回归还应该知道多元逻辑回归的条件概率由sigmoid变为softmax。
（2）逻辑回归的广义线性模型解释
首先要知道什么是广义线性模型：广义线性模型是指输出概率是指数分布族的y|x;θ~ExpoentialFamily(η)，而且指数分布族的自然参数η的是x的线性组合。这个我掌握的不是很好，但是如果面试的时候讲出来效果应该不错。
（3）逻辑回归输出值是不是概率
答案是肯定的，解释在这里，其实用广义线性模型的思路说更好，但是实在是对概念掌握的不好。

5、SVM支持向量机
（1）支持向量机公式推导，要详细到KKT条件。
（2）可以进一步结合核函数和GLM引出核机的概念。

6、概率图模型
有向图、无向

三、重要概念

1、监督学习的生成模型和判别模型
这可以说是一个最基础的问题，但是深挖起来又很复杂，面试的时候应该说出几个有亮点的部分。
（1）基本说法
生成模型是由数据学习联合概率分布P(X,Y)，然后再求出条件概率分布P(Y|X)，典型的生成模型有朴素贝叶斯和马尔科夫模型。
判别模型就是直接学习判别函数或者是条件概率分布，应该是更直接一些。两者各有优缺点。
（2）进阶区分
* 应该说生成模型的假设性更强一些，因为通常是从后验分布的角度思考问题，通常对x的分布进行了一些假设。
* 训练过程中，对于判别模型通常是最大化对数似然，对生成模型则是最大化联合对数似然函数
* 因为生成模型对于特征的分布都做出了一定的假设（如高斯判别模型假设特征分布满足多元高斯分布），所以如果对于特征的分布估计比较正确的情况下，生成模型的速度更好准确性也更高。
* 生成模型在训练数据的时候对于每一类数据的都是独立估计的（也就是每一类的参数不同），这也就说明如果有新类别加入的情况下，是不需要对原有类别进行重新训练的
* 对于半监督学习，生成模型往往更有用
* 生成模型有一个大的缺点就是不能对特征进行某些预处理（如特征映射），因为预处理后的数据分布往往有了很大的变化。

2、频率学派的一些基本理论
（1）期望损失（风险函数）、经验损失（经验风险）、结构风险
期望损失：理论上知道模型后得到的平均损失较期望损失（依赖于真实分布），但是模型正是我们要求的
经验损失：经验损失指针对模型的抽样值（训练集）进行平均的损失估计，根据大数定律当训练数据足够的时候经验损失和期望损失是等价的
结构风险：经验损失是假设经验分布和自然分布相同时得到的，但是这样会造成过拟合，所以引入了正则化，惩罚模型复杂度。
（2）极大似然MLE、极大后验MAP
因为我们有的时候利用经验损失求解的时候会遇到不好求解的问题（如不连续0-1）这是可以用对数极大似然估计等价的对参数进行分析。
同理最大后验利用先验概率达到惩罚模型的作用。如l2-norm岭回归对应高斯先验、L1对应拉普拉斯先验。

文章标签： 深度学习 面试
个人分类： 机器学习 深度学习 学习数据挖掘进程
相关热词： 华为面试笔试 图像处理面试笔试 fpga面试笔试 面试笔试数据结构 腾讯面试笔试

上一篇 面试笔试整理3：深度学习机器学习面试问题准备（必会）
下一篇 面试笔试整理5：项目问题准备