互信息(Mutual Information)近似计算

matricer

August 8, 2018

我们考虑将类别型特征的AUC的近似计算方法平移到MI上。类别型特征的AUC的近似计算方法依赖于下述命题:

命题1:设类别型特征的统计信息为$\{(n_i, p_i)\}_{i=1,\dots,N}$,其中,$N$为互不相同的类别数。$n_i$为第$i$个类别的计数,$p_i$为第$i$个类别的正样本计数。记$s_i = \frac{p_i}{n_i}$为第$i$个类别的正样本命中率。我们将$s_i$精确相等的类别进行统计信息叠加,Ji Ru Guo $s_i = s_j$,则将(n_i, p_i)与(n_j, p_j)叠加为$(n_i + n_j, p_i + p_j)$。记叠加前的auc为auc_o,叠加后的auc为auc_m。我们有$auc_o = auc_m$。

我们可以将上述命题推广到互信息上。

命题2:设类别型特征的统计信息为$\{(n_i, p_i)\}_{i=1,\dots,N}$,其中,$N$为互不相同的类别数。$n_i$为第$i$个类别的计数,$p_i$为第$i$个类别的正样本计数。记$s_i = \frac{p_i}{n_i}$为第$i$个类别的正样本命中率。我们将$s_i$精确相等的类别进行统计信息叠加,Ji Ru Guo $s_i = s_j$,则将(n_i, p_i)与(n_j, p_j)叠加为$(n_i + n_j, p_i + p_j)$。记叠加前的mi为mi_o,叠加后的mi为mi_m。我们有$mi_o = mi_m$。

为了证明命题2,只需证明如下命题: Ru Guo $s_i = s_j$,则

$$\left(\frac{p_i}{c}\log\frac{\frac{p_i}{c}}{\frac{n_i}{c}r_p} + \frac{n_i-p_i}{c}\log\frac{\frac{n_i-p_i}{c}}{\frac{n_i}{c}r_n}\right)+\left(\frac{p_j}{c}\log\frac{\frac{p_j}{c}}{\frac{n_j}{c}r_p} + \frac{n_j-p_j}{c}\log\frac{\frac{n_j-p_j}{c}}{\frac{n_j}{c}r_n}\right)$$

$$=\left(\frac{p_i+p_j}{c}\log\frac{\frac{p_i+p_j}{c}}{\frac{n_i+n_j}{c}r_p} + \frac{n_i+n_j-p_i-p_j}{c}\log\frac{\frac{n_i+n_j-p_i-p_j}{c}}{\frac{n_i+n_j}{c}r_n}\right)$$

我特征的AUROC的近似计算依赖于下述命题。命题1:设类别型特征的统计信息为