

机器学习 讨论题

讨论小组： 2班 第二小组 （序号22~41）

Regression T4.2 Lasso回归和岭回归各有什么特点，适用于怎样的场合？

答： Lasso回归采用L1正则化，能稀疏矩阵，进行庞大特征数量下的特征选择。适用于高维特征数据，尤其是线性关系为稀疏的。

岭回归采用L2正则化，能够降低参数范数的总和，平滑权重，有效的防止模型过拟合，解决非满秩下求逆困难的问题。适用于非满秩下求逆困难的情况。

Ensemble T12-2为何通常集成学习的预测性能会优于其他的机器学习算法？

答： 集成学习是将若干个学习器组合之后产生的一个学习器，它能够保证弱分类器的多样性，集成不稳定的算法也能得到一个比较明显的性能提升。弱分类器间存在一定的差异性，这会导致分类的边界不同，也就是说可能存在错误。那么将多个弱分类器合并后，就可以得到更加合理的边界，减少整体的错误率，实现更好的效果。

对于数据集过大或者过小的情况，可以分别进行划分和有放回的操作产生不同的数据子集，然后使用数据子集训练不同的分类器，最终再合并成为一个大的分类器。

如果数据的划分边界过于复杂，可以使用线性模型很难描述情况，那么可以训练多个模型，然后再进行模型的融合。对于多个异构的特征集的时候，很难进行融合，可以考虑每个数据集构建一个分类模型，然后将多个模型融合。因此集成学习的预测性会优于其他机器学习算法。

Deep learning T13-1 卷积神经网络和多层感知机之间有什么异同？

答： 相同点：

- 卷积神经网络由多层感知机发展而来，每个隐藏层的神经元/卷积核都在提炼学习特征，越深层提炼出的特征越抽象和复杂

不同点：

- 卷积神经网络引入卷积核提取特征，重视空间信息的保留。多层感知机则不然
- 卷积神经网络的感受野里只有局部小特征，多层感知机只有全局大特征
- 卷积神经网络多用于图像处理，其理解图片是一块一块比对小特征的过程，多层感知机则是全局性大特征比对

T14.2

朴素贝叶斯分类是以**贝叶斯定理**为基础并且假设特征条件之间相互独立的方法，先通过已给定的训练集，以特征词之间独立作为前提假设，学习从输入到输出的联合概率分布，再基于学习到的模型，输入 X 求出使得后验概率最大的输出 Y

设有样本数据集 $\{d_1, d_2, \dots, d_n\}$, 对应样本数据的特征属性集为

$$X = \{x_1, x_2, \dots, x_d\}$$

类变量为

$$Y = \{y_1, y_2, \dots, y_m\}$$

, 即 D 可以分为 y_m 类别。其中 x_1, x_2, \dots, x_d 相互独立且随机, 则 Y 的**先验概率** $P_{prior} = P(Y)$ Y 的后验概率 $P_{post} = P(Y|X)$, 由朴素贝叶斯算法可得, 后验概率可以由先验概率 $P_{prior} = P(Y)$ 、证据 $P(X)$ 、类条件概率 $P(X|Y)$ 计算出:

$$P(Y|X) = \frac{P(Y) P(X|Y)}{P(X)}$$

朴素贝叶斯基于各特征之间相互独立, 在给定类别为 y 的情况下, 上式可以进一步表示为下式:

$$P(X|Y = y) = \prod_{i=1}^d P(x_i|Y = y)$$

由以上两式可以计算出后验概率为:

$$P_{post} = P(Y|X) = \frac{P(Y) \prod_{i=1}^d P(x_i|Y)}{P(X)}$$

由于 $P(X)$ 的大小是固定不变的, 因此在比较后验概率时, 只比较上式的分子部分即可。因此可以得到一个样本数据属于类别 y_i 的朴素贝叶斯计算:

$$P(y_i|x_1, x_2, \dots, x_d) = \frac{P(y_i) \prod_{j=1}^d P(x_j|y_i)}{\prod_{j=1}^d P(x_j)}$$

产生式模型与判别式模型 T4 产生式模型和判别式模型各有哪些应用？

答：判别式模型（Discriminative Model）是直接对条件概率 $p(y|x;\theta)$ 建模。常见的判别式模型有线性回归模型、线性判别分析、支持向量机SVM、神经网络、boosting、条件随机场等。生成式模型（Generative Model）则会对 x 和 y 的联合分布 $p(x,y)$ 建模，然后通过贝叶斯公式来求得 $p(y|x)$ ，然后选取使得 $p(y_i|x)$ 最大的 y_i ，常见的生成式模型有隐马尔可夫模型HMM、朴素贝叶斯模型、高斯混合模型GMM、LDA、高斯、混合多项式、专家的混合物、马尔可夫的随机场

PCA T2 PCA的核心思想是什么？

答：PCA核心思想是将原始 n 维特征投影到 k 维上，这投影后的 k 维数据之间包含了原数据绝大多数信息。

先求出原数据的 n 阶协方差矩阵，其主对角线上的元素是各个维度上的方差(即信息量)，其他元素是两两维度间的协方差(即相关性)。通过将协方差矩阵对角化，除去不同维度之间的相关性，使得各个维度上自身的方差最大化。提取方差（信息量）最大的 k 维特征，便完成了PCA的过程。