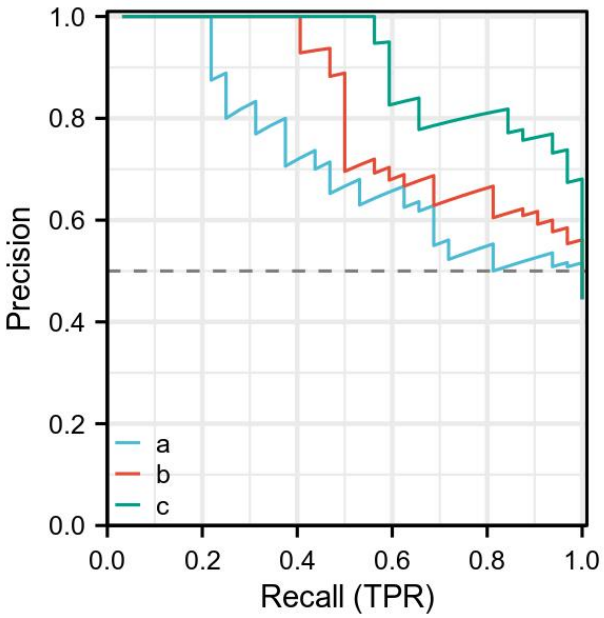
临床意义 - 诊断 PR 曲线



网址：https://www.xiantao.love

更新时间：2023.07.11

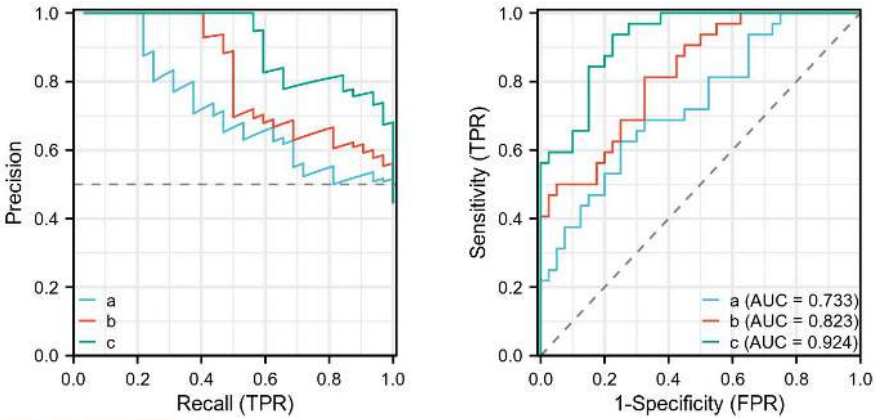
目录

基本概念

- PR 曲线是反映精确率(Precision)与召回率(Recall)之间关系的曲线
  - 精确率(Precision)：Precision = TP/(TP+FP)
  - 召回率(Recall)：Recall = TP/(TP+FN)
- PR 曲线与 ROC 曲线
  - ROC 曲线：受试者工作特征曲线（Receiver Operating Characteristic Curve，ROC 曲线）和 ROC 曲线下的面积（Area Under ROC Curve，AUC）常用于诊断试验的评估，评估预测准确率情况。例如一组数据的结局为 group1 和 group2，变量为 a、b 和 c，也就是评估 a、b 和 c 在预测 group1 和 group2 上的结局，哪个的准确性更高。ROC 曲线图是反映敏感性与特异性之间关系的曲线。AUC 取值范围一般在 0.5 和 1 之间，使用 AUC 值作为评价标准是因为很多时候 ROC 曲线并不能清晰的说明哪个分类器的效果更好，而作为一个数值，对应 AUC 更大的分类器效果更好
    - 真阳率（True Positive Rate, TPR）| 敏感度（Sensitivity）：检测出来的真阳性样本数除以所有真实阳性样本数：TPR = TP/(TP+FN)
    - 假阳率（False Positive Rate, FPR）| 1-特异度：检测出来的假阳性样本数除以所有真实阴性样本数：FPR = FP/(FP+TN)
    - 真阴性率（特异度，Specificity）：检测出来的真阴性样本数除以所有真实阴性样本数
  - 特点：
    - ROC 曲线由于兼顾正例与负例，所以适用于评估分类器的整体性能，相比而言 PR 曲线完全聚焦于正例
    - 不考虑 ROC 方向的问题：在 ROC 空间，ROC 曲线越凸向左上方向效果越好。与 ROC 曲线左上凸不同的是，PR 曲线是右上凸效果越好

- ◆ 类别不平衡问题（正例负例不平衡）中 ROC 曲线确实会作出一个比较乐观的估计，而 PR 曲线则因为 Precision 的存在会不断显现 FP 的影响
- ◆ 评估在相同的类别分布下正例的预测情况，则宜选 PR 曲线

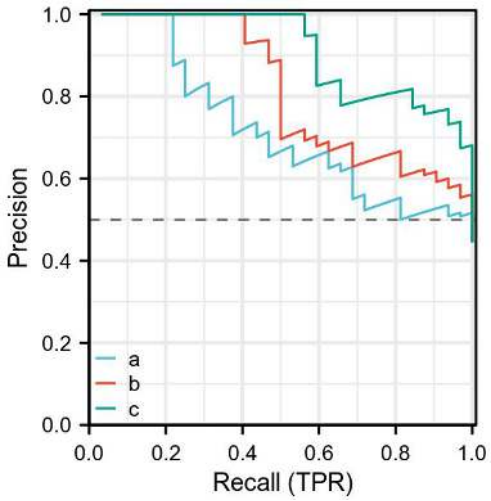
➢ 图形构成：左侧 PR 曲线，右侧 ROC 曲线



应用场景

多应用在医学领域，判断某种因素对于某种疾病的诊断是否有诊断价值。

结果解读



诊断 ROC 曲线

- 横坐标 X 轴为召回率，也称为真阳性率，X 轴越接近 1 准确率越高
  - Recall = TP/(TP+FN)
- 纵坐标 Y 轴称为精确率，Y 轴越大代表准确率越好
  - Precision = TP/(TP+FP)
- PR 曲线与 ROC 曲线的区别：
  - ROC 曲线兼顾正例和负例，故适用于评估分类器的整体性能，而 PR 曲线完全聚焦于正例
  - 类别不平衡问题（正例负例不平衡）中 ROC 曲线确实会作出一个比较乐观的估计，而 PR 曲线则因为 Precision 的存在会不断显现 FP 的影响
  - 从结果来看:
    - ① 当一个变量的值是促进结局(事件)发生的趋势时，该分子的 AUC 会>0.5，此时面积越大(AUC 值越接近于 1)说明结果越好，此时 ROC 曲线下面积越往左上角凸起，说明结果越好，而 PR 曲线则是曲线下面积越往右上角凸起，说明结果越好

- ② 反之，当一个变量的值是促进结局(事件)发生的趋势相反时，该分子的 AUC 会<0.5，此时面积越小(AUC 值越接近于 0)说明结果越好

数据格式

	A	B	C	D
1	outcome	a	b	c
2	group1	1.585854594	1.17428046	2.674787643
3	group1	2.205293427	0.86192791	2.003079333
4	group1	2.199553804	2.31587217	1.281605297
5	group1	1.241118417	1.574637672	1.866428136
6	group1	2.016991788	1.953333649	1.84722131
7	group1	2.391270613	1.089195069	2.149648096
8	group1	0.620790581	0.837543584	1.864922823
9	group1	2.442848378	1.736095106	1.213975139
10	group1	1.636013122	2.414536228	2.946673422
11	group1	1.420847315	2.405175261	1.14110224
12	group1	1.562684913	1.335404124	1.97343619
13	group1	1.034827449	1.373806171	2.070316063
14	group1	2.097985186	2.184310223	2.076315129
15	group1	1.481034166	2.078691615	1.766842006
16	group1	1.020168922	2.388650332	1.862112528
17	group1	0.811700221	1.879728769	2.313832162
18	group1	0.633664805	1.716647154	2.46027284

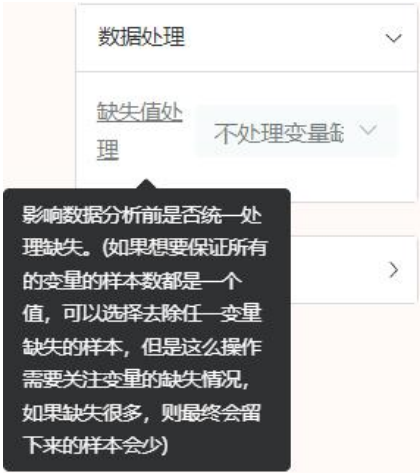
- 第 1 列结局变量（必须是二分类），缺失值不能超过第一列长度的 85%。第 1 列中分类的前后出现的顺序会被参考的顺序，先出现的分类会被当做参考组
- 至少需要 2 列数据，一次最大只支持 11 列数据(10 个自变量)(更多的变量建议是分成 2 张图进行展示)，最少需要 6 行，最多不能超过 10000 行
- 每一行为 1 个样本
- 除第一列外，其他列必须都是数值（为待分析的变量），填入每个样本对应的变量的值

例如：如果是表达谱数据，则 abc 代表想要分析的分子，每一行代表 1 个样本，第 1 列代表样本所属于的分组（想要预测的结局，比如正常 vs 异常）

参数说明

（说明：标注了颜色的为常用参数。）

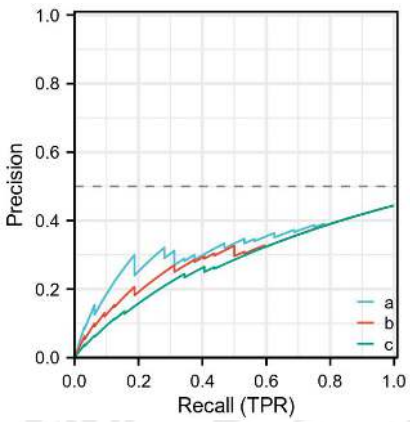
数据处理

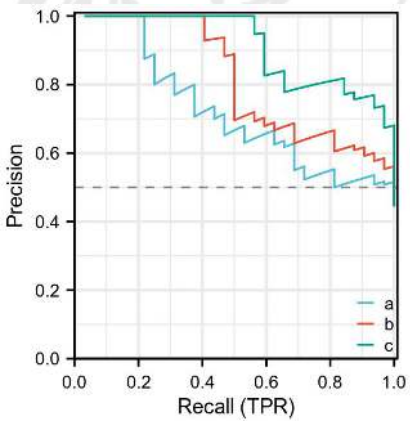


- 缺失值处理：默认是单因素后多因素前处理变量缺失，也可以选择单因素分析前统一删除缺失值

统计

- 方向：可以选择自动、正向或者反向，如下：





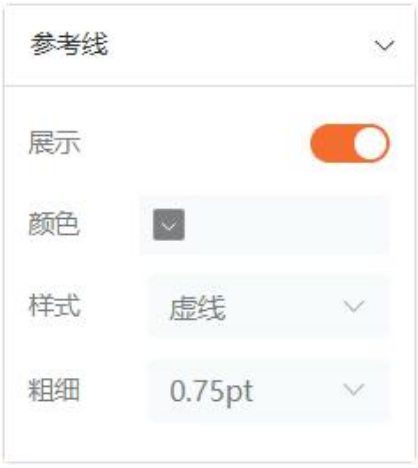
线

- 颜色：可以修改每条曲线的颜色
- 样式：可以修改每条曲线的样式(线条类型)，默认是实线，也可以选择虚线
- 粗细：可以修改每条曲线的线条粗细，默认是 0.75pt
- 不透明度：可以修改每条曲线的不透明度，默认是 1，0 是完全透明，1 是完全不透明

点

- 展示：可以选择是否展示曲线上的点
- 填充色：可以修改点的填充色
- 描边色：可以修改点的描边色
- 样式：点的样式，可以选择圆形、三角形等形状选择
- 大小：点的大小，默认 0.3
- 不透明度：点的不透明度，默认是 1，0 是完全透明，1 是完全不透明

参考线



- 展示：可以选择是否展示 PR 曲线中的参考线，默认为展示
- 颜色：当选择展示参考线时，可以修改参考线的颜色
- 样式：当选择展示参考线时，可以修改参考线的样式，默认为虚线，还可以选择实线类型
- 粗细：当选择展示参考线时，可以修改参考线的线条粗细

标题

- 大标题：大标题内容
- x 轴标题：x 轴标题内容
- y 轴标题：y 轴标题内容

补充：在要换行的中间插入\n。如果需要上标，可以用两个英文输入法下的大括号括住，比如 {{2}}；如果需要下标，可以用两个英文输入法下的中括号括住，比如 [[2]]

图注

- 是否展示：图注内容是否展示
- 图注标题：可以填入图注标题
- 图注位置：默认是左下，还可以选右等

风格

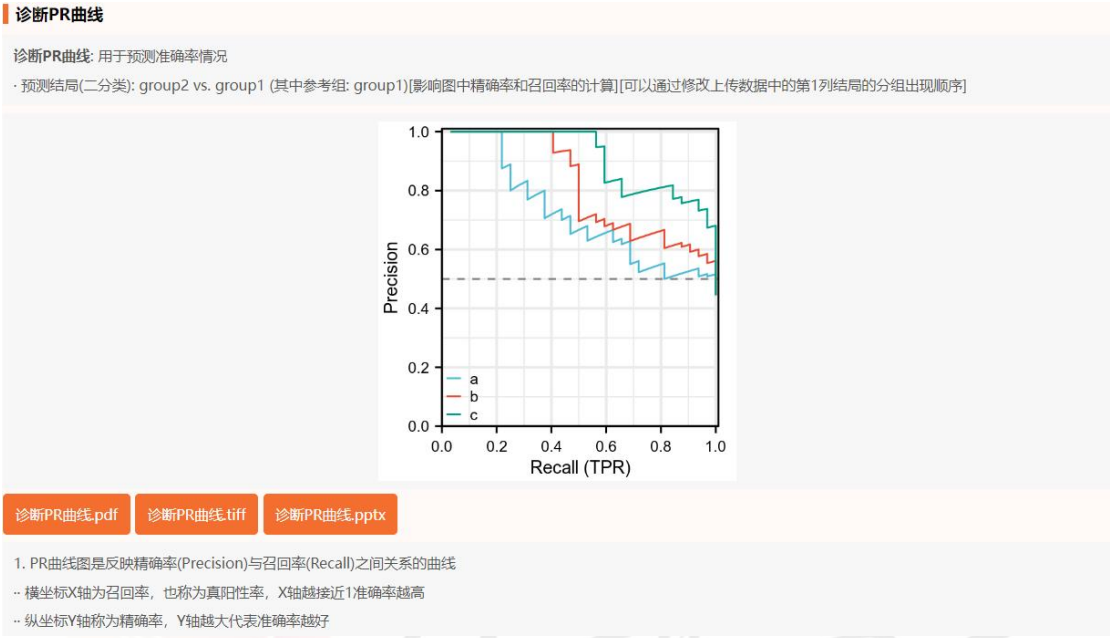
- 边框：是否在图中添加边框
- 网格：是否在图中添加网格线
- 文字大小：图中的文字部分的大小（包括标签文字和刻度数），默认是 7pt

图片

- 宽度：图片横向长度，单位为 cm
- 高度：图片纵向长度，单位为 cm
- 字体：可以选择图中文本内容字体

结果说明

主要结果



主要结果格式为图片格式，提供 PDF、TIFF 、PPT 格式下载。

补充结果

1. 统计描述表：上传数据的一些基本情况

统计描述

各个组常见「统计描述指标」

结局	变量	数目	最小值	最大值	中位数(Median)	四分位距(IQR)	下四分位	上四分位	均值(Mean)	标准差(SD)
group1	a	40	0.58634	2.4428	1.4531	0.82236	1.1199	1.9423	1.5112	0.55454
group1	b	40	0.55563	2.4946	1.8003	1.0719	1.137	2.2089	1.6463	0.61858
group1	c	40	1.0473	2.9467	2.0733	0.66516	1.7701	2.4352	2.0366	0.52699
group2	a	32	0.1287	1.9882	0.99585	0.87824	0.62288	1.5011	0.99804	0.55444
group2	b	32	0.022259	1.9617	0.77933	0.85985	0.32455	1.1844	0.84336	0.57161
group2	c	32	0.030092	1.8936	0.9231	0.96898	0.42143	1.3904	0.93236	0.54622

统计描述.xlsx

2. AUC 结果表

AUC结果表

预测变量	预测结局	曲线下面积(AUC)	置信区间(CI)
a	反向	0.733	0.617 - 0.849
b	反向	0.823	0.730 - 0.917
c	反向	0.924	0.868 - 0.981

预测结局中，正向或者反向会影响真/假阳性和真/假阴性的区分(如果统计-方向参数选择的是“自动”，则会对结局的方向会进行调整保证曲线都是往上凸(pROC包提供))(如果选择“正向”或者“反向”，则图形有可能会向下凹)

在AUC > 0.5的情况下，AUC越接近于1，说明该变量在预测结局上诊断效果越好。

AUC在0.5～0.7时有较低准确性，AUC在0.7～0.9时有一定准确性，AUC在0.9以上时有较高准确性。

AUC = 0.5时，说明该变量不起作用，无诊断价值。

3. ROC 信息表

ROC信息表

预测变量	cut-off值	灵敏度	特异度	准确率	真阳个数	真阴个数	假阳个数	假阴个性	阳性预测值	阴性预测值	约
a	1.1155	0.625	0.75	0.69444	20	30	10	12	0.66667	0.71429	0.
b	1.2633	0.8125	0.675	0.73611	26	27	13	6	0.66667	0.81818	0.
c	1.7437	0.9375	0.775	0.84722	30	31	9	2	0.76923	0.93939	0.

各预测变量在各自最佳cut-off值下部分ROC相关信息和数据。

方法学

统计分析和可视化均在 R 4.2.1 版本中进行

涉及的 R 包：pROC[1.18.0] 用于 PR 分析

1. 使用 pROC 包对数据进行 PR 分析

2. 结果用 ggplot2 进行可视化

3. pROC 包默认会对数据的结局顺序进行校正(保证结果是往上凸)

如何引用

生信工具分析和可视化用的是 R 语言，可以直接写自己用 R 来进行分析和可视化即可，可以无需引用仙桃，如果想要引用仙桃，可以在致谢部分（Acknowledge）致谢仙桃学术（www.xiantao.love）。

方法学部分可以参考对应说明文本中的内容以及一些文献中的描述。

常见问题

1. AUC 会出现< 0.5 的情况吗?

答：一般情况下，pROC 分析结果中 AUC 面积是在 0.5-1 之间。

2. 1 个组的时候为什么没有给出统计学检验的 p 值?

答:

PR 曲线与 ROC 曲线相似，一般是看 AUC 的大小的，只有当存在有多个曲线的时候才会进行检验比较。如果只有 1 条曲线，是没办法进行统计检验的，除非是跟 0.5 的对角线比，这种比较其实是没有意义的，这种只要 AUC 的下限没有跨过 0.5，那么这个曲线肯定是有意义的，所以单个曲线是没有统计学比较的意义。

3. 数据的结局是以哪个作为阴性（参考）？哪个作为阳性（实验）？

答:

默认上传数据的第一列（二分类）以第一个出现的分类组参考，后出现的分类作为实验。这个方向会影响最终的真阳、真阴、假阳、假阴个数。如果需要反过来，可以在<统计>-<方向>参数中进行修改。