

# به نام خدا

## معرفی دیتاست و متغیرهای اصلی

برای مشاهده عملکرد معیارهای بررسی پیچیدگی مسئله ی دسته بندی که در ادامه شرح میدهم از دیتاست آماده ی iris استفاده کرده ام  
پیش از آنکه به معیارها بپردازیم لازم است دیتاست را بگونه ای مناسب برای محاسبات دریاوریم

cls\_index

براین متغیر مشخص شده است که کدام رایه ها متعلق به چه کلاسی هستند  
بعنوان مثال در این دیتاست سه کلاس داریم  
همچنین این متغیر برای خود دارای سه رایه است  
از آنجا که این دیتاست دارای ۱۵۰ عضو است سه رایه ای که در این لیست هستند هر یک ۱۵۰ عضو دارند، در رایه  
اول اعضای کلاس اول برچسب true دارند و اعضای بقیه ی کلاس ها برچسب false  
این قضیه در مورد رایه دوم برای کلاس دوم و رایه سوم برای کلاس سوم ادامه دارد

cls\_n\_ex

در این لیست تعداد اعضای هر کلاس مشخص شده است  
عضو اول نمایانگر تعداد اعضای کلاس اول است و به همین ترتیب الی آخر

ovo\_comb

برای بررسی پیچیدگی دیتاست، بهتر است در صورتی که بیشتر از دو کلاس داریم هر بار مسئله را شکسته و به دو کلاس  
تقلیل دهیم. در این لیست مشخص میکنیم که چه کلاسهایی باید دو به دو بررسی شود. مثلا در مورد دیتاست iris داریم  
[[ (0, 1), (0, 2), (1, 2) ]]

## Feature-based Measure

### Maximum Fisher's Discriminant Ratio (F1)

$$r_{f_i} = \frac{\sum_{j=1}^{n_c} n_{c_j} (\mu_{c_j}^{f_i} - \mu^{f_i})^2}{\sum_{j=1}^{n_c} \sum_{l=1}^{n_c} (x_{li}^j - \mu_{c_j}^{f_i})^2},$$

numerator function

به زای فیچر خواسته شده، در تک تک کلاس ها اختلاف میانگین فیچر در هر کلاس را از میانگین فیچر به زای همه  
ی داده ها حساب میکند و به توان دو میرساند و در تعداد اعضای کلاس ضرب میکند، سپس همه ی این اعداد را با  
یکدیگر جمع میکند تا صورت کسر ساخته شود

denominator function

به زای فیچر مشخص شده، اختلاف هر مقدار را از میانگین فیچر در آن کلاس کم کرده و مربع آن را بدست می آوریم و  
حاصل جمع این عبارت را به زای همه ی نمونه ها و در همه ی کلاسهها بدست می آوریم

compute\_rfi function

به ازای هر ویژگی با استفاده از دو تابعی که در بالا گفته شد مقدار rfi را محاسبه میکنیم و درون یک لیست ریخته و بلزمیگردانیم

ft\_F1 function

ماکزیمم مقدار لیستی که با استفاده از تابع بالا بدست آمده را بدست می آوریم و با استفاده از آن با توجه به رابطه زیر مقدار F1 محاسبه میشود

$$F1 = 1/(1 + \max)$$

در واقع rfi به ازای هر فیچر نشاندهنده ی آن است که تا چه میزان داده ها توسط این ویژگی تفکیک پذیرند و به طور کلی هرچه این معیار کوچکتر باشد یعنی با دیتاست راحت تری سر و کار داریم و میتوان فیچری را یافت که با استفاده از آن فیچر بتوان هاپیرپلینی بدست آورد و داده هارا تفکیک نمود

---

## The Directional-vector Maximum Fisher's Discriminant Ratio (F1v)

dVector function

این تابع به ازای دو کلاسی که بعنوان آرگومان ورودی اندیس آنها را دریافت کرده مقادیر W, d, B را محاسبه میکند و با استفاده از آنها dF را ساخته و به عنوان خروجی برمیگرداند

در قدم اول به ازای هر فیچر میانگین مقادیر را در هر کلاس محاسبه میکند

سپس مقدار W را که بعنوان Scatter matrix شناخته میشود با توجه به نسبت جمعیتی داده های دو کلاس و کواریانس محاسبه میکند

سپس با توجه به W و براساس روابط زیر مقادیر d, B, dF را محاسبه میکند

$$\mathbf{W} = p_{c_1} \Sigma_{c_1} + p_{c_2} \Sigma_{c_2},$$

$$\mathbf{d} = \mathbf{W}^{-1}(\mu_{c_1} - \mu_{c_2}),$$

$$\mathbf{B} = (\mu_{c_1} - \mu_{c_2})(\mu_{c_1} - \mu_{c_2})^t$$

$$dF = \frac{\mathbf{d}^t \mathbf{B} \mathbf{d}}{\mathbf{d}^t \mathbf{W} \mathbf{d}},$$

$$F1v = \frac{1}{1 + dF}$$

به ازای تمامی جفت کلاسها مقدار dF و براساس آن F1v محاسبه میشود و نهایتا میانگین تمام F1v ها بعنوان مقدار نهایی بازگردانده میشود  
در مورد این معیار نیز همانند معیار F1 هر چه مقدار F1v کوچکتر باشد نشان دهنده ی آن است که با کلاس ساده تری

## Volume of Overlapping Region (F2)

$$F2 = \prod_i^m \frac{\text{overlap}(f_i)}{\text{range}(f_i)} = \prod_i^m \frac{\max\{0, \min \max(f_i) - \max \min(f_i)\}}{\max \max(f_i) - \min \min(f_i)},$$

توابع `_minmin`, `_maxmin`, `_minmax`, `_maxmax` برای یافتن مقادیر مورد نیاز پیاده سازی شده اند

کلاس ها را به صورت دو به دو به این توابع میدهم و هر کدام از این توابع لیستی به طول فیچرهای دیتاست برای ما برمیگرداند. سپس با ساختن لیستی تمام صفر به طول تعداد فیچرها و با توجه به لیستهای خروجی توابع `_minmax` و `_maxmin` صورت کسر مورد نظر را بدست می آوریم و با توجه به لیستهای خروجی توابع `_maxmax`, `_minmin` مخرج کسرها و بدین ترتیب برای هر جفت کلاس معیار `F2` را محاسبه میکنیم مقدار نهایی `F2` میانگین مقادیر بدست آمده به ازای هر جفت کلاس است

مقادیر کم `F2` بیانگر کم بودن اورلپ بین کلاسهای گوناگون است و بالعکس. همچنین میدانیم داشتن اورلپ بین کلاسها به معنی رویارویی با یک دیتاست سخت است

---

## Maximum Individual Feature Efficiency (F3)

با استفاده از توابع `_minmax`, `_maxmin` که پیش از این تعریف کردیم، به ازای هر جفت کلاس این مقادیر را حساب حساب کرده و درون تابع `compute_f3` براساس این مقادیر و با کمک تابع `and` منطقی حاصل جمع زیر را بدست می آوریم

$$n_o(f_i) = \sum_{j=1}^n I(x_{ji} > \max \min(f_i) \wedge x_{ji} < \min \max(f_i))$$

سپس از آنجایی که حاصل یک لیست به طول تعداد فیچرهاست کمترین مقدار را بدست آورده و بعنوان حاصل برمیگردانیم. نهایتاً براساس تعداد اعضای هر کلاس و طبق رابطه زیر معیار `F3` را محاسبه کرده و در صورتیکه بیشتر از یک کلاس داشته باشیم میانگین آنها به عنوان `F3` نهایی برمیگردانیم

$$F3 = \min_{i=1}^m \frac{n_o(f_i)}{n},$$

کمینه حاصل جمع یادشده توسط تابع `compute_f3` بدست می آید و برای محاسبه `F3` از تابع `ft_F3` استفاده میکنیم

مقادیر کم `F3` بیانگر کم بودن اورلپ بین کلاسهای گوناگون است و بالعکس. همچنین میدانیم داشتن اورلپ بین کلاسها به معنی رویارویی با یک دیتاست سخت است

---

## Colective Feature Efficiency (F4)

برای این معیار باید فیچرهایی که در کلاس ها بیشترین تمایز را بوجود می آورند را بیابیم و تمام نمونه هایی که توسط آنها تفکیک پذیرند را از داده ها حذف کنیم تا زمانی که هیچ مثالی براساس فیچرها قابل تفکیک نباشد و براساس تعداد داده ها نسبت به تعداد کل داده های دو کلاس معیار F3 این را بدست می آوریم

اگر تعداد کلاسها از دو کلاس بیشتر باشد میانگین این معیار به زای هر جفت کلاس بعنوان مقدار گزارش میشود

به راحتی توسط تابع `compute_f3` میتوانیم این نواحی مشترک که به راحتی تفکیک پذیر نبوده اند را پیدا کنیم و در هر ایتريشن عملیات را بر روی همین نواحی ادامه دهیم

همچنین در هر ایتريشن تاثیرگذارترین فیچر را یافته و آنرا از داده ها حذف کنیم

این فرآیند تا زمانی ادامه می یابد که حداقل یک مثال یا یک فیچر در مجموعه داده ها باقی مانده باشد

بزرگ بودن معیار F4 به معنی وجود درصد بیشتری داده هایی است که با توجه به فیچرهایی که داریم تفکیک پذیر نیستند و این بدین معناست که برای دسته بندی این دیتاست با مسئله ی سخت تری مواجه هستیم