统计机器学习
Statistical Machine Learning

魏莱

上海海事大学信息工程学院

2019 年 4 月 4 日

第十一章：特征选择与稀疏学习

1. 子集搜索与评价

前一章，我们谈论了维数约简问题。所谓的维数约简，是在保持数据集某种结构的情况下，去寻找一个低维的子空间，在这个子空间中，原数据样本的低维嵌入能够更好的揭示数据集的某种结构信息。

总结关于维数约简的两个特点：

1. 假设样本 $\mathbf{x} \in \mathbb{R}^D$ 的低维嵌入 $\mathbf{y} \in \mathbb{R}^d$，是其在低维空间中的一种表示，并且 $d \ll D$;
2. 由于 $\mathbf{y} = \mathbf{W}^t\mathbf{x}$，$\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_d] \in \mathbb{R}^{D\times d}$ 为投影矩阵，则 $y_i = \mathbf{w}_i^t\mathbf{x} = \sum_{j=1}^{D} w_{ij}x_j$。可见，$\mathbf{y}$ 的第 i 个分量（特征）是由 $\mathbf{x}$ 的分量（特征）线性组合而成。

由此可见，维数约简可以看做是对原数据样本的特征进行了“抽取”，选择其中某些重要的（或其组合）来表示原样本。

一个学习任务来说，给定属性集，其中有些属性（特征）可能很关键、很有用，另一些属性（特征）则可能没什么用。对当前学习任务有用的属性称为“**相关特征**”**(relevant feature)**，没什么用的属性称为“**无关特征**”**(irrelevant feature)**。还有一类特征称为“**冗余特征**”**(redundant feature)**，它们所包含的信息能从其他特征中推演出来。

1. 子集搜索与评价

从给定的特征集合中选择出相关特征子集的过程，称为“**特征选择**”**(feature selection)**。特征选择是一个重要的“**数据预处理**”**(data preprocessing)**过程，在现实机器学习任务中，获得数据之后通常先进行特征选择，然后再训练学习器。这么做的原因有两个：

1. 在现实任务中由于属性过多，会遇到维数灾难问题，若能从中选择出重要的特征，使得后续学习过程仅需在部分特征上构建模型，则维数灾难问题会大为减轻。
2. 去除不相关特征往往会降低学习任务的难度。

需注意的是，特征选择过程必须确保不丢失重要特征，否则后续学习过程会因为重要信息的缺失而无法获得好的性能。

1. 子集搜索与评价

从初始的特征集合中选取一个包含了所有重要信息的特征子集，在没有任何领域知识作为先验假设的情况下，只能遍历所有可能的子集。但这样做会遭遇组合爆炸，特征个数稍多就无法进行。可行的做法是产生一个“候选子集”，评价出它的好坏，基于评价结果产生下一个候选子集，再对其进行评价，这个过程持续进行下去，直至无法找到更好的候选子集为止。

从上面阐述可以看出，特征选择包括两个关键环节：如何根据评价结果获取下一个候选特征子集（子集搜索）？如何评价候选特征子集的好坏（子集评价）？

1. 子集搜索与评价

1. 子集搜索。子集搜索有“前向”（forward）搜索以及“后向”（backward）搜索。假设给定特征集合 $\{a_1, a_2, \cdots, a_D\}$,

   A. 前向搜索方法：

   STEP 1. 将每个特征看做一个候选子集，对每个子集进行评价，选择最优子集，作为候选子集；

   STEP 2. 在原属性集合中，选择一个不属于候选子集的属性，加入候选子集；

   STEP 3. 若加入新属性后，候选子集性能没有增加，跳出循环，否则跳转 STEP 2。

B. 后向搜索方法：

STEP 1. 将所有特征组成的集合看做一个候选子集；

STEP 2. 在候选集合中，删除一个属性；

STEP 3. 若删除属性后，候选子集性能降低，跳出循环，否则跳转 STEP 2。

注：特征选择的后向搜索方法，有时也称为**属性约简（attribute reduction）**。

1. 子集搜索与评价

从上面子集搜索的前向及后向策略可以看出，算法需要评价候选子集的性能，那么如何进行子集评价呢？实际上，我们在决策树一章中已经谈到过这个问题。

给定数据集 $\mathbf{D}$，条件属性集合 $\mathbf{A} = \{A_1, A_2, \cdots, A_m\}$。在决策树中，计算属性 A_k 的信息增益：

$$I_{A_k} = H(\mathbf{D}) - H(\mathbf{D}|A_k), \tag{1}$$

其中，$H(\mathbf{D}|A_k) = \sum_{i=1}^{v_k} \frac{|\mathbf{D}^{A_k^i}|}{\mathbf{D}} H(\mathbf{D}^{A_k^i})$，$A_k$ 具有 v_k 个可能的取值，$\mathbf{D}^{A_k^i}$ 为属性 $A_k = v_k^i$ 的样本集合。

假设候选子集为 $\tilde{\mathbf{A}} \subseteq \mathbf{A}$，那么根据 $\tilde{\mathbf{A}}$ 可以将数据集进行划分，因此可以计算 $\tilde{\mathbf{A}}$ 的信息增益：

$$I_{\tilde{\mathbf{A}}} = H(\mathbf{D}) - H(\mathbf{D}|\tilde{\mathbf{A}}), \tag{2}$$

其中，$H(\mathbf{D}|\tilde{\mathbf{A}}) = \sum_{i=1}^{k} \frac{|\mathbf{D}^k|}{\mathbf{D}} H(\mathbf{D}^k)$，$\mathbf{D}^k$ 为根据子集 $\tilde{\mathbf{A}}$ 将数据集进行划分得到的第 k 个子集。

若学习任务为无监督学习，则上式变成：

$$I_{\tilde{\mathbf{A}}} = H(\mathbf{A}) - H(\mathbf{A}|\tilde{\mathbf{A}}), \tag{3}$$

1. 子集搜索与评价

将特征子集搜索机制与子集评价机制相结合，即可得到特征选择方法。常见的特征选择方法大致可分为三类: 过滤式 (filter)、包裹式 (wrapper) 和嵌入式 (embedding)。

2. 过滤式特征选择

过滤式方法先对数据集进行特征选择，然后再训练学习器，特征选择过程与后续学习器无关。这相当于先用特征选择过程对初始特征进行“过滤”，再用过滤后的特征来训练模型。过滤式方法主要各种统计学方法，比如 P-value 等等。

2. 包裹式特征选择

与过滤式特征选择不考虑后续学习器不间，包裹式特征选择直接把最终将要使用的学习器的性能作为特征子集的评价准则。换言之，包裹式特征选择的目的就是为给定学习器选择最有利于其性能、“量身走做”的特征子集。

一般而言，由于包裹式特征选择方法直接针对给定学习器进行优化，因此从最终学习器性能来看，包裹式特征选择比过滤式特征选择更好；但另一方面，由于在特征选择过程中需多次训练学习器，因此包裹式特征选择的计算开销通常比过滤式特征边择大得多。

3. 嵌入式特征选择与 L_1 正则化

在过滤式和包裹式特征选择方法中，特征选择过程与学习器训练过程有明显的分别。与此不同，嵌入式特征选择是将特征选择过程与学习器训练过程融为一体，两者在同一个优化过程中完成，即在学习器训练过程中自动地进行了特征选择。

给定数据集 $\mathbf{D} = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)\}$，其中 $\mathbf{x}_i \in \mathbb{R}^D, y_i \in \mathbb{R}$，$y_i$ 是样本 $\mathbf{x}_i$ 的类别标签。考虑平方误差损失函数：

$$\min_{\mathbf{w}} \sum_{i=1}^{n} (y_i - \mathbf{w}^t \mathbf{x}_i)^2 \tag{4}$$

若能够求得 $\mathbf{w}$，则对新样本 $\mathbf{x}^*$ 其类别标签 $y^* = \mathbf{w}^t \mathbf{x}^*$。因此，$\mathbf{w}$ 一方面可以看做用于特征选择，另一方面用于分类。

3. 嵌入式特征选择与 L_1 正则化

按照第一章内容，式（4）容易导致过拟合。为了降低 $\mathbf{w}$ 的复杂性，常常对 $\mathbf{w}$ 加上一定正则算子，如 L_2 范数正则以及 L_1 范数正则化：

$$\min_{\mathbf{w}} \sum_{i=1}^{n} (y_i - \mathbf{w}^t \mathbf{x}_i)^2 + \lambda \|\mathbf{w}\|_2^2 \tag{5}$$

$$\min_{\mathbf{w}} \sum_{i=1}^{n} (y_i - \mathbf{w}^t \mathbf{x}_i)^2 + \lambda \|\mathbf{w}\|_1 \tag{6}$$

（5）式很容易求解，现在来讨论（6）式求解。

考虑一般性问题：

$$\min_{\mathbf{w}} f(\mathbf{w}) + \lambda \|\mathbf{w}\|_1 \tag{7}$$

将 $f(\mathbf{w})$ 在 $\mathbf{w}_k$ 附近 Taylor 展开，得到

$$\begin{aligned} f(\mathbf{w}) &\simeq f(\mathbf{w}_k) + \langle \nabla f(\mathbf{w}_k), \mathbf{w} - \mathbf{w}_k \rangle + \frac{L}{2} \|\mathbf{w} - \mathbf{w}_k\|^2 \\ &= \frac{L}{2} \|\mathbf{w} - (\mathbf{w}_k - \frac{1}{L} \nabla f(\mathbf{w}_k))\|_2^2 + const \end{aligned} \tag{8}$$

其中 $const$ 是与 $\mathbf{w}$ 无关的常数，$\langle \cdot, \cdot \rangle$ 表示内积运算。显然，上式在如下 $\mathbf{w}_{k+1}$ 处有最小值：

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \frac{1}{L} \nabla f(\mathbf{w}_k)) \tag{9}$$

于是，借助于梯度下降法思想，最小化式（7），可以在更新 $\mathbf{w}$ 时，同时考虑 L_1 范数约束。令 $\mathbf{z} = \mathbf{w}_k - \frac{1}{L}\nabla f(\mathbf{w}_k)$，$\mathbf{w}_{k+1}$ 应该满足：

$$\mathbf{w}_{k+1} = \arg\min_{\mathbf{w}} \frac{L}{2}\|\mathbf{w} - \mathbf{z}\|_2^2 + \lambda\|\mathbf{w}\|_1 \tag{10}$$

将 $\mathbf{w}$ 按分量展开，可得

$$\mathbf{w}_{k+1}^i = \begin{cases} \mathbf{z}^i - \lambda/L, & \lambda/L < \mathbf{z}^i; \\ 0, & |\mathbf{z}^i| \le \lambda/L; \\ \mathbf{z}^i + \lambda/L, & \mathbf{z}^i < -\lambda/L. \end{cases} \tag{11}$$

其中 $\mathbf{w}_{k+1}^i, \mathbf{z}^i$ 分别是 $\mathbf{w}_{k+1}, \mathbf{z}$ 的第 i 个分量。

假设数据集组成的矩阵 $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \in \mathbb{R}^{D \times n}$，其中 $\mathbf{x}_i$ 为一数据样本。假设任一样本 $\mathbf{x}_i$ 都可以由其他样本线性表示，则可以得到如下问题：

$$\min_{\mathbf{w}} \sum_{i=1}^{n} \|\mathbf{x}_i - \sum_{j=1}^{n} w_{ij}\mathbf{x}_j\|_2^2 \tag{12}$$

同样，我们可以要求每一样本用尽量少的样本来表示，于是可以得到所谓的**稀疏表示（sparse representation）**问题：

$$\begin{aligned} & \min_{\mathbf{w}_1,\cdots,\mathbf{w}_n} \quad \sum_{i=1}^{n} \|\mathbf{x}_i - \sum_{j=1}^{n} w_{ij}\mathbf{x}_j\|_2^2 + \sum_{i=1}^{n} \lambda\|\mathbf{w}_i\|_1 \\ = \quad & \min_{\mathbf{w}_1,\cdots,\mathbf{w}_n} \quad \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{X}\mathbf{w}_i\|_2^2 + \sum_{i=1}^{n} \lambda\|\mathbf{w}_i\|_1 \end{aligned} \tag{13}$$

通过上面这个问题，可以这样思考，样本 $\mathbf{x}_i$ 在 $\mathbf{X}$ 张成的空间中坐标为 $\mathbf{w}_i$。

对于高维数据来说，样本维数通常远远高于样本数量，即 $D \gg n$。因此通过稀疏表示实际上也完成了一定的特征提取问题。进一步，如果我们能学习一个更好的数据集 $\mathbf{B} = [\mathbf{b}_1, \cdots, \mathbf{b}_k]$ 代替 $\mathbf{X}$，则上述问题可以写成

$$\min_{\mathbf{w}_1, \cdots, \mathbf{w}_n} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{B}\mathbf{w}_i\|_2^2 + \sum_{i=1}^{n} \lambda \|\mathbf{w}_i\|_1 \tag{14}$$

$\mathbf{B}$ 常常称为一个字典，上式即为字典学习问题。上式中需要学习得到 $\mathbf{B}$ 以及 $\mathbf{w}_1, \cdots, \mathbf{w}_n$。

可采用变量交替优化的策略来求解式上式。

STEP 1. 固定字典 $\mathbf{B}$，则对于任一 $\mathbf{w}_i$，求解下式：

$$\min_{\mathbf{w}_i} \|\mathbf{x}_i - \mathbf{B}\mathbf{w}_i\|_2^2 + \lambda\|\mathbf{w}_i\|_1 \tag{15}$$

STEP 2. 固定所有 $\mathbf{w}_i$，求解下式：

$$\min_{\mathbf{B}} \|\mathbf{X} - \mathbf{B}\mathbf{W}\|_F^2 \tag{16}$$

这里 $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_n]$。

求解式（16）可以用 KSVD 方法，即

$$\begin{aligned} \min_{\mathbf{B}} \|\mathbf{X}-\mathbf{B}\mathbf{W}\|_F^2 &= \min_{\mathbf{B}} \|\mathbf{X}-\sum_{i=1}^{k}\mathbf{b}_j\mathbf{w}_i\|_F^2 \\ &= \min_{\mathbf{b}_j} \|\mathbf{X}-\sum_{i\neq j}^{k}\mathbf{b}_i\mathbf{w}_i-\mathbf{b}_j\mathbf{w}_j\|_F^2 \\ &= \min_{\mathbf{b}_j} \|\mathbf{E}_j-\mathbf{b}_j\mathbf{w}_j\|_F^2 \end{aligned} \tag{17}$$

上式的含义即在更新 j 列时，其他列都是固定的，于是 $\mathbf{E}_j=\mathbf{X}-\sum_{i\neq j}^{k}\mathbf{b}_i\mathbf{w}_i$ 也是固定的。通过对 $\mathbf{E}_j$ 进行奇异值分解，求得最大奇异值所对应的正交向量就可以求得 $\mathbf{b}_j$。

注意，直接对 $\mathbf{E}_j$ 进行奇异值分解会同时修改 $\mathbf{b}_j, \mathbf{w}_j$，因此会破坏 $\mathbf{W}$ 的稀疏性。为了避免这样的情况发生，KSVD 对 $\mathbf{E}_j, \mathbf{w}_j$ 分别处理：$\mathbf{w}_j$ 仅保留非零元素，$\mathbf{E}_j$ 仅保留 $\mathbf{b}_j, \mathbf{w}_j$ 的非零元的乘积项，然后再进行奇异值分解，得到 $\mathbf{w}_j$，再将原先删掉的 0 插补回去。

在上述字典学习过程中，用户能通过设置词汇量 k 的大小来控制字典的规模，从而影响到稀疏程度。

5. 矩阵补全

基于部分信息来恢复全部信息的技术在许多现实任务中有重要应用。例如，网上书店通过收集读者在网上对书的评价，可根据读者的读书偏好来进行新书推荐，从而达到走向广告投放的效果。显然，没有哪位读者读过所有的书，也没有哪本书被所有读者读过，因此，网上书店所搜集到的仅有部分信息。

	《笑傲江湖》	《万历十五年》	《人间词话》	《云海玉弓缘》	《人类的故事》
赵大	5	?	?	3	2
钱二	?	5	3	?	5
孙三	5	3	?	?	?
李四	3	?	5	4	?

5. 矩阵补全

能否将上表中通过读者评价得到的数据当作部分信号，恢复出完整信号呢？矩阵补全 (matrix completion) 技术可用于解决这个问题，其形式为：

$$\begin{aligned} \min_{\mathbf{X}} \quad & rank(\mathbf{X}) \\ s.t. \quad & (\mathbf{X})_{ij} = (\mathbf{A})_{ij}, (i,j) \in \Omega \end{aligned} \tag{18}$$

$\mathbf{X}$ 表示需恢复的稀疏信号，$rank(\mathbf{X})$ 表示矩阵 $\mathbf{X}$ 的秩，$\mathbf{A}$ 是观测到信号，Ω 是一个指标集。

求解上式问题是一个 NP 问题，其凸的替代问题可以表示成：

$$\begin{aligned} \min_{\mathbf{X}} \quad & \|\mathbf{X}\|_* \\ s.t. \quad & (\mathbf{X})_{ij} = (\mathbf{A})_{ij}, (i,j) \in \Omega \end{aligned} \tag{19}$$

其中 $\|\mathbf{X}\|_*$ 为 $\mathbf{X}$ 的核范数，$\|\mathbf{X}\|_* = \sum_{i=1}^{D,n} \delta_i(\mathbf{X})$ 为 $\mathbf{X}$ 的所有奇异值的和。上式可以通过半正定规划问题来求解。

Thanks !