机器学习基本概念

cloud

2016.12.27

1 概念

1.0/1 损失函数

当样本点被正确分开在间隔之外损失为 0，否则为 1。合页（hinge）损失函数相当于线性支持向量机原始最优化问题。

2.AdaBoost

AdaBoost 是基学习器的线性组合，只适应于二分类任务。它提高被前一轮弱分类器错误分类样本的权值，而降低被正确分类样本的权值，使得分类错误的样本在后一轮分类器中得到更多的关注。它采用加权多数表决的方法，即加大分类误差率小的弱分类器的权值，使其在表决中起较大的作用，减小分类误差率大的弱分类器权值，使其在表决中起较小的作用。

Boosting

先从初始训练集训练出一个基学习器，再根据基学习器的表现对训练样本的分布进行调整，使得先前基学习器做错的训练样本在后续受到更多关注，然后基于调整后的样本分布来训练下一个基学习器；如此重复进行，直至基学习器数目达到事先指定的值 T，最终将这 T 个基学习器进行加权结合，主要关注降低偏差。

3.Bagging

对于包含 m 个样本的数据集，有放回的采样 m 个样本并重复 T 次。将这 T 个采样集训练出的基学习器进行结合。通常对分类任务采用简单投票法，对回归任务采用简单平均法，主要关注降低方差。

随机森林

RF 在 Bagging 基础上，进一步在决策树训练过程中引入随机属性的选择。传统决策树在选择划分属性时是在当前结点的属性集合中选择一个最优属性，而在 RF 中，对基决策树的每个结点，先从该结点的属性集合中随机选择一个包含 k 个属性的子集，然后从这个子集中选择一个最优属性进行划分，推荐 $k = logd$。

4.ID3/C4.5/CART

ID3　信息增益：$g(D,A)=H(D)-H(D|A)$

C4.5　信息增益比：$g_R(D,A)=\frac{g(D,A)}{H_A(D)}$　$H_A(D)=-\sum_{i=1}^{n}\frac{|D_i|}{|D|}log\frac{D_i}{D}$，其中 n 是特征 A 的取值数目

CART　基尼系数：$Gini(D,A)=\sum_{i=1}^{n}\frac{|D_i|}{|D|}Gini(D_i)$　$Gini(D)=1-\sum_{k=1}^{K}(\frac{|C_k|}{|D|})^2$，其中 n 是特征 A 的取值数目，K 是类的个数。

5.EM 算法

第一步是期望 E 步，利用当前估计的参数值来计算对数似然的期望值；第二步是最大化 M 步，寻找能使 E 步产生的似然期望最大化的参数值，然后新得到的参数值重新被用于 E 步直到收敛到局部最优解。

6.F1

$F1$ 是基于查准率和查全率的调和平均：$\frac{1}{F_1}=\frac{1}{2}(\frac{1}{P}+\frac{1}{R})$，查准率/准确率是指预测结果为正例中实际为正例的比例，查全率/召回率是指实际是正例中被预测出来的比例。

ROC

ROC 是以横轴为假正率（真实为假例的样本中被预测成正例的比例），纵轴为真正率（真实为正例的样本中被预测出来正例的比例）曲线下面积。

7.Fisher 判别分析/LDA

给定训练样例集，设法将样例投影到一条直线上，使得同类样本的投影点尽可能接近，异类样例的投影点尽可能远离。

8.Jensen 不等式

对于任意凸函数 $f(x)$，有 $f(E(x))\leqslant E(f(x))$。

9.KTT 条件

对原始问题和对偶问题，则原始问题和对偶问题有相同解的充分必要条件是节满足下面的 KTT 条件。

10.KL 散度/相对熵/信息散度

用于度量两个概率分布之间的差异。

11.k 折交叉验证

将数据集划分成 k 个大小相似的互斥子集，每次用 $k-1$ 个子集的并集作为训练集，余下的哪个子集作为测试集；最终返回 k 个测试结果的均值。

12.kNN

给定测试样本，基于某种距离度量找出训练集中与其最靠近的 k 个训练样本，然后基于这 k 个邻居的信息来进行预测；通常在分类任务中采用投票法，在回归任务中使用平均法；还可以基于距离远近进行加权平均或加权投票，距离越近样本权重越大。

13.k-means

从数据集中随机选择 k 个样本作为初始均值向量，计算样本与各均值向量的距离然后将其划入最近的簇；重新取平均计算均值向量，循环直至均值向量不变。

14.L_1 正则化

求的是系数绝对值之和，对应的回归模型是岭 $(ridge)$ 回归，更容易求得稀疏解，降低过拟合风险。

15.L_2 正则化

求的是系数平方和后开根，对应的回归模型是 $lasso$ 回归，降低过拟合风险。

16.MCMC

首先设法构造一条马尔可夫链，使其收敛至平稳分布恰为待估计参数的后验分布，然后通过这条马尔可夫链来产生符合后验分布的样本，并基于这些样本来进行估计，这里马尔可夫转移概率的构造至关重要，不同的构造方法将产生不同的 $MCMC$ 算法。

17.PCA

对所有样本进行中心化（即当前样本向量减去样本向量的平均）；计算样本的协方差矩阵，对协方差矩阵做特征值分解；取最大的 d 个特征值对应的特征向量，输出这 d 个特征向量组成的投影矩阵。

18.Stacking

先从初始数据集训练出初级学习器，然后生成一个新数据集用于训练次级学习器，在这个新数据集中，初级学习器的输出被当作样例输入特征，而初始样本的标记仍被当作样例标记。

将训练集 D 划分为 k 个大小相似的集合，对其中的每个集合用 T 个初级学习算法训练训练集得到 T 个模型，然后用这 T 个模型预测每个训练集对应的验证集的样本，每个样本得到预测结果 T 个，将预测的 T 个结果和验证集真实结果作为次级训练集。

19. 奥卡姆剃刀

若有多个假设与观察一致，则选最简单的那个。

20. 半监督学习

让学习器不依赖外界交互，自动地利用未标记样本来提升学习性能。

21. 层次聚类

试图在不同层次对数据集进行划分，从而形成树形聚类结构，数据集的划分分为自底向上和自顶向下的策略。自底向上首先把每个样本看成初始聚类簇，然后在算法运行的每一步找出距离最近的两个聚类簇进行合并并不断循环，直至达到预设的聚类簇个数；自顶向下过程相反。

22. 重赋权法

在训练过程的每一轮中，根据样本分布为每个训练样本重新赋予一个权重。

重采样法

在每一轮学习中，根据样本分布对训练集重新进行采样，再用重采样而得的样本集对基学习器进行训练。

23. 狄利克雷（Dirichlet）分布

是关于一组 d 个连续变量 μ_i 属于 $[0,1]$ 且加和为 1 的概率分布。当 d 为 2 时退化为贝塔 $(Beta)$ 分布。

24. 对偶问题

将约束最优化问题表示为广义拉格朗日函数的极大极小问题，称为原始问题的对偶问题，定义对偶问题的最优值为原始问题的解。

25. 多变量决策树

多变量决策树是将多个变量组合起来作为条件进行划分。

26. 多分类

即将多分类任务拆为若干个二分类任务，通过投票选择测试样本属于的类别，有一对一，一对多，多对多的情况。

27. 二次规划

是一类典型的优化问题，包括凸二次优化和非凸二次优化，在此类问题中，目标函数是变量的二次函数，而约束条件是变量的线性不等式。

28. 泛化

学得模型适应于新样本的能力，称为泛化能力。

29. 分层采样

保留类别比例的采样方式通常称为分层采样。

30. 概率模型

将学习任务归结于计算变量的概率分布，在概率模型中利用已知变量推测未知变量的分布称为推断，其核心是如何基于可观测变量推断出未知变量的条件分布。假定所关心的变量集合为 Y，可观测变量集合为 O，其他变量的集合为 R，生成式模型考虑联合分布 $P(Y,R,O)$，判别式模型考虑条件分布 $\mathsf{P}(Y,R|O)$，给定一组观测变量值，推断就是要由 $P(Y,R,O)$ 或 $P(Y,R|O)$ 得到条件概率分布 $\mathsf{P}(Y|O)$。

31. 概率图模型

是一类用图来表达变量相关关系的概率模型，它以图为表示工具，最常见的是用一个结点表示一个或一组随机变量，结点之间的边表示变量间的概率相关关系，分为有向无环图表示变量间的依赖关系和无向图表示变量间的相关关系。

32. 感知机学习算法

当一个实例点被误分类，即位于分离超平面错误一侧时，即调整 w,b 的值，使分离超平面向该误分类点的一侧移动，以减少该误分类点与超平面的距离，直至超平面越过该误分类点使其被正确分类。

33. 高斯混合模型

该分布共由 k 个混合成分组成，每个混合成分对应一个高斯分布。假设样本的生成过程由高斯混合分布给出，首先根据混合系数定义先验分布选择高斯混合成分，即每个混合系数表示选择其对应高

斯分布的概率，然后根据被选择的混合成分概率密度函数进行采样从而生成相应的样本。

34. 规范化

是将不同变化范围的值映射到相同的固定范围中，常见的是 $[0,1]$，此时亦称归一化。

35. 吉布斯采样

假定 $X = x^{(1)}, x^{(2)}, ..., x^{(N)}$，目标分布为 $p(X)$，在初始化 X 的取值后，通过循环执行以下步骤来完成采样：随机或以某个次序选取某变量 $x^{(i)}$；根据 X 中除 $x^{(i)}$ 外的变量的现有取值，计算条件概率 $p(x^{(i)}|X - x^{(i)})$；根据 $p(x^{(i)}|X - x^{(i)})$ 对变量 $x^{(i)}$ 采样，用采样值替代原值。

36. 结构风险

结构风险用于描述模型的复杂度，经验风险用于描述模型与训练集的契合程度；由于结构风险引出正则化问题，范数是常用的正则化项，L_2 范数倾向于 w 的分量取值尽量均衡，即非零分量个数尽量稠密，而 L_0 和 L_1 范数则倾向于 w 的分量尽量稀疏，即非零分量个数尽量少。

37. 拉格朗日乘子法

是一种寻找多元函数在一组约束下的极值的方法，通过引入拉格朗日乘子，可将有 d 个变量与 k 个约束条件的最优化问题转化为具有 $d + k$ 个变量的无约束优化问题。

38. 拉普拉斯修正

为了避免其它属性携带的信息被训练集中未出现的属性值抹去，在估计概率值时通常要进行平滑，即在原来求概率的基础上分子加 1 分母加 N，在原来求条件概率的基础上分子加 1 分母加对应属性的数目 N_i。

39. 密度聚类

从样本密度的角度来考察样本之间的可连续性，并基于可连接样本不断扩展聚类簇以获得最终的聚类结果。

40. 模拟退火

模拟退火在每一步都以一定的概率接受比当前解更差的结果，从而有助于跳出局部最小，在每步迭代的过程中，接受次优解的概率要随着时间的推移而逐渐降低，从而保证算法稳定。

41. 判别式模型

给定 x，通过直接建模 $P(c|x)$ 来预测 c，这样得到的是判别式模型；先对联合概率分布 $P(x, c)$ 建模，然后由此获得 $P(c|x)$, 这样的到的是生成式模型。

42. 泛化误差

可分解为偏差，方差与噪声之和；偏差度量了学习算法的期望预测真实结果的偏离程度，即刻画了学习算法本身的拟合能力；方差度量了同样大小的训练集变动所导致学习性能的变化，即刻画了数据扰动所造成的影响；噪声则表达了在当前任务学习上任何学习算法所能达到的期望泛化误差的下界，即刻画了学习问题本身的难度。

43. 奇异值分解

任意实矩阵 A 都可分解为 $A = U\Sigma V$。

44. 软间隔

所有样本都必须划分正确称为硬间隔，软间隔是允许某些样本不满足约束。

45. 梯度下降法

是一种常见的一阶优化方法，是求解无约束优化问题最简单经典的方法之一。

46. 条件随机场（CRF）

是一种判别式无向图模型，条件随机场试图对多个变量在给定观测值后的条件概率进行建模。

47. 有序属性

就是属性之间是有距离度量，可以直接在属性值上计算距离的叫作有序属性；不能在离散属性上直接计算属性值距离的叫作无序属性。

48. 支持向量回归

假设我们能容忍 $f(x)$ 与 y 之间最多有 c 的偏差，即仅当 $f(x)$ 与 y 之间的差别绝对值大于 c 时才计算损失；这相当于以 $f(x)$ 为中心，构建了一个宽度为 $2c$ 的间隔带，若训练样本落入此间隔带，则认为是被预测正确的。

49. 最小二乘法

基于均方误差最小化来进行模型求解的方法称为最小二乘法。

50. 坐标下降法

是一种非梯度优化方法，它在每步迭代中沿一个坐标方向进行搜索，通过循环使用不同的坐标方向来达到目标函数的局部最小。

2 参考文献

1. 周志华的机器学习

2. 李航博士的统计学习方法